Language Testing in Asia

RESEARCH Open Access

The application of Kunnan's test fairness framework (TFF) on a reading comprehension test

M. Moghadam[1*] and F. Nasirzadeh[2]

* Correspondence: Moghaddam.m@fasau.ac.ir
[1]Faculty of Science, Fasa University, Fasa, Iran
Full list of author information is available at the end of the article

Abstract

The present study tries to investigate the fairness of an English reading comprehension test employing Kunnan's (2004) test fairness framework (TFF) as the most comprehensive model available for test fairness. The participants of this study comprised 300 freshman students taking general English course chosen based on the availability sampling, three test developers, and seven university officials, administering the test. The main instrument is the teacher-made reading comprehension test to examine its validity, reliability, and differential item functioning (DIF). Furthermore, to examine the other modules of TFF, namely, access, administration, and social consequence, a questionnaire and semi-structured interviews with the test developers and test administrators were applied. In order to analyze the data, exploratory factor analysis is used to evaluate test validity employing Minitab software. Moreover, *t* test and ANOVA were used to examine the disparate impact of the test using SPSS package. Furthermore, the Mantel-Haenszel procedure is applied to determine the DIF, while coding the required formulas in R programming. The frequencies of different aspects of access and test administration are explained and consolidated by qualitative data gleaned from interview sessions. Examining the data with respect to TFF modules, it was concluded the test must be enhanced in terms of validity, while it is totally fair. The statistical procedure and the mixed research design implemented in the present study can be a sound model to be applied by test developers to enhance the test fairness of the exams.

Keywords: Test fairness, Kunnan's test fairness framework, Differential item functioning, Validity

Introduction

Fairness is one of the key issues that concern people about any testing procedure. According to Flaugher (1973), the theory of test fairness was shown to be a complex issue, a discovery that had immediate real-life implications, in a time of increased attention to the fair treatment of minority groups. In this way, different scholars such as Shohamy (2001), Weir (2005), and Fulcher and Davidson (2007) claimed that testing must be under careful examination because tests are mostly used for making high-

stake decisions. Therefore, according to Stobart (2005), effort must be made so that the tests be as fair as possible for the groups who want to take them.

Test fairness has not been paid due attention for a long time. Gradually, measurement professionals "began to pay increasing attention" to test and item fairness almost at the beginning of 1970s (Cole & Zieky, 2001, p. 370). However, Kunnan (2010) mentioned that test fairness entered the forefront of investigations and discussions in the field of language assessment in 1990s. As Baharloo (2013) mentioned, fair judgment is an essential requirement for measurement professionals to be aware of the concept and its characteristics.

Fairness is such a complicated and multi-faceted issue that a variety of definitions has been proposed to clarify its broad and controversial nature. According to Webster's Ninth New Collegiate Dictionary (1988), fairness means, being free from having favor toward either or any side. However, Xi (2010) believed that fair testing mainly focuses on comparing testing practices and test results across different groups. Therefore, test fairness mainly arises from the way group differences are perceived and treated. Thus, Xi (2010) defines fairness "as comparable validity for all the identifiable and relevant groups across all stages of assessment, from assessment conceptualization to the use of assessment results" (p. 154).

Baharloo (2013) claimed that fairness is not confined to the content of a test and covers other aspects of testing as well. Although some fairness models have been proposed, to the best of the researcher's knowledge, none lends itself easily to practical investigation of fairness. Therefore, the current study is an endeavor to provide a comprehensive portrait of test fairness by discussing Kunnan's (2004) framework as the most comprehensive model available for test fairness and practically examining its modules using a real data obtained through a teacher-made reading exam.

To this end, the purpose of the present study is to investigate the fairness of a locally designed English reading comprehension test which the students of higher education must pass as their general English basic course. This test is designed by the instructors of the course jointly, and it aims to evaluate the students' adroitness in implementing the reading strategies taught during the term. In this way, the content of the exam is determine by the materials covered during the term.

The significance of this study lies in seeking ways to develop a fair general reading comprehension test which is the basic course in higher education at bachelor level for all majors. Given that it is impossible for a test to be perfectly fair, and that some tests are patently unfair, test takers are always prone to suffer the adverse test effects. Thus, it is worth pointing out that the outcome of the present study can be of great benefit to the examinees. Secondly, considering the Iranian EFL context, the present research is the first-ever study aiming at a systematic, in-depth investigation into test fairness based upon a step by step statistical procedure, which in turn can motivate others to examine test fairness. Thirdly, although, as Xi (2010) mentioned, conceptual frameworks for fairness have gained considerable momentum in recent years, empirical studies motivated by these frameworks lag far behind.

As the main body of the present study lies upon the Kunnan's test fairness framework (TFF), in the following section, the modules and the content of this framework are elaborated.

Theoretical framework

Although the concept of test fairness was dramatically soared after the birth of test evaluation approaches, there was no coherent framework to evaluate test fairness until 2000s. Among the ethics and principle-based approach to test fairness, Kunnan (2000) proposed an ethics-inspired rationale entitles the test fairness framework (TFF) with a set of principles and sub-principles. This framework focused on three test qualities, including validity, absence of bias, and social consequences. Then, Kunnan developed a new framework in 2004 in which qualities of access and administration were also emphasized.

Throughout TFF, Kunnan (2004) viewed fairness in terms of the whole system of a testing practice, not just the test itself. Therefore, multiple facets of fairness that includes multiple tests uses (for intended and unintended purposes), multiple stakeholders in the testing process (test takers, test users, teachers, and employers), and multiple steps in the test development process (test design, development, administration, and use) are implicated. Two general principles of justice and beneficence and sub-principles underlying this framework are explained further.

The principle of justice tries to ensure that a test ought to be fair to all test takers. This includes two sub-principle which state that any test ought to have comparable construct validity in terms of its test-score interpretation for all test takers and secondly, it ought not to be biased against any test taker groups, in particular by assessing construct-irrelevant matters.

Furthermore, the principle of beneficence noted that a test ought to bring about good in society, that is, it should not be harmful or detrimental to society. In this way, a test ought to promote good in society by providing test score information and social impacts that are beneficial to society and it ought not to inflict harm by providing test-score information or social impacts that are inaccurate or misleading.

Based on the mentioned principles, TFF has five main qualities: *validity*, *absence of bias*, *access*, *administration*, and *social consequences*. Figure 1 illustrates the TFF within the circle of tests and testing practice where validity is at the center of the framework and the other qualities have their distinct role besides validity.

As it can be seen, one of the important components of this framework is validity which can be described appraising content representativeness or coverage evidence,

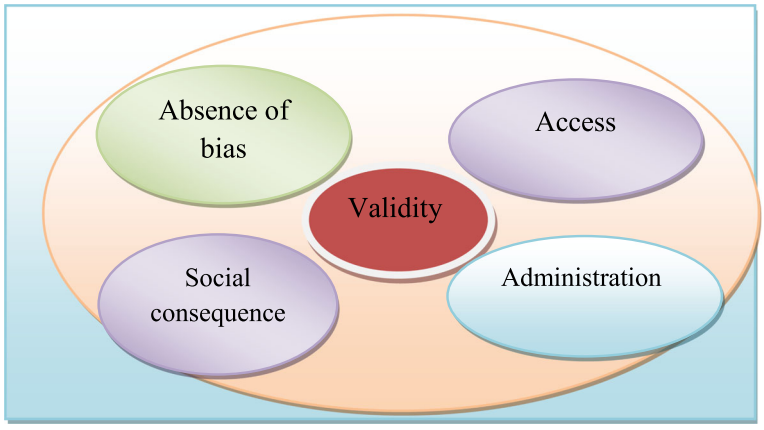

Fig. 1 The main components of Kunnan's test fairness framework

construct or theory-based validity evidence (construct validity), criterion-related validity evidence (criterion validity), and reliability.

The other module includes absence of bias which takes into account the content or the language of the test, disparate impact, and standard setting. This, firstly, requires that the content, language or dialect of the test that is offensive or biased to test takers from different backgrounds (based on gender, race and ethnicity, religion, age, native language, national origin, and sexual orientation) should be modified. Secondly, different performances and resulting outcomes by test takers from different group memberships should be determined through differential item/test functioning (DIF/DTF). In addition, a differential validity analysis should be conducted in order to examine whether a test predicts success better for one group than for another.

With respect to the standard setting, test scores should be examined in terms of the criterion measure and selection decisions. Test developers and score users need to be confident that the appropriate measure and statistically sound and unbiased selection models are in use. These analyses should indicate to test developers and score users that group differences are related to the abilities that are being assessed and not to construct-irrelevant factors.

Considering the third module, access, the test developers should make sure that the test enjoys educational, financial, geographical, and conditions or equipment access. This means that a test should be accessible to test takers in terms of opportunity to learn the content, financially affordable to test takers, accessible in terms of distance to test takers, offers certified test takers with physical and learning disabilities with appropriate test accommodations, and finally, ensures that whether test takers are familiar with test-taking equipment, procedures, and conditions.

Administration, as the next module in TTF, can be explained in terms of providing appropriate physical conditions such as optimum light, temperature and facilities, uniformity in test administration by observing consistency across test sites, equivalent forms and instructions, and proper test security.

The final module of TTF includes social consequences. This can be observed when sufficient evidence regarding washback and remedies can be collected. This required the test developers to consider the effect of a test on instructional practices and try to reverse the detrimental consequences of a test such as re-scoring and re-evaluation of test responses, and legal remedies, once needed.

Literature review

Literature on test fairness indicates that fairness has had different conceptualizations in its relatively short history. Moreover, some philosophical approaches toward test fairness divided fairness into procedural and substantive fairness (Kane, 2010), while Davis (2010) considered test fairness chimerical and Xi (2010) investigated fairness argument in relation to a validity argument.

Given the contentious debate on the relation between the concepts of fairness and validity, Xi (2010) referred to three conceptualizations of fairness mentioned by Willingham and Cole (1997). The categorizations conceptualized fairness as an independent test quality bearing little or no relation to validity; secondly, fairness as a broad test quality subsuming validity and finally, fairness as an important

aspect of validity. This type of categorizations, in turn, give birth to a variety of approaches to test fairness.

In line with the purpose of the present study, Kunnan (1997) claimed that language testing research had always explored fairness within the concepts of validity and reliability. He argued:

> Although validation studies are said to generally take on the role of investigating the fairness of tests and testing practices, an examination of about 100 validation studies shows that the themes addressed by the researchers are not particularly concerned with fairness. (p. 85)

On the other hand, some other testing professionals hold the view that fairness has been dealt with within concepts like bias, justice, and equality. However, the related literature reveals that bias and justice do not build up an accurate and comprehensive picture of test fairness. The 1960s through 1980s represented a period of expansion of bias-based approach. These studies began with the narrow focus on test and item bias studies and then developed into the technical literature now known as DIF (differential item functioning) studies. As Kunnan (2008) argues, "the bias-based approach is fragmentary at best as all tests are not evaluated using a uniform concept of fairness" (p. 231).

Notwithstanding the biased-based approach unaccountability, in the late1980s, the time was ripe for the emergence of full-fledged approaches to fairness investigation. Thus, the code-based approach, as a major initiative to examine the uniform concept of fairness, was introduced in 1988. The 1988 Code presents standards for educational test developers and users in four areas: developing and selecting tests, interpreting scores, striving for fairness, and informing test takers. Here is the excerpt from Section C, striving for Fairness divided into two parts, one for test developers and one for test users:

> Test developers should strive to make tests that are as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions. Test users should select tests that have been developed in ways that attempt to make them as fair as possible for test takers of different races, gender, ethnic backgrounds, or handicapping conditions. (1988 Code, p. 4-5)

In the 1990s, standard-based approach came into play. In the 1999 Standards, the concern for fairness in testing is pervasive and the treatment accorded to the topic here cannot do justice to the complex issues involved. Four principal ways were supposed in terms of test fairness.

First, fairness is linked to absence of bias. Second, fairness is defined as providing all examinees with even-handed treatment in the testing process. According to the third definition, "examinees of equal standing with respect to the proposed construct should on average earn the same test score, irrespective of group membership." (the 1999 Standards, p. 74). Finally, fairness is defined as equality of chances to learn the materials incorporated in an achievement test. (Kunnan, 2008). With respect to the empirical studies on test fairness, although scanty, a number of empirical research articles which made use of Kunnan's test fairness framework are reported.

Zheng and Cheng (2008) appraised College English Test (CET), a large-scale language test with distinctive Chinese characteristics, based on Kunnan's (2004) concept of test fairness. They explained CET in almost every particular module and reported on a number of studies conducted on CET. Finally, they generally found CET fair to test Chinese students' English language ability. This study enjoys some preliminary statistical procedures, although most of the analysis is based on comparing the formats of other similar tests in terms of fairness.

Pishghadam and Tabataba'ian (2011) investigated the relationship between IQ and test format considering test fairness. Wechsler's IQ test and a reading test which included four test formats were used as the instruments of their study. The results of the correlational study, the *t* test between high and low IQ groups regarding certain test formats, and the regression equations showed the effect of test formats and level of IQ on test fairness. The mentioned study examined the effect of two factors on test fairness; however, the fairness of the test itself was not taken into account.

In the same way, Khoii and Shamsi (2012) investigated the effects of different test formats on the measurement of grammatical knowledge by comparing the construct validities of two different formats of error-identification grammar tests and a multiple-choice grammar test. They compared these formats in MA entrance exam with TOEFL error identification test and SAT test. The results of the study lent support to the idea that the test method facet could be a strong source of bias in language testing affecting the fairness of the decisions in relation to admitting students to specific academic programs.

Loh and Shih (2016) reviewed the English language test as part of Primary School Leaving Examination in Singapore. Describing the characteristics and background of the test, they adopted Kunnan's test fairness framework to examine the fairness of the test. They found that the validity of the test had been enhanced because of improvements to the test design in the past two decades and also sought the proactive access arrangements and stringent, efficacious administration of the test in place. Since they adopted an evidence-based approach to examine fairness, they examined secondary sources to base their arguments; therefore, no statistical procedures were taken into account.

Hamid, Hardy, and Reyes (2019) examined test takers' perceptions and evaluations of the fairness, justice, and validity of global tests of English, with a particular focus upon the International English Language Testing System (IELTS). Based on relevant literature and on self-reported test experience data gathered from test takers from 49 countries, they demonstrated how test takers experienced fairness and justice. They found that the participants expressed concerns about whether IELTS was a vehicle for raising revenue and for justifying immigration policies, thus raising questions about the justness of the test. Their research foregrounded the importance of focusing attention upon the socio-political and ethical circumstances over the large-scale, standardized testing situations with respect to test fairness.

In sum, the foregoing illustrates the point that despite the fact that test fairness research has taken massive steps in providing us with an ample scope of fairness, it has nearly come to nought when it comes to empirical investigation of the notion of fairness. It should also be noted that there is only scanty evidence either supporting or contradicting the abovementioned conceptual frameworks in practice. A yet more

discouraging point is that little effort has been made to exploit wider fairness frameworks like that of Kunnan (2002, Kunnan, 2004) to evaluate test fairness. Because of the reported dearth of literature, this study made an attempt to investigate test fairness of a teacher-made reading comprehension test based on the mentioned framework using a sound statistical procedure based on a mixed research design.

Method

Participants

The participants of this study were composed of three groups. The first group was 300 freshman students of Fasa University, including 170 female and 130 male students, taking a general English test to certify their proficiency and readiness in English and to pass their course. Eliminating the outliers, the results taken from 81 examinees (29 male and 52 female) were analyzed in the current study. Their age range is between 18 and 24. Since they had been accepted in the university entrance exam quite recently, they were homogenous in terms of their general score range to be accepted in Fasa University and the percentage of their English proficiency scored in the national entrance exam was between 40 and 52. This group of participants is chosen based on the availability sampling and their consent to use their scores for the purpose of the study is taken.

The second group of participants was three test developers who developed the general English test, the main instrument of the present study, for the university students each year. They were instructors of general English course at Fasa University with the experience of teaching and test designing between 5 and 10 years. They had designed and administered the present test for 3 years (6 academic semesters). They were familiar with the basics of testing, and they have MA degree in ELT.

The third group includes 7 university officials, administering the test and preparing the settings for the students to participate in exam session. They were interviewed in this study to elicit the information concerning the administration procedure.

Instrument

The main instrument is the teacher made general English proficiency test consisted of 45 items on general vocabulary (15 items) and reading comprehension (30 items). The purpose of using this test is to examine the students' performance scores to determine the validity, reliability, and DIF of the test which will be reported in the following sections.

To determine the educational, geographical, and economical access and administrational conditions of this test, a questionnaire is designed and the test takers are required to answer each question on a Likert scale from very high to very low. The questionnaire contains 30 items, selected from an item pool collected by the researchers based on the fairness framework specifications and the points mentioned in other studies. The selected items were assessed in terms of validity and reliability by scattering the first draft of the questionnaire on thirty participants in a pilot phase. The reported Cronbach alpha coefficient was 0.84 and the test was valid in terms of face validity, checked by three adroit instructors, and construct validity checking the exploratory factor analysis.

Moreover, to consolidate the results of the study a semi-structured interview was conducted to collect data from test developers, test administrators, and test takers. The items addressed the administrational conditions of the test, different aspects of the test takers' access, and the social consequences predicted by the test developers. The contents of the questions were identified and edited by the researchers based on the TTF and demographic information obtained in the questionnaire.

Data collection

The data needed for this study is collected from different sources. Firstly, the scores of students' performances on the test are used to determine the validity and reliability of the test. The responses to all items of the test are inserted into SPSS, assigning 1 to the correct responses and 0 to the incorrect ones. The sum of all correct responses is labeled as total score.

Moreover, to check the absence of bias, the quantitative scores are used to calculate the DIF of the test items. Secondly, the data obtained from the questionnaire items were used to ask students about their access to the test, the way the test is administered, and the social consequences. At last, the qualitative data obtained from the interview with the test takers, test designers, and test administrators were transcribed, and the primary contents were codified and analyzed to consolidate the results gained in the second phase of the study.

Data analysis

To analyze the obtained data, the following procedure is implemented. To check the validity of the designed test, or more specifically, construct validity, exploratory factor analysis (EFA) was employed to identify hypothetical factors that account for the patterns of correlations that are observed in test scores. To calculate EFA, firstly, the assumption of the test of sphericity was determined by Bartlett's test and Kaiser-Myer Olkin test. Then, to determine the number of extracted factors, the eigenvalues obtained from the initial extraction using the criteria of substantive importance and the Scree test were examined.

After identifying the number of factors, rotated factor structures and oblique factor solutions were examined (assuming that the factors are likely to have a strong positive or negative correlation). Finally, to observe the validity, test performance should be explained with respect to the extracted factors. Minitab software was implemented in line with SPSS to examine the required procedures.

To examine the second module of TTF, absence of bias, three main criteria were inspected. Firstly, the test should be checked in terms of content or language. The content or language or dialect that is offensive or biased to test takers from different backgrounds (based on gender, race and ethnicity, religion, age, native language, national origin) or choice of dialect that is biased to test takers should be checked. This can be done through the interview with the test developers and a sample of test takers.

The second criterion is disparate impact which is identified using a two-phased procedure. In the first phase, differences among test taker groups should be identified. Group memberships are self-reported by the examinees in the questionnaire. The aim is to test the hypothesis that there is no difference between test taker groups on one or

more variables. The first step is to examine test scores using *t* test or ANOVA to see if there are mean score differences on variables of interest. If there are score differences between groups on variables of interest, the next step is to examine the descriptive statistics of the groups on those variables to find out if the differences are statistically significant.

Thirdly, to determine the disparate impact, differential item functioning (DIF) should be explored. From Classical Test Theory, the Mantel-Haenszel procedure is employed to determine the DIF. Mantel-Haenszel Statistic (MH) is preferred as an alternative to the IRT methods, since as Clauser and Mazor (2005) mentioned, in MH statistic, examinees are typically matched on an observed variable (such as total test score) and then counts of examinees in the focal and reference groups getting the studied item correct or incorrect are compared. MH methods comprise a highly flexible methodology for assessing the degree of association between two categorical variables, whether they are nominal or ordinal, while controlling for other variables.

In the procedure of employing MH statistic, the crosstabulation of frequencies concerning different ethnic groups and gender, the tests of conditional independence, and Kuder-Richardson 20 (KR-20) were examined. Moreover, to explain the internal consistency, point-biserial correlation coefficient was calculated. Since the last two procedures could not be examined in SPSS, the required formulas were codded in R coding program. The detail of the formulas and the step-by-step procedures are reported in the next section to facilitate the interpretation of the results.

In order to explain test access and administration, as the third and fourth modules in TFF, the examinees' answers to the questionnaire items were scored based on Likert's scale and inserted into SPSS. The frequency of the students' agreement and disagreement on each of the components of access and administrational conditions are examined and reported. At last, to consolidate the quantitative data, the contents of the interview with the test developers and test administrators were transcribed and analyzed based on the main themes and contents, to gather information on the way the test is administered and the results are qualitatively reported.

Results and discussions

The test takers' performance on each item (sum, 45 items) was entered into SPSS with the value of 1 for the correct answer and 0 for the incorrect one. All the items assessed test takers' reading comprehension abilities, specifically the reading strategies they had used to comprehend a text. Then, based on the themes and the contents of the items, the items were categorized into 8 groups, each related to one of the reading strategies.

These strategies were seeking purpose (asking for the authors' purpose in the text), determining tone (asking for determining the literary tone of a text), vocabulary knowledge (assessing learners' ability to find the meaning of the words based on the context), making inference (evaluating learners' inference ability), asking for detail (asking for the details stated directly in the text), answering unstated questions (searching for the learners' ability to exclude those parts of information not mentioned in the text), finding pronoun referents (examining their ability to find the reference of the pronouns), and main idea questions (asking for the gist/theme of the text).

The mean of test takers' performance on those items related to each category was entered into SPSS. As it can be seen, a kind of categorizing was done by the researchers

in terms of grouping the items into 8 categories or factors as the first step in determining the test validity.

Determining the validity

As mentioned earlier, to determine the validity of the present test, different statistical procedures tested. For the sake of space, the tables of the rudimentary analysis related to factor analysis are not reported. The means of test takers' performances on each of the abovementioned strategies were calculated. Then two statistical procedures of *t* test and Friedman were employed.

T test was implemented to determine whether each of these components is effective on the overall test performance. To this point, the mean of test takers' performance in each of these components was compared with the half of the overall performances in each component, to infer whether it is higher than half or not. Therefore, *t* test at the level of 0.05 is implemented using Minitab software. As it can be seen in the following table, just in three components, i.e., unstated facts, inference, and pronouns, the mean of scores were higher than half (Tables 1 and 2).

Then, to determine the order of these components, Friedman test was employed. The results showed that pronoun, inference, unstated, vocabulary, detail, main idea, purpose, and tone are respectively the components which gain the highest score from the maximum to the minimum. The low score in the components such as tone and purpose can be due to weak students' performance, or the fact that the test does not cover that component appropriately.

Based on the results it can be concluded that the test can be interpreted as valid in covering the material taught to the students in just the three components. Therefore, the test does not cover all the materials; hence, it is not valid.

Determining the reliability of the test

Besides validity, the internal consistency of the items is checked. The test takers' answers to each item were inserted into SPSS and Cronbach's alpha coefficient was executed to determine the reliability of the scale. As it can be seen in Table 3, the Cronbach alpha coefficient is 0.74 which is acceptable. This shows that the test results are reliable.

Table 1 One-sample *T* test results

Variable	*N*	Mean	St. Dev	SE mean	Upper bound	*T*	*P*
Detail	81	0.4494	0.1499	0.0167	0.4771	− 3.04	0.002
Tone	81	0.1975	0.4006	0.0445	0.2716	− 6.80	0.000
Purpose	81	0.2716	0.4476	0.0497	0.3544	− 4.59	0.000
Unstated	81	0.5136	0.2529	0.0281	0.5603	0.48	0.685
Vocabulary	81	0.4373	0.1956	0.0217	0.4735	− 2.89	0.003
Main idea	81	0.4072	0.3085	0.0343	0.4642	− 2.71	0.004
Pronoun	81	0.5741	0.2694	0.0299	0.6239	2.47	0.992
Inference	81	0.5741	0.3882	0.0431	0.6459	1.72	0.955

Table 2 The order of components

	Mean rank
Detail	4.60
Unstated	5.27
Vocabulary	4.61
Mainidea	4.20
Pronoun	5.67
Inference	5.60
Purpose	3.27
Tone	2.78

Determining the disparate impact

As it was mentioned, Mantel-Haenszel procedure was employed to determine the differential item functioning (DIF). When a test item unfairly favors one group over another, it can be said to be biased. Such items exhibit DIF, a necessary but not a sufficient condition for item bias. According to Abdul Aziz, Mohamad, Shah, and Din (2016), evidence is needed to ensure that inferences made from performance assessments are equally valid for different subgroups in the population.

According to Clauser and Mazor (1993), differential item functioning is present when examinees from different groups have differing probabilities or likelihoods of success on an item, after they have been matched on the ability of interest. In the same line, as Holland and Thayer (1988) mentioned, MH compares the probabilities of a correct response in the focal and reference groups for the examinees of the same abilities. The procedure is implemented by first dividing examinees into levels based on their abilities. In the present study, the total test score, i.e., the sum of all correct responses, is used to match examinees.

As it was mentioned, there were 45 items in the test which was administered to 81 test takers consisting of 29 males and 52 females. Based on the test takers' self-report ethnographic information, they are from three main ethnic groups, i.e., Fars (61 persons), Lor (9 persons), and Tork (11 persons). Since the number of Tork and Lor students are few, Lor and Tork students are considered as one group of non-fars students.

In the first phase, MH statistics was employed to test the independent hypothesis concerning the relationship between gender and total score controlling the ethnicity of the test takers. To this end, the total scores are grouped into focal (with total scores higher than 0 and less than 23) and reference group (with total scores higher and equal to 23 and less or equal to 45). The null hypothesis and alternative hypothesis are defined as follows:

H0: Total score and gender are independent.

H1: Total score and gender are not independent.

Table 4 shows the cross-tabulation of frequencies of the three ethnic groups based on their gender and the total score groups.

Table 3 Reliability statistics of the reading test

Cronbach's alpha	Cronbach's alpha based on standardized items	*N* of items
0.749	0.768	45

Table 4 Cross-tabulation of frequencies of the three ethnic group based on grade, gender, and ethnicity

Ethnicity			Gender		Total
			Male	Female	
Fars	Grade	0	7	29	36
		1	14	11	25
	Total		21	40	61
Other	Grade	0	6	10	16
		1	2	2	4
	Total		8	12	20

As it can be inferred from Table 5, the p value of the MH test is less than 0.05, signifying that there is a correlation between the total score and gender, while the ethnicity variable is controlled. This implies that gender affects test takers' performance on the whole test though their ethnicity is controlled; therefore, differential test functioning is evident. Thus, gender appears to function as a bias in the performance of examinees on the test. This means that the test can function differently for different gender groups.

Besides differential test functioning (DTF), differential item functioning (DIF) must be explored. To this end, the MH test is used to investigate the independent hypothesis between gender and the jth item, $j = 1,\ldots,45$, while the total score is being controlled. The focal and reference groups were organized, as mentioned above.

As depicted in Table 6, p value and chi-square statistic of the MH test for all 45 items are reported. As it can be seen, just in item 33, p value is less than 0.05; thus, the null hypothesis is rejected. Therefore, there is no independence between item 33 and gender when the total score is controlled. Gender can be considered as a bias in item 33 (it is noteworthy that for item 33, p value is equal to 0.04, which weekly rejects the null hypothesis). In all the other items, there is no DIF which means that the items do not function differently for different gender groups.

As it was discussed, except one item, there was no DIF for the other items, but the test functioned differently for the different gender groups. Because of the existence of DTF, it is necessary to elaborate which gender group excelled in performance on the test. Therefore, R programming is employed. The required formulas are written in this coding program, and the following outputs are attained and illustrated in Table 7.

As it can be seen, males performed better on the test than female test takers considering their score mean. The results are also reliable using the Kudar-Richardson 20 (KR-20) formula. This shows the homogeneity or the internal consistency of the test for the two gender groups. The formula for KR-20 for a test with k test items numbered $i = 1$ to K is

Table 5 Tests of conditional independence

	Chi-squared	Df	Asymp. Sig. (2-sided)
Cochran's	8.189	1	0.004
Mantel-Haenszel	6.680	1	0.010

Table 6 MH tests of conditional independence (Gender* Item *j*th* Grade)

Items	Chi-square	Df	Asymp. Sig (2-sided)
Item 1	0.075	1	0.785
Item 2	0.692	1	0.405
Item 3	0.001	1	0.974
Item 4	0.00	1	0.985
Item 5	0.916	1	0.339
Item 6	0.658	1	0.417
Item 7	0.020	1	0.889
Item 8	2.212	1	0.137
Item 9	0.074	1	0.786
Item 10	0.039	1	0.844
Item 11	0.008	1	0.929
Item 12	3.172	1	0.075
Item 13	0.141	1	0.707
Item 14	2.116	1	0.146
Item 15	0.038	1	0.845
Item 16	0.080	1	0.777
Item 17	1.053	1	0.305
Item 18	0.027	1	0.869
Item 19	3.006	1	0.083
Item 20	0.529	1	0.467
Item 21	0.013	1	0.910
Item 22	1.582	1	0.208
Item 23	0.790	1	0.374
Item 24	1.145	1	0.285
Item 25	1.074	1	0.300
Item 26	3.063	1	0.080
Item 27	0.633	1	0.426
Item 28	0.149	1	0.700
Item 29	0.140	1	0.708
Item 30	0.964	1	0.326
Item 31	3.796	1	0.051
Item 32	0.756	1	0.385
Item 33	4.213	1	0.040
Item 34	1.272	1	0.259
Item 35	0.709	1	0.400
Item 36	1.245	1	0.264
Item 37	0.198	1	0.656
Item 38	0.060	1	0.806
Item 39	1.707	1	0.191
Item 40	0.662	1	0.416
Item 41	0.000	1	0.995
Item 42	0.274	1	0.601
Item 43	0.036	1	0.851
Item 44	0.057	1	0.812
Item 45	0.034	1	0.855

Table 7 Summary statistic for the male and female examinees

Group	Mean	Standard deviation	KR-20	Number of examinees
Male	23.7931	7.148323	0.838096	29
Female	19.0961	6.206447	0.7653911	52

$$KR = \frac{K}{K-1}\left[1 - \frac{\sum_{i=1}^{K} p_i q_i}{S_X^2}\right]$$

where p_i is the proportion of correct responses to test item i, q_i is the proportion of incorrect responses to test item i (so that $p_i + q_i = 1$), and

$$S_X^2 = \frac{\sum \left(X_i - \overline{X}\right)^2}{n-1}$$

In this phase of the study, point biserial correlation coefficient is applied. According to Guilford and Fruchter (1973) and Brown (1988), point-biserial correlation coefficient (symbolized as *rpbi*) can be used where the researcher is interested in understanding the degree of relationship between a naturally occurring nominal scale, i.e., gender, and an interval (or ratio) scale. However, language testers most commonly use *rpbi* to calculate the item-total score correlation as another way of estimating item discrimination. Here, the correlation is between a nominal scale (the correct or incorrect answer on each item usually coded as 1 or 0) and the interval scale, which is the total scores on the test.

The goal of this analysis is to estimate how highly each item is correlated with the total scores (the *rpbi* shown in the third column of Table 8). The *rpbi* shows the degree to which each item is separating the better students on the whole test from the weaker students. Thus, the higher the *rpbi*, the better the item is discriminating. Moreover, applying this procedure leads to a shorter, more efficient, norm-referenced version of the test, by selecting those items with the highest point-biserial correlation coefficients from among the whole items. The formula for point biserial correlation coefficient is as the following.

$$r_{pbi} = \frac{M_p - M_q}{S_t}\sqrt{pq}$$

where

M_p stands for whole test means for students answering the item correctly,

M_q refers to the whole test means for students answering item incorrectly,

P is the proportion of students answering correctly,

q is the proportion of students answering incorrectly, and

S_t stands for the standard deviation for the whole test.

This formula is coded in R program and the outputs are obtained and shown in Table 8.

Table 8 shows the p values and point-biserial correlations for the two groups. The range of p values for the males is from 0.000 to 0.897, whereas the range of p values for females is from 0.154 to 0.808.

For male test takers, items 1, 5, 7, 10, 13, 16, 18, 19, 23, 24, 25, 27, 28, 29, 30, 34, 40, and 45 and for females, items 3, 16, 18, 25, 26, 27, 28, 29, 32, and 35, the correlation

Table 8 Proportions (*p*) passing the item and point-biserial correlation (*r*) for the male and female examinees

Item	Male		Female	
	P	*R*	*P*	*R*
1	0.379	0.579	0.212	0.204
2	0.276	0.223	0.154	0.045
3	0.621	0.395	0.462	0.404
4	0.276	− 0.090	0.289	0.147
5	0.690	0.492	0.462	0.244
6	0.862	0.114	0.673	0.236
7	0.724	0.446	0.673	0.036
8	0.690	0.293	0.808	0.179
9	0.517	0.445	0.346	0.397
10	0.379	0.579	0.231	0.607
11	0.241	− 0.029	0.269	0.290
12	0.000	0.000	0.192	− 0.133
13	0.552	0.614	0.365	0.378
14	0.138	0.110	0.308	0.163
15	0.241	− 0.220	0.346	− 0.218
16	0.793	0.605	0.750	0.347
17	0.621	0.226	0.442	0.079
18	0.690	0.731	0.539	0.572
19	0.759	0.423	0.442	0.314
20	0.690	− 0.019	0.539	0.290
21	0.414	0.122	0.365	0.314
22	0.483	0.308	0.577	0.328
23	0.621	0.594	0.346	0.377
24	0.793	0.509	0.538	0.309
25	0.483	0.781	0.423	0.466
26	0.552	0.294	0.212	0.597
27	0.586	0.544	0.577	0.476
28	0.517	0.648	0.269	0.471
29	0.690	0.679	0.442	0.413
30	0.897	0.445	0.692	0.370
31	0.483	0.096	0.289	− 0.071
32	0.621	0.335	0.404	0.439
33	0.345	0.224	0.096	0.173
34	0.897	0.402	0.673	0.278
35	0.655	0.568	0.423	0.434
36	0.310	0.384	0.404	0.251
37	0.552	0.187	0.442	0.172
38	0.379	0.231	0.308	0.157
39	0.586	0.279	0.327	0.285

Table 8 Proportions (*p*) passing the item and point-biserial correlation (*r*) for the male and female examinees *(Continued)*

Item	Male		Female	
	P	*R*	*P*	*R*
40	0.690	0.533	0.673	0.339
41	0.138	− 0.142	0.154	0.028
42	0.172	− 0.447	0.327	− 0.038
43	0.448	0.278	0.423	0.261
44	0.517	0.223	0.500	0.352
45	0.517	0.426	0.539	0.020

between the items and the total score is very high. Therefore, these items appear to be spreading the students out in the same way as the total scores are. In this sense, the point biserial correlation coefficient indicates that these items discriminate well among the students in this group.

For male test takers, the correlation between the items (items 4, 11, 15, 20, 41, and 22) and the total scores, and for the females, the items 12, 15, 31, and 42, has a very high negative value. These items appear to spread the students out opposite to the way the total scores are. In other words, it shows that these items discriminate in a very different way from the total scores. With respect to item 12, the correlation between this item and the total scores is zero; therefore, this item does not discriminate the students in this particular group.

Access

The purpose of policy makers is to ensure balanced educational opportunities for all students. Meanwhile, assessing the progress of all students attempting to gain success in academic arena is of high respect. Therefore, the participation of students with disabilities and students with limited English proficiency, those from different ethnicities speaking different first languages and accent is important.

With regard to the educational access, test takers were asked in the questionnaire whether the materials included in the test covered those contents taught in the course sessions and whether the exam provide enough opportunity to learn those materials. The students declared that the test was accessible to test takers in terms of opportunity to learn the content and to become familiar with the types of tasks and cognitive demands. As it can be seen in Table 9, 68% of those who passed the exam and 32% of those who failed believed that the content coverage of the items was acceptable. Since the number of those who believe in the content coverage or discoverage in both fail

Table 9 The frequency of responses on different access components

	Coverage		Chance		Answering		Format	
	Bad	Good	Bad	Good	Bad	Good	Bad	Good
Fail	66.7%	68%	87.5	62.5	80.5	57.5	94.4	56.5
Pass	33.3%	32%	12.5	37.5	9.5	42.5	6.5	43.5
Total	15	50	16	48	21	45	18	46

Table 10 The frequency of responses on administrational conditions

	Light		Temperature		Facility		Place	
	Bad	Good	Bad	Good	Bad	Good	Bad	Good
Fail	100	66.7	40	71.2	78.8	61.1	50	69.4
Pass	0	33.3	60	28.8	22.2	33.9	50	30.6
Total	2	63	5	59	9	56	2	62

and pass groups is somehow equal, it can be concluded that the conditions were equal for all test takers.

Also, 62.5% of those who fail the exam and 37.5% of those who pass the exam believed that the exam had provided enough chance to learn the materials. Based on the results of the interview with the sample of 20 students, 30% of them believed that although all the points had been included in the test, but the item distribution was not fair. As an example, one of them declared that "there were near 20 questions out of 45 related to 'directly stated questions', but regarding 'determining the tone or the authors' purpose' just one item was included".

Regarding the financial access, since all test takers were the student of state university, they are not required to pay for the exam so there was no financial burden on the students. Therefore, all students have financial access to the test.

Geographical access is somehow equal for all test takers. The test takers were the residents in the dormitories which are located in the campus of the university; therefore, there is the same distance for all students to access the site of examination. Based on the demographic information elicited in the questionnaires, there was no student suffering from a physical disability.

57.5% of those who fail the exam and 42.5% of those who pass the exam declared that they were familiar with the process of taking the exam and answering the questions based on the strategies and instructions they were taught in the course sessions. This point illuminates that most of the students were familiar with the exam procedure; therefore, there is equal condition and educational access for all students.

Administration

As it was mentioned, a fair testing situation must enjoy appropriate physical conditions for test administration such as optimum light, temperature, and facilities. Analyzing test takers' responses to the questionnaire items, it was found that 66.7% of those who passed the exam and 33.3 of those who failed believed, the light of the exam site was appropriate. Also, based on the observation of the exam site by the exam developers and the interview with the exam administrators, this point was ensured.

71.2% of failed students and 28.8% of those who passed believed in an appropriate temperature of the exam conditions. The exam administrators declared that all the site was ventilated with a central cooling system and the temperature of all parts of the building is equal and appropriate. The results are tabulated in the following table (Table 10).

Besides the physical conditions, the uniformity across test sites, equivalent forms, test manuals, and instructions were taken into account. To this end, the students were asked whether the exam site, the format of the questions, the materials being taught to

them, and the condition in which they are placed are different for them in comparison to other test takers.

With regard to the place of examination, as it can be seen in Table 11, 69.4% of failed students and 30.6% of those who passed the exam agreed that the place was acceptable and appropriate, while 67.7% of those who failed the exam and 32.2% of those who passed the exam believe that there was no difference in the format of the question relative to other students. Moreover, 46.8% of the failed 35.2% of those who passed asserted that the physical conditions were equal for all test takers.

Based on the direct observation of test developers and the researcher in the process of item writing, exam development, and administration, the notion of test security was ensured. The number of proctors was equal and sufficient for all test takers accommodated in different classrooms and their seats are randomly numbered to decrease the likelihood of any cheating chance.

Social consequences

To determine the social consequences of the test, washback and remedies must be elaborated. Christopher (2008) defined washback effect as the impact or influence of assessment practices, tests, exams, or any other kind of assessment, on all the individuals involved in the teaching-learning process.

Through the interview with the course instructors, it was concluded that the test does not encourage an examination-oriented approach to teaching and learning. They claimed that they try to teach the materials based on the predetermined curriculum prescribed by the ministry of science, research and technology supervising the state universities in Iran. In this way, the instructors' syllabi were inspected and it was found that the main objectives in the classroom were making learners aware of reading comprehension strategies and no direct mention of the final exams key points was not made.

Moreover, the instructors concentrated on the development of innovative teaching practices, such as the in-depth, critical engagement with texts not just sticking to test preparation purposes such as promoting rote learning of clichéd phrases and strategies to answer the comprehension check questions in the exam session. They asserted that they tried to focus on teaching concepts well and guiding students to apply knowledge and skills instead of excessive drilling and practicing examination strategies.

As it can be inferred, most of the mentioned parts have been based on anecdotal evidence from the interviews and personal observations; however, more research is necessary to further investigate this subject. As Loh and Shih (2016) pointed out, every test requires tailor-made research to examine its washback.

Table 11 The frequency of responses on administrational conditions

	Format difference		Condition difference	
	Bad	Good	Bad	Good
Fail	100	67.7	80	64.8
Pass	0	32.3	20	35.2
Total	1	62	10	54

The other main area of inquiry for social consequences is remedies regarding the harmful effects of a test. This refers to remedies offered to test takers to "reverse the detrimental consequences of a test such as re-scoring and re-evaluation of test responses, and legal remedies for high-stakes tests." (Kunnan, 2004, p. 39). The key fairness questions here are whether the social consequences of a test and/or the testing practices are able to contribute to societal equity.

According to Loh and Shih (2016), the notion of remedies required the test administrators to ensure the examinees that they have been fairly assessed and to meet the needs of test takers who encounter any unforeseen circumstances such as illness during the test (p. 164). To this end, several remedies have been applied to safeguard this aspect of test fairness.

A second exam was predicted with the same format and range of test difficulty for those who could not attend the exam session due to the acceptable excuses such those who are medically certified as being too ill to take the test. Also, there was another possibility to use the mean of their class activities and quiz scores as their final examination score. Moreover, the examinees had the opportunities to review their exam sheets a week after their exam session.

All in all, based on the analysis done employing the five modules of Kunnan's TFF, it was seen that the mentioned reading comprehension test must be enhanced in terms of its validity and the items which made the test function differently considering the different gender groups must be amended. With respect to other modules, this test was fair enjoying and acceptable level of access, while it was administered appropriately under satisfying conditions. Moreover, the social consequences were taken into consideration in terms the effect of washback and the predicted remedies.

The results of the study are in line with the results gleaned in Loh and Shih's (2016) study, although the present study is peculiar in terms of employing a step by step statistical procedure in two-phased research design using both qualitative and quantitative data collection and analysis. Loh and Shih (2016) implemented an evidence-based approach toward test fairness. Moreover, the other studies mentioned previously, e.g., Khoii and Shamsi (2012) and Pishghadam and Tabataba'ian (2011), tried to explore the effect of other factors such as test format on test fairness; however, the present study considered the examination of the notion of fairness using TFF.

Conclusion

In the present study, an attempt was made to determine the test fairness of a general English reading comprehension test at Fasa University based on Kunnan's test fairness framework (TFF). Checking the modules of TFF, it was concluded that the test must be enhanced in terms of validity, although the reliability of the test was acceptable by checking the internal consistency of the items. The Mantel-Haenszel procedure was employed to determine the differential test functioning (DTF). For this, it was concluded that gender affects test takers' performance on the whole test, although their ethnicity was controlled; therefore, differential test functioning was evident. This means that gender appears to function as a bias in the performance of examinees on the test. To amend the test toward fairness, the DIF for the *j*th item was determined using the Mantel-Haenszel procedure for each item. As it was discussed, except for one item, there was no DIF for the other items. In this way, redrafting the biased item, the

fairness of the test can be improved. With regard to geographical, financial, and educational access, as well as physical conditions for test administration and social consequences, the examined test represents an acceptable level of fairness.

With respect to the practical implications of the current study, as it was mentioned, the step-by-step statistical procedure employed in this study can be a sound model for the test developers to ensure the fairness of their designed tests by examining TFF throughout the test development and administration processes. Although some researchers have tried to implement Kunnan's TFF in an evidence-based approach, the mixed qualitative and quantitative design of the present study can be a stimulus to be replicated on other large-scale examinations with a more populated, diverse test takers in terms of ethnicity, age range, and other demographic features.

Acknowledgements
Hereby, I would like to acknowledge the educational staff at Fasa University who provide permitted conducting a research on the general English reading test given to their students every year. Also, the esteemed participants and the official staffs who permitted us to interview them are highly acknowledged.

Authors' contributions
The author listed has contributed sufficiently to the project to be included as author. The first author dealt with the theoretical issues, surveying the related studies, and writing and editing the manuscript, while the second author contributed in the analysis section, statistical procedures, and interpretation. The authors read and approved the final manuscript.

Authors' information
Dr. Meisam Moghadam has earned his Ph.D. in TEFL from Shiraz University and is currently the assistant professor at Fasa University, Iran. His areas of interest include language teaching and testing, teacher education, and psycholinguistics.
Dr. Fariba Nasirzadeh, Ph.D. was graduated from Shiraz University majored in statistics. He is now a lecturer at Jahrom University, Iran. His current areas of interest mostly involve spatial functional data analysis, data science, data mining, and time series.

Funding
This study was not supported by any funding.

Availability of data and materials
The required data for analysis are tabulated throughout the manuscript. The exam sheets and the questionnaires were kept confidential since a consent was confirmed throughout the study and interviews to the participants to keep their information confidential.

Competing interests
To the best of our knowledge, no conflict of interest, financial, or other, exists. The authors declare that they have no competing interests.

Author details
[1]Faculty of Science, Fasa University, Fasa, Iran. [2]Department of Statistics, Jahrom University, Jahrom, Iran.

Received: 9 April 2020 Accepted: 6 July 2020
Published online: 15 July 2020

References

Abdul Aziz, J., Mohamad, M., Shah, P. M., & Din, R. (2016). Differential item functioning in online learning instrument. *Creative Education*, *7*, 180–188.
Baharloo, A. (2013). Test fairness in traditional and dynamic assessment. *Theory and Practice in Language Studies*, *3*(10), 1930–1938.
Christopher, N. M. (2008). Social and educational impact of language assessment in Nigeria. *Nordic Journal of African Studies*, *17*(3), 198–210.
Clauser, B., & Mazor, K. M. (2005). Using statistical procedures to identify differential item functioning test items. *Educational Measurement Issues and Practice*, *17*(1), 31–44.
Cole, N. S., & Zieky, M. J. (2001). The new faces of fairness. *Journal of educational Measurement*, *38*(4), 362–384.
Davies, A. (2010). Test fairness: A response. *Language Testing*, *27*(2), 171–176.
Flaugher, R. L. (1973). *The new definitions of test fairness in selection: Developments and Implications, research memorandum*. Princeton, New Jersey: Educational testing Service.
Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. New York: Routledge.
Hamid, M. O., Hardy, I., & Reyes, V. (2015). Test-takers' perspectives on a global test of English: Questions of fairness, justice and validity. *Language testing in Asia*, *9*(16), 1–20.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (p. 129–145). Lawrence Erlbaum Associates, Inc.
Kane, M. (2010). Validity and fairness. *Language Testing*, *27*(2), 177–182.
Khoii, R. & Shamsi, N. (2012). A fairness issue: Test method facet and the validity of grammar subtests of high-stakes admissions tests. *Literacy Information and Computer Education Journal (LICEJ)*, Special Issue, *1*(1).
Kunnan, A. J. (1997). Connecting validation and fairness in language testing. In A. Huhta et al. (Eds.), *Current developments and alternatives in language assessment*, (pp. 85–105). Jyvaskyla, Finland: University of Jyvaskya.
Kunnan, A. J. (2000). Fairness and justice for all. In A. J. Kunnan (Ed.), *Fairness and validation in language assessment*, (pp. 1–13). Cambridge: CUP.
Kunnan, A. J. (2004). Test fairness. In M. Milanovic, & C. Weir (Eds.), *European language testing in a global context*, (pp. 27–48). Cambridge: CUP.
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. Papers from the ALTE Conference in Berlin, Germany*, (pp. 229–251). Cambridge: CUP.
Kunnan, A. J. (2010). Test fairness and Toulmin's argument structure. *Language Testing*, *27*(2), 183–189.
Loh, T. L., & Shih, C. M. (2016). The English language test of the Singapore Primary School Leaving Examination. *Language Assessment Quarterly*, *13*(2), 156–166.
Pishghadam, R., & Tabataba'ian, M. S. (2011). IQ and test format: A study into test fairness. *Iranian Journal of Language Testing*, *1*(1), 17–29.
Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, *18*(4), 373–391.
Stobart, G. (2005). Fairness in multicultural assessment systems. *Assessment in Education.*, *12*(3), 275–287.
Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. London: Palgrave.
Willingham, W. W., & Cole, N. (1997). *Gender and Fair Assessment*. Mahwah, NJ: Lawrence Erlbaum.
Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*(2), 147–170.
Zheng, Y., & Cheng, L. (2008). Test review: College English test (CET) in China. *Language Testing*, *25*(3), 408–417.

Publisher's Note