frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 29 January 2025
DOI 10.3389/feduc.2025.1518315

OPEN ACCESS

EDITED BY
Xiaomin Zhu,
Changshu Institute of Technology, China

REVIEWED BY
Semirhan Gökçe,
Ömer Halisdemir University, Türkiye
Meenakshi Sharma Yadav,
King Khalid University, Saudi Arabia

*CORRESPONDENCE
Mengting Kong
✉ rachel3979@163.com

RECEIVED 28 October 2024
ACCEPTED 16 January 2025
PUBLISHED 29 January 2025

CITATION
Li K, Kong M, Li L and Lu J (2025) Test fairness of the in-house College English examination for Chinese non-English major undergraduates: a case study.
Front. Educ. 10:1518315.
doi: 10.3389/feduc.2025.1518315

Test fairness of the in-house College English examination for Chinese non-English major undergraduates: a case study

Kangxi Li, Mengting Kong*, Liye Li and Jingwen Lu

School of Foreign Studies, Hefei University of Technology, Hefei, China

Introduction: The in-house College English examination is a high-stakes assessment for College English, a compulsory course for most non-English major undergraduates in China. However, the fairness of such examinations has not received due attention.

Methods: This paper employs Kunnan's Test Fairness Framework to investigate common issues of test fairness related to the in-house College English examination, using a university in East China as an example. The data analyzed include the College English examination results of 5,680 non-English major undergraduates from 2018 to 2020.

Result: The research finds various fairness problems in terms of validity, absence of bias, access, administration, and social consequences. These issues can be attributed to underlying factors such as lack of language assessment literacy, time, and funding on the part of examiners or administrators.

Discussion: Hence, this research proposes a holistic approach to improve test fairness, involving all stakeholders and all procedures in the in-house College English examination. Collaborating with external experts on language tests and lowering the stakes of in-house examinations are suggested as effective measures to mitigate test unfairness issues.

KEYWORDS

test fairness, in-house College English examination, case study, Kunnan's Test Fairness Framework, longitudinal study

Introduction

In-house College English examinations are standardized tests for the course of College English administered by colleges or universities themselves, which cater to their specific talent cultivation objectives and the needs of the students. In view of the stark differences between various higher learning institutions, it is of great necessity to develop a school-based College English examination system that complements the nationwide English proficiency tests such as CET-4 and CET-6, and those international examinations such as TOEFL, IELTS and GRE. The in-house examinations, including placement tests prior to College English instructions and summative tests for each semester, can help evaluate students' English proficiency more accurately and provide feedback for teachers as well as administrative bodies in the college. They also play a decisive role in GPA, scholarship, further study and even job seeking for college students.

In China, top universities like Tsinghua University, Shanghai Jiao Tong University, and Zhejiang University have issued their own in-house college English proficiency examination syllabuses. Similarly, many other universities have been implementing their own in-house college English examinations for years. A recent survey by Sun et al. (2020) across 98 universities in different regions of China, ranging from national key universities to

provincial and municipal key universities, and ordinary universities, revealed that the majority (96.9%) of them have both summative and placement examinations, with national key universities slightly outpacing other categories.

However, according to Jin (2015) research, it is evident that among various methods of English proficiency testing, teachers and students generally prefer the national unified examinations designed and implemented by professional institutions, which enjoy a high level of support. In contrast, tests developed independently by universities receive relatively lower approval. Furthermore, another survey reveals that the development and implementation of these exams still face certain difficulties, including a lack of time and funding for exam development and administration, as well as a shortage of teachers with the necessary knowledge and skills in testing (Jin, 2020).

Despite the ubiquity of in-house college English examinations in China and their significance for students' academic and career trajectories, research on the fairness of in-house college English examinations, especially empirical studies, remains scarce. The few representative studies, such as those by Fan and Ji (2013), Jia et al. (2013), and Guo and Lin (2016), have been limited to the facets of reliability and validity, thus leaving much to be explored.

Therefore, to address this research gap, the present study aims to conduct a longitudinal investigation of the fairness of in-house college English examinations in a state key university in East China. Drawing on Kunnan's Test Fairness Framework, this study seeks to identify the major unfairness issues, explore their underlying causes, and propose practical countermeasures.

Literature review

Test fairness

Test fairness has begun to capture the attention of the academic circle as early as the 1960s when it was first studied as item bias (Angoff, 1993). The persisting nature of this issue can be seen in the fact that the issue of test fairness or justice has been repeatedly revisited by the Language Testing Research Colloquium (LTRC) throughout the past two decades, particularly in 1997 and 2019.

Up till now, though various interpretations have been proposed with regard to the three interrelated concepts, i.e., fairness, justice and validity (Davies, 2010; Kane, 2010; Kunnan, 2000; Kunnan, 2004; Kunnan, 2008; Kunnan, 2014; Roever and McNamara, 2006; McNamara and Ryan, 2011), there is still no consensus on the definition of test fairness among researchers and test developers. Generally speaking, fairness has been conceptualized either as an independent test quality, as all-encompassing, or as directly linked to validity (Xiaoming, 2010).

In addition to the differences in definitions, people also seem to approach test fairness from different perspectives, ranging from sociology (Camilli, 2006; Jensen, 2006), standards or norms (Kane, 2010) and stakeholders (Brown, 1996; Hamp-Lyons, 1997).

From the sociological perspective, test fairness is essentially an all-encompassing concept influenced by diverse factors such as test development, implementation, interpretation and use of tests scores, which have legal, ethical, political as well as economic consequences.

From the perspective of standards or norms, test fairness mainly focuses on whether the procedures of a test meet the required specifications. According to *Standards for Educational and Psychological Testing* (AERA, 2014), test fairness comprises a series of sub-standards for test design, development, administration, and scoring procedures, test score interpretations, etc.

From the perspective of stakeholders, the interests of all parties concerned should be factored in to ensure the fairness of a test. Teachers should design the form and items of tests that are suitable for specific test takers, who can fully demonstrate their abilities regardless of learning backgrounds and cognitive styles. Also, they should keep the students' parents informed of the test results. Only when the test developers, users, teachers and examinees interact with each other in a positive manner can the test fairness be attained (Fan, 2014).

Although some test fairness models have been proposed, few of them lend themselves easily to empirical investigations. A well established and widely used theoretic model is the Test Fairness Framework (TFF) formulated by Kunnan (2004) and Kunnan (2008), which comprises five major qualities: validity, absence of bias, access, administration, and social consequences (Kunnan, 2004). To elaborate, validity ensures that a test accurately measures what it claims to measure, providing meaningful and appropriate interpretations for its intended use. Absence of bias means that no group is unfairly favored or disadvantaged based on irrelevant characteristics such as gender or ethnicity, ensuring comparable outcomes for all individuals with similar abilities. Access guarantees that all potential test-takers have equal opportunities to prepare for and take the test, including accommodations for those with disabilities. Administration focuses on delivering the test consistently and fairly through standardized procedures and qualified administrators. Social consequences consider the broader impact of the test on society and individuals, including both intended and unintended effects on policies, teaching practices, and personal opportunities.

This framework evaluates test fairness in terms of the whole system of a testing practice rather than the test itself (Kunnan, 2004; Kunnan, 2005), which implies that multiple factors are involved in the issue, including test uses (either intended or unintended), stakeholders (test-takers, test users, teachers, and employers, et al.), and test development (test design, development, administration, and use).

Kunnan's TFF has broadened the scope of test fairness, and hence has been adopted by many academic researches. For large-scale assessments like the IELTS, Hamid et al. (2019) found that many test-takers questioned the test's ability to accurately measure their language skills and were skeptical about its fairness in relation to the potential uses for income increase or immigration. With regard to small-scale assessments, Moghadam and Nasirzadeh (2020) investigated the fairness of a locally developed English proficiency test and found that the test was acceptably fair from the perspectives of access, administration, and social consequences. Wallace and Qin (2024) concluded that test-takers perceived the testing procedures and their interactions with teachers during the test as fair, aligning with Kunnan's framework of test fairness and justice.

When it comes to in-house College English examinations, despite their omnipresence, there are insufficient researches, especially empirical ones, on their fairness. The few representative

researches to date include Fan and Ji (2013), Jia et al. (2013), and Guo and Lin (2016), with the scope of discussions limited to reliability and validity only.

Research questions

In view of the universality and significance of in-house college English examinations, and the general lack of empirical researches on their fairness, this study keeps track of the in-house college English examinations from a state key polytechnic university in East China, spanning four semesters between 2018 and 2020, and analyzes the issue of test fairness, its cause and effect, and how to tackle it, using Kunnan's TFF theory as its theoretic framework. Specifically, the study addresses the following research questions:

1 Are there some significant unfair practices in the in-house college English examinations of this university, which appear to be universal among other higher-learning institutions?
2 If so, what are the underlying reasons for the lack of fairness in these tests? And what effective measures may be taken to ensure the fairness of in-house college English tests, so as to produce a positive wash-back effect on English teaching and learning?

Methods

Participants

The research subjects of this longitudinal study are 5,680 non-English major undergraduates admitted to a national key polytechnic university in East China in 2018. The recruitment period for this study commenced on September 1, 2018, and concluded on July 1, 2020. During their freshman to sophomore years, all of the subjects would take the course in College English, a compulsory course which is divided into four sections (hereinafter referred to as CE-1 to CE-4 respectively). Each section has a corresponding terminal examination. Those who fail in the examinations need to take the make-up test, or to retake the course instead. Only those students who are taking CE-3 and above are qualified to take CET-4, which is an important nationwide English proficiency test administered by the National Education Examinations Authority (NEEA) in China, and only those who score over 425 are allowed to take CET-6. The overall scheme of College English teaching and examinations is as follows:

It can be seen from Figure 1 that freshmen are divided into two levels after the placement test. The level 2 students, having been waived CE-1, need to take 3 semesters of College English courses only, while the level 1 students must take the entire series of College English sequentially spanning four semesters. In other words, all students must take the placement test, but there are only three in-house examinations for the level 2 students, and four examinations for the level 1 students. This study kept records on all of the in-house College English examinations undertaken by students of both levels, using their scores in the National College Entrance Examination (NCEE), CET-4 and CET-6, if available, as the reference.

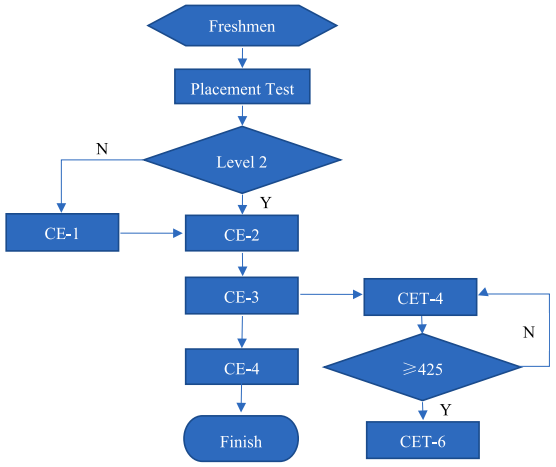

FIGURE 1
The college English curriculum map.

Data collection

Participants in this study included 5,680 students from various majors across the university. The data collection period spanned from September 1, 2018, to July 1, 2020, capturing a comprehensive snapshot of student performance over nearly two academic years. To accommodate different preferences and ensure broad participation, both online and paper-based testing formats were utilized. The sample was carefully selected to represent a diverse range of disciplines, including humanities, sciences, engineering, and business, ensuring that the findings are generalizable across different fields of study. Participants ranged from freshmen to seniors, allowing for insights into how test performance evolves throughout a student's academic journey.

Qualitative data

To deepen our analysis and provide a richer understanding of the test experience, we conducted interviews with a subset of the 5,680 students and teachers involved in the study. This qualitative approach allowed us to gather nuanced insights into their perceptions and experiences, complementing the quantitative data collected. By integrating qualitative data, we gained deeper insights into the practical implications of the test, highlighting areas for improvement and reinforcing the overall robustness of our research findings.

Enrollment analysis

As a state key university, its student body come from different regions with different educational backgrounds. According to statistics, the top five sources of undergraduate students in 2018 are Anhui Province (1501), Henan Province (308), Shandong Province (301), Hebei Province (291) and Jiangsu Province (271), adding up to 47.04% of the overall enrollment of undergraduate program. Conversely, the last five provinces or regions account for only 3.45% in total. It can

be seen that the number of students is negatively correlated with the distance between their homes and the province where the university is located. The vast majority of students are concentrated in the local province and the adjacent areas, especially in Central and East China.

Faced with disparities in educational and English proficiency levels, it is crucial to prioritize enrollment analysis. Enrollment analysis serves not only to provide deeper insights into the backgrounds and advantages of students in diverse regions but also aids in the adaptation of education policies and curriculum designs to cater to the unique needs and characteristics of students in various areas. Consequently, emphasizing the importance of enrollment analysis and research not only heightens the relevancy and practicality of studies but also furnishes valuable insights and backing for regional educational advancement.

Provincial autonomous test papers chiefly originate from regions with relatively robust economies and ample educational resources in China, indicating that students in these areas may generally possess higher English proficiency levels compared to those in western and inland regions. The primary intent of autonomous test papers was to adapt to the varying educational standards across regions, permitting modifications in test content and difficulty levels according to local characteristics and circumstances, thereby galvanizing provincial education management and development initiatives.

Nevertheless, concerns persist regarding autonomous test papers. Local test paper teams often fall short in comparison to national standardized test papers, leading to issues related to discriminatory power, reliability, validity, and test question difficulty, potentially resulting in inconsistent test paper quality among provinces and undermining the fairness and comparability of test outcomes to some extent. Therefore, excluding autonomous test paper areas from research aims to ensure the rigor and objectivity of the research. By eliminating the variables associated with autonomous test paper regions, research can more precisely evaluate disparities in regional education and candidate capabilities, yielding more insightful data and findings essential for devising effective education policies.

It should be pointed out that there is a stark difference between the English proficiency of students from different regions in China. In light of the fact, three nationwide test papers for NCEE (Paper I, II, and III in Table 1) are adopted by the Ministry of Education of China, with the tests sharing the same syllabus and structure, but in varied difficulty levels. Besides, another five test papers are developed by the provincial or municipal educational authorities in Jiangsu Province, Zhejiang Province, Shanghai, Tianjin and Beijing, respectively. These five test papers are tailored for students in these regions because, on average, these students tend to have a significantly higher level of English proficiency than students from other areas.

In the following analyses where NCEE scores are involved, we would exclude the students from the last 5 regions in Table 1, since their examinations are not comparable, but paper I, II, and III, due to their proximity by nature, are considered as equivalent to each other.

Test development

The in-house test items are subjectively selected by examiners from a pre-designed question bank, aiming to maintain a basic consistency in difficulty and question type with the CET-4 exam. However, due to varying levels of expertise among examiners, the selection process often relies on personal judgment. To ensure quality, examiners consult with department heads and curriculum leaders, such as the department chair and the head of the teaching research office, to adjust and control the difficulty of selected questions based on specific circumstances. The in-house test question types at this institution are gradually transitioning from the college entrance exam format towards the CET-4/6 format, with the ultimate goal of aligning or closely matching the question types and scoring ratios of CET-4/6.

Results

Validity

Reliability is a necessary condition for validity, which shall be analyzed prior to further discussions. Table 2 shows the Cronbach's alpha values of the in-house College English examinations. Please note that this table does not include the reliability of CE-1, and in order to eliminate the impact of raters on reliability, the subjective items (translation and writing) in each examination are excluded from the scores. It can be seen that all of the in-house examinations have Cronbach's alpha values well above 0.600, which indicates that those examinations are internally consistent and hence reliable.

Content validity

The content validity of the in-house examinations is evaluated with reference to external examinations such as Paper I, II, III of NCEE, CET-4, and CET-6. The composition of each in-house test paper is shown in Table 3.

Table 3 shows that, both the question types and the percentage of scores for each part in the in-house examinations have shifted gradually over time, forming a continuum between NCEE and CET-4/6. Specifically, Vocabulary and Structure, a frequently used question type in NCEE, still lingers in the placement test and CE-1 test. However, in the subsequent examinations like CE-2/3/4, the composition of each test paper has been revised to be identical to that of CET-4/6, except for the fact that the score ratios have been adjusted a bit, with less weight given to listening and more weight to Reading and Writing. Since CET-4/6 are often used as one of the primary indicators for assessing the quality of College English teaching, it is

TABLE 1 Types of English test papers in NCEE for 2018 non-English major undergraduates.

	Paper I	Paper II	Paper III	Beijing	Shanghai	Tianjin	Jiangsu	Zhejiang
Number	3,427	934	564	26	22	111	271	218
Mean	126.5	123.06	122.39	123.77	112.95	123.3	96.01	126.4
Std. dev	10.39	12.82	17.19	6.72	8.16	7.34	5.34	9.47

TABLE 2 Reliability of each in-house College English examinations.

Examination	Placement	CE-1	CE-2	CE-3	CE-4	CE-2	CE-3	CE-4
Level	All Levels	1	1	1	1	2	2	2
Cronbach's α	0.643	0.704	0.752	0.719	0.699	0.715	0.657	0.719

TABLE 3 Composition of the in-house examinations in comparison with external examinations.

Question type	Listening	Reading	Writing	Translation	Language use
NCEE	20%	26.7%	23.3%	N/A	30%
Placement	40%	40%	N/A	N/A	20%
CE-1	25%	40%	N/A	15%	20%
CE-2	25%	40%	20%	15%	N/A
CE-3	25%	40%	20%	15%	N/A
CE-4	25%	40%	20%	15%	N/A
CET-4/6	35%	35%	15%	15%	N/A

Language use refers to Cloze, Vocabulary and Structure, Error Correction, etc., which are categorized as Language Use in NCEE test papers.

TABLE 4 Criterion-related validity of the in-house examinations.

		Correlations			
		Placement	NCEE	CET-4	CET-6
Placement	Pearson Correlation	1	0.603**	0.673**	0.565**
	Sig. (2-tailed)		0.000	0.000	0.000
	N	4,702	4,702	4,590	3,649
NCEE	Pearson Correlation	0.603**	1	0.592**	0.469**
	Sig. (2-tailed)	0.000		0.000	0.000
	N	4,702	4,702	4,590	3,649

Correlation is significant at the 0.01 level (2-tailed).

reasonable for this university to calibrate the in-house examination constructs against the CET-4/6 standards.

However, Table 3 also shows that the rule has not been followed stringently and consistently in the in-house examinations. The most noticeable of all is the placement test, which only examines three aspects including Listening (40%), Vocabulary and Structure (20%), and Reading (20%). Those important language abilities like writing and translation are not examined in the test paper.

The study also finds that the duration of the College English placement test is only 90 min, which is significantly shorter than that of CET-4/6 (125 min each). This inevitably leads to a reduced number of questions (60 multiple-choice items only, and no subjective items included), another significant factor influencing test reliability and validity.

Criterion-related validity

As a longitudinal study, we are able to compare the in-house examinations with the external examinations that are taken prior to, during or after CE-1/4 as the criteria for validity. Since there is little time difference between NCEE and the placement test, the former can be used to evaluate the concurrent validity of the placement test. Meanwhile, CET-4/6 are usually taken by students of CE-3 and CE-4, thus they are taken as the criteria for the predictive validity for the in-house examinations.

Table 4 shows the Pearson correlation between the in-house examinations and external ones. All students (4702) who took the NCEE Paper I, II, III and the placement test are counted in. The main findings are as follows:

1 There is a moderate or strong correlation between the placement test and external examinations ($r_{max} = 0.673$, $r_{min} = 0.565$), indicating that the placement test is concurrently and predicatively valid.
2 The correlations between the placement test and CET-4/6 ($r_1 = 0.673$, $r_2 = 0.565$) are only slightly higher than those between NCEE and CET-4/6 ($r_1 = 0.592$, $r_2 = 0.469$), which implies that NCEE can and should complement with the placement test in categorizing students into different English proficiency levels for subsequent College English courses.
3 Compared with its strong correlation with CET-4 ($r = 0.673$), the placement test still need to be further improved in order to be correlated with CET-6 ($r = 0.565$).

The analysis above shows that the placement test is a valid means of distinguishing students for hierarchical teaching in the future, yet

there is still much room for further improvement in the content and length of the placement test.

After the placement test, students are subsequently divided into two levels, with 3,949 students in level 1 to take the course in CE-1 and 1,186 students in level 2 to take CE-2. Tables 5, 6 list the correlation between the final examinations of the two levels and CET-4/6.

By comparing Tables 5, 6, it can be seen that:

1 There are moderate or strong correlations between the final examinations and CET-4/6, which are the predictive criteria for the former in this research, while the Pearson correlation values fluctuate dramatically somewhere between 0.440 and 0.687.
2 Most of the final examinations, with the exception of CE-3 for level 2 students, consistently show a stronger correlation with CET-4 than with CET-6. This pattern is more noticeable for level 1 students than for level 2 students.

The correlation analysis shows that the validity of the final examinations for level 1 students is significantly higher than those for level 2 students. The reason for this difference may be that the difficulty level of the final examinations in this university is often gauged against CET-4. However, for level 2 students, who usually pass CET-4 with ease (98.40%), the final examinations seem too simple to discriminate between their language ability.

To sum up, all in-house examinations meet the basic validity requirements, having consistent reliabilities, moderate or strong correlations with external examinations like NCEE and CET-4/6. Still, there are some issues to be solved with respect to the content and difficulty level. On one hand, some in-house examinations, especially the placement test and the final examination for CE-1, deviate significantly from the standards of CET-4/6. On the other hand, most of the final examinations are intended to be close to CET-4 in terms of difficulty, and thus students in level 2 find them too easy to be valid enough.

Absence of bias

Another criterion of TFF is the absence of bias, which requires that, the content, language and standard of the test, etc. shall not be biased

TABLE 5 Correlation between final examinations for Level 1 students and CET-4/6.

		Correlations					
		CE-1	CE-2	CE-3	CE-4	CET-4	CET-6
CE-1	Pearson Correlation	1	0.740**	0.687**	0.650**	0.687**	0.504**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000	0.000
	N	3,949	3,942	3,922	3,941	3,704	2,102
CE-2	Pearson Correlation	0.740**	1	0.688**	0.659**	0.660**	0.488**
	Sig. (2-tailed)	0.000		0.000	0.000	0.000	0.000
	N	3,942	3,943	3,916	3,935	3,699	2097
CE-3	Pearson Correlation	0.687**	0.688**	1	0.655**	0.653**	0.507**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000
	N	3,922	3,916	3,923	3,915	3,682	2086
CE-4	Pearson Correlation	0.650**	0.659**	0.655**	1	0.646**	0.558**
	Sig. (2-tailed)	0.000	0.000	0.000		0.000	0.000
	N	3,941	3,935	3,915	3,942	3,697	2099

TABLE 6 Correlation between final examinations for Level 2 students and CET-4/6.

		Correlations				
		CE-2	CE-3	CE-4	CET-4	CET-6
CE-2	Pearson Correlation	1	0.564**	0.936**	0.465**	0.441**
	Sig. (2-tailed)		0.000	0.000	0.000	0.000
	N	1,186	1,186	1,186	1,186	1,078
CE-3	Pearson Correlation	0.564**	1	0.494**	0.458**	0.483**
	Sig. (2-tailed)	0.000		0.000	0.000	0.000
	N	1,186	1,186	1,186	1,186	1,078
CE-4	Pearson Correlation	0.936**	0.494**	1	0.440**	0.405**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000
	N	1,186	1,186	1,186	1,186	1,078

Correlation is significant at the 0.01 level (2-tailed).

against any group of test takers. Specifically, the test must not contain offensive or biased content, language or dialect for testees from different backgrounds, like gender, race and ethnicity, religion, age, mother tongue, nationality, so as to ensure that the difference in scores is due to the measured language ability rather than the factors mentioned above.

In the present study, we find little, if any, evidence of bias in the test papers. However, this does not mean that the entire hierarchical teaching mode, which is established on the basis of the in-house examinations, is not without bias. In fact, such hierarchical teaching turns out to be unfriendly towards some subgroups of students, particularly the male and the ethnic minorities.

Gender differences

This paper analyzes the relationship between gender and English performance of 4,749 students who took NCEE Paper I, II, or III. The results show that the distributions of NCEE total scores, and the scores of NCEE English and placement test differ remarkably between male and female students (the significance level $p < 0.05$).

It can be seen from Table 7 that, similar to most polytechnic universities in China, the proportion of male students is significantly higher than that of female students. In the NMT, the average total score of male students who are admitted to this university is significantly higher than that of female students by a margin of 9.12 points (the total score is 750), but they are inferior to females in NCEE by 3.09 points (the total score is 150). After admission to the university, however, the gap in English has been further consolidated, or even amplified to a much greater extent of 3.63 points in the placement test (the total score is 100).

We further analyze the gender imbalance by using the Pearson Chi-Square test (χ^2). Table 8 shows that only 16.5% of male students are assigned to level 2, while the proportion for female students has risen to 28.6%. The χ^2 hypothesis test demonstrates that the students' placement results is not independent of the gender factor ($p < 0.05$) and that the correlation is quite weak (Cramers' V = 0.130).

After the placement test, a perennial difference between male and female students can also be observed in the in-house and external examinations using independent *t*-test (see Tables 9, 10). Whether in level 1 or level 2, female students have maintained a significant edge over male students in average scores. For level 1 students, the average of male students in the final examinations is consistently 1–2 points lower than those of female students. For level 2 students, the gender gap continues to exist. With regard to external examinations, the differences in CET4/6 average scores are 23.259/16.955, while for level 1 students, and 15.5580/28.3041 for level 2 students.

Ethnic differences

In the present research, we only take Tibetan and Uygur students as an example. A total of 71 students are singled out to investigate the impact of ethnicity on examination outcomes (see Table 11).

The above table shows that the students of Tibetan or Uygur origin are much inferior to other ethnic groups when it comes to performance on English tests. The average score in NCEE is about 50 points lower than that of students from other ethnic origins. Besides, all of them are placed in level 1, and have only slim chances of passing the final examinations for CE-1/2/3/4. Worst of all, only a fraction of those students (12.68%) finally scrape through CET-4, and none of them stand a chance of passing CET-6 with a score over 425.

Overall, it can be concluded that the hierarchical teaching mode in this university is unfavorable for ethnic minority students, who are classified into a group (level 1 in this case) that are ill-suited for their English foundation and as a result fail most of the in-house and external examinations.

As discussed in the introduction, test fairness depends not only on the design of the exam content but also on students' access to the exam and the impartiality of its administration. By delving into opportunities for participation, management processes, and social impacts, we can better understand how these factors interact to create a more equitable testing environment.

Access

The concept of access refers to equality in educational, financial, geographic and personal access, as well as familiarity with equipment and condition of the test (Kunnan, 2004). In this case study, all the in-house examinations are available to the students free of charge. Even if they fail and hence have to retake one of the courses or the make-up examinations, no additional fees are incurred. Also, to ensure the equity of access, the school authorities provide the final examinations for students of the same level on all campuses simultaneously. The school radio broadcasting station will play the listening materials for College English days before the test to familiarize them with the examination procedures.

TABLE 7 NMT, NCEE, and placement test statistics by gender.

	Group statistics				
	Gender	N	Mean	Std. deviation	Std. error mean
NMT	Male	3,670	595.73	28.498	0.470
	Female	1,079	586.61	53.972	1.643
NCEE	Male	3,670	125.01	10.052	0.166
	Female	1,079	128.10	13.601	0.414
Placement Test	Male	3,540	60.01	11.882	0.200
	Female	1,033	63.64	12.650	0.394

TABLE 8 Chi-square test of students' gender and placement results.

			Gender * Placement Crosstabulation		
			Placement results		Total
			Level 2	Level 1	
Gender	Male	Count	604	3,066	3,670
		% within gender	16.5%	83.5%	100.0%
	Female	Count	309	770	1,079
		% within gender	28.6%	71.4%	100.0%
Total		Count	913	3,836	4,749
		% within gender	19.2%	80.8%	100.0%

	Chi-square tests				
	Value	Df	Asymptotic significance (2-sided)	Exact sig. (2-sided)	Exact sig. (1-sided)
Pearson chi-square	79.657	1	0.000		
Continuity correction	78.875	1	0.000		
Likelihood ratio	74.293	1	0.000		
Fisher's exact test				0.000	0.000
N of valid cases	4,749				

		Symmetric measures	
		Value	Approximate significance
Nominal by nominal	Phi	−0.130	0.000
	Cramer's V	0.130	0.000
N of valid cases		4,749	

	Group statistics				
	Gender	N	Mean	Std. deviation	Std. error mean
NMT	Male	3,670	595.73	28.498	0.470
	Female	1,079	586.61	53.972	1.643
NCEE	Male	3,670	125.01	10.052	0.166
	Female	1,079	128.10	13.601	0.414
Placement test	Male	3,540	60.01	11.882	0.200
	Female	1,033	63.64	12.650	0.394

		Symmetric measures	
		Value	Approximate significance
Nominal by nominal	Phi	−0.130	0.000
	Cramer's V	0.130	0.000
N of valid cases		4,749	

Administration

The university's in-house examinations follow a typical top-down approach to ensure fairness in administration. Several school departments at different hierarchical levels are involved in the process, among which the school registration office is in charge of scheduling examination and invigilation, while the College English teaching department is responsible for major tasks ranging from developing test papers, analyzing, and scoring, etc. Students' scores will then be used by the various schools, respectively, for ranking students in GPA, scholarship, further study, etc. Only when all parties concerned closely collaborate with each other can the in-house examinations be properly administered. In this research, we mainly focus on the role of College English teaching department in the administration of tests.

TABLE 9 Gender differences in final examinations and CET-4/6 for Level 1 students.

		Independent samples test for Level 1 students						
		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
CE-1	Equal variances assumed	2.345	0.126	−4.900	3,895	0.000	−2.2141	0.4518
	Equal variances not assumed			−4.685	1249.507	0.000	−2.2141	0.4726
CE-2	Equal variances assumed	10.868	0.001	−4.284	3,889	0.000	−1.678	0.392
	Equal variances not assumed			−3.985	1205.921	0.000	−1.678	0.421
CE-3	Equal variances assumed	3.588	0.058	−5.687	3,871	0.000	−2.105	0.370
	Equal variances not assumed			−5.394	1218.049	0.000	−2.105	0.390
CE-4	Equal variances assumed	3.473	0.062	−7.142	3,888	0.000	−2.826	0.396
	Equal variances not assumed			−6.797	1236.952	0.000	−2.826	0.416
CET-4	Equal variances assumed	1.203	0.273	−9.905	3,698	0.000	−23.259	2.348
	Equal variances not assumed			−9.674	1236.448	0.000	−23.259	2.404
CET-6	Equal variances assumed	3.393	0.066	−5.909	2096	0.000	−16.995	2.876
	Equal variances not assumed			−5.741	1000.741	0.000	−16.995	2.960

TABLE 10 Gender differences in final examinations and CET-4/6 for Level 2 students.

		Independent samples test for Level 2 students						
		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
CE-2	Equal variances assumed	0.203	0.653	−6.565	1,184	0.000	−3.1946	0.4866
	Equal variances not assumed			−6.709	818.132	0.000	−3.1946	0.4762
CE-3	Equal variances assumed	3.700	0.055	−6.156	1,184	0.000	−3.1142	0.5059
	Equal variances not assumed			−6.340	835.263	0.000	−3.1142	0.4912
CE-4	Equal variances assumed	4.108	0.043	−6.130	1,184	0.000	−3.3025	0.5387
	Equal variances not assumed			−6.387	861.195	0.000	−3.3025	0.5171
CET-4	Equal variances assumed	0.716	0.398	−4.532	1,184	0.000	−15.5580	3.4326
	Equal variances not assumed			−4.663	833.106	0.000	−15.5580	3.3361
CET-6	Equal variances assumed	0.627	0.429	−5.779	1,076	0.000	−28.3041	4.8978
	Equal variances not assumed			−5.899	792.955	0.000	−28.3041	4.7979

TABLE 11 Examination statistics of students from Tibetan and Uygur ethnic groups.

	NCEE	Placement	CE-1	CE-2	CE-3	CE-4	CET-4	CET-6
Average	76.65	40.14	43.25	45.68	48.48	50.35	372	335
Pass rate	14.08%	5.63%	9.86%	14.08%	15.49%	35.21%	12.68%	0%

Test development

The liability for the test development in this university is held by the College English teaching department. For a long time, there exist two opposing views on the nature of these terminal examinations, that is, whether they are achievement tests or proficiency tests. As neither side can prevail over the other, a compromise has to be made, with several test items including texts or questions taken from the textbooks or exercise books, and other test items chosen from test banks developed by third-party test developers.

In 2018, the teaching department decided to transform the in-house examinations from achievement-oriented to proficiency-oriented. Taking the final examinations for level 2 students as an example, except for the translation task for CE-2, all examination questions are randomly selected from the test banks. A comparison of the translation items is shown in Table 12.

In contrast to CE-3/4 for level 2 students, the translation item of CE-2 is the highest in average (16.733), skewness (−1.808) and kurtosis (10.473), and the discrimination and difficulty is only 0.03 and 0.16, respectively. Overall, the data suggest that the translation item designed by the teaching faculty rather than by a third-party test bank compromises the quality of examinations.

Inter-rater reliability

As in CET-4/6, the university's in-house College English examinations include two subjective items, writing and translation, which are rated by teachers themselves. That may give rises to an issue about inter-rater reliability. We take the scores of writing and translation (20 and 15% respectively) in CE-2/3/4 for level 2 students as an example.

Although the teaching department has issued a detailed scoring rubric for each subjective item, which is essentially the same as those in CET-4/6, there are still significant differences between raters. In Table 13, the largest score difference is 6.05 points (the total score is 35 points) in CE-4. Meanwhile, the standard deviation also varies significantly from rater to rater. Generally speaking, it is observable that the average scores are negatively correlated with the standard deviations.

Social consequences

The in-house College English examinations have both direct and indirect social consequences. Their indirect consequences, as mentioned previously, are linked to students' accessibility to taking CET-4/6, which in turn somehow affects their job-seeking and further study prospects. In the current job market, many enterprises regard CET-4/6 scores as one of the major factors in recruiting talents. For students who intend to go to graduate school in China, a high CET-6 score often helps them stand out from other candidates.

In addition to indirect social consequences, the in-house College English examinations have direct social consequences due to the fact that College English is an essential compulsory course in the curriculum for all of the non-English major undergraduates. The high-stakes nature of these in-house examinations is partly due to the sheer number of undergraduate students in the university, and partly due to the various issues that the test results are linked to, like GPA, scholarship and other crucial matters on and off campus.

To make matters worse, the school authorities, in an attempt to motivate students' English learning, have decided that more weight shall be given to the scores of the final examinations for level 2 students (weight coefficient 1.1). This equating method implies that level 2 students will always be given a huge edge over their level 1 counterparts in school-wide competitions. Table 14 lists the average scores of the students from both levels (1,186 in level 2 and 3,950 in level 1) in the final examination.

As can be seen, the scores of level 2 students are significantly lower than their level 1 counterparts. When the scores of level 2 students are multiplied by the coefficient of 1.1, the gap between the two groups will be further widened. This study holds that the above score conversion method is arbitrary and unfair.

TABLE 13 Subjective items of CE-2/3/4 for Level 2 students by different raters.

Rater	Average			Std. dev.		
	CE-2	CE-3	CE-4	CE-2	CE-3	CE-4
No. 1	29.23	26.01	28.59	2.36	3.33	2.58
No. 2	28.92	26.17	28.18	2.56	3.9	3.11
No. 3	28.64	23.96	25.11	3.03	4.27	4.28
No. 4	27.97	27.98	30.59	1.83	3.14	2.05
No. 5	27.96	23.04	26.91	3.6	4.44	3.01
No. 6	27.7	23.83	25.46	3.65	3.35	2.41
No. 7	27.07	23.9	24.54	2.29	3.19	4.3

TABLE 14 Comparison between the scores of both levels by semester.

	1st Semester	2nd Semester	3rd Semester	4th Semester
Level 1 Average	64.33	64.36	64.05	63.57
Level 2 Average	77.47	70.34	70.02	N/A

TABLE 12 Translation items in CE-2/3/4 for Level 2 students.

	Descriptive statistics						
	N cases	Average	Std Dev.	Skewness		Kurtosis	
				Statistics	Standard error	Statistics	Standard error
CE-2	1,179	16.733	1.9425	−1.808	0.071	10.473	0.142
CE-3	1,170	9.311	2.5974	−0.795	0.072	0.556	0.143
CE-4	1,125	11.757	1.6737	−1.384	0.073	5.333	0.146

Discussion

This study analyzes the five aspects of fairness of the in-house College English examinations under the framework of TFF theory. To sum up, the major problems of the in-house College English examinations are listed as follows:

1. These examinations are reliable and valid enough in terms of internal consistency, content validity and criterion-related validity. However, they are not so well-suited for students in level 2 as for level 1, and are poorly correlated with CET-6, due to the fact that most in-house examinations tend to align with CET-4 rather than CET-6.
2. No bias has been found in the in-house examinations per se, but in the broad sense, these examinations are biased against male and ethnic minority students, since they usually help consolidate or even exacerbate the performance gap between these subgroups and their counterparts.
3. The in-house examinations are accessible to all non-English major undergraduates free of charge. Yet these examinations, the placement test in particular, can determine the time of the students' access to CET-4/6.
4. The in-house examinations are administered by the College English teaching department in collaboration with a few other school departments. The test items designed by the teaching staff are not as good as those taken from a third-party test bank. Also, different raters, who are teachers in this case study, do not have the same scoring standards.
5. The in-house examinations are linked to a variety of students competitions and evaluations either on campus or off campus. Its direct and indirect social consequences can be vividly found during college and well after graduation.

Since in-house College English examinations are often independently administered by the colleges and universities concerned, it is hard to assert that the problems mentioned above are universal to other higher learning institutions in China. However, we strongly believe that test fairness, whatever forms it may assume, is a universal concern in the majority of colleges and universities in China. According to an exhaustive survey of 672 colleges and universities from many regions in China. Jin (2020) found that 85% of these institutions have their own in-house College English examinations, with the major difficulties including lack of faculty with relevant language assessment literacy (32%), lack of time and funds necessary for the development and implementation of in-house examinations (52%), negligence of the importance of in-house examinations (27%), etc.

These underlying factors, which inevitably affect the fairness of in-house College English examinations, are also found in our case study. To begin with, the language assessment literacy of teachers, who play pivotal roles such as test developers, examiners and raters, has much to be desired. Although the in-house College English examinations have been conducted for years, few of its College English teaching faculty specialize in language testing, or have ever taken long-term training programs for language testing in recent years.

Furthermore, this study also finds that the in-house College English examinations are under-resourced. Since the in-house examinations are considered merely an imperative procedure for College English courses, as an obligation for teachers despite the heavy workload they have already had, neither the school authorities nor teachers have given barely enough funds or time for examination-related researches.

To make matters worse, teachers seem to be unaware of the existence and significance of test fairness. First, only when the data of all test takers are collected, analyzed and compared can we have an overall picture about test fairness. Since most teachers have only access to the results of their classes, they are usually ignorant of how serious the problem may be. In our case study, it is by comparing different levels, genders, ethnic groups and raters that we are able to find some evidence of unfairness. Also, the fact that the in-house examinations have serious impacts is not fully understood by teachers. Most teachers have the illusion that their mission is completed after the students' scores are submitted to the school's examination database. What they fail to notice, however, is that these examinations are high-stake in nature, whose scores will be used by other parties in various decision-making processes, and can have significant implications for students' academic and professional career in the future.

As for the measures that can be taken to ensure test fairness, the present study provides the following suggestions:

First, we should adopt a holistic approach to test fairness of the in-house College English examinations. As an all-encompassing concept, test fairness shall be ensured only when all parties and procedures involved are linked to each other in an organic way. This calls for an overarching unit that play the role as test developer, supervisor, coordinator and assessor, etc. Develop comprehensive policies outlining roles and responsibilities, conduct regular training sessions on best practices, implement pilot tests to refine processes, and establish continuous feedback channels from all stakeholders to improve future iterations.

Second, we should adhere to the essence of in-house College English examinations. These examinations are supposed to serve the talent cultivation objectives, and cater to the realities of the student body in the university concerned. Those factors should be borne in mind as supreme standards in our test practice. We need to avoid two extremes, i.e., blindly following the test banks developed by third-party professionals, or solely relying on the institution's teaching faculty to design examinations independently, which are inefficient at best, and unfair at worst. In fact, given the general shortage of talents with language assessment literacy, we may advocate for a language testing consortium that rally the talents from different colleges and universities, and share their intellectual outputs among institutions with similar teaching and learning backgrounds. We will conduct surveys and focus groups to understand the specific needs of students and faculty, tailor test content to align with institutional objectives and student realities, and periodically review and update test formats based on feedback and evolving educational requirements. This approach ensures that assessments remain fair, relevant, and supportive of the institution's goals.

Finally, we may need to lower the stakes of individual examinations. Today, we can measure students' English proficiency and performance through a variety of means, ranging from the more scientific external tests such as CET-4/6, TOEFL and IELTS tests, to the more dynamic formative or classroom assessment. The in-house College English examinations, which are summative in nature, should be incorporated with other assessments to form a continuum that facilitate learning and teaching. Besides, we may well lower the stakes of in-house examinations by reducing the proportion of the scores in the final assessments, and by implementing a dynamic hierarchical teaching mode in which students can shift between different levels on the basis of their performances. We will revise grading policies to reflect the new assessment structure, ensuring they align with our goals for fairness and effectiveness. A clear communication plan will inform students and faculty about these changes through detailed guidelines and training sessions. Additionally, we will regularly monitor the system's effectiveness and make necessary adjustments to ensure continuous improvement and optimal performance.

Conclusion

Finding

In the context of the increasingly widespread use of in-house College English examinations in China, due attention has not been given to the test fairness yet. This empirical research is a two-year longitudinal study on the in-house College English performances of the non-English major undergraduates in a specific Chinese university. Under the guidance of Kunnan's TFF theory, we have found that the fairness is a pervasive problem in the in-house College English examinations, which, though in heterogeneous forms, are due to some underlying factors including lack of talents with language assessment literacy, short of time and funds, or negligence of the high-stakes nature of those examinations.

To ensure the fairness of test, this research suggests a holistic approach, rather than a piecemeal one, that involves all of the parties and procedures in the in-house College English examinations. If necessary, we may resort to external experts, who can form alliances to provide high-quality test services. However, the specific talent cultivation goals and students' language ability must be factored in. Also, the in-house College English examinations should be reshaped as medium-stakes or low-stakes forms of assessment, especially when the test fairness is still an issue for the time being.

This work, though contributive to the field, exhibits several methodological limitations that warrant acknowledgment. Primarily, the utilization of Kunnan's Test Fairness Framework (TFF) as the underpinning theoretical construct and analytical instrument is not accompanied by a thorough elucidation of its nuanced application and empirical operationalization within the context of this investigation. Secondly, the study's purview is circumscribed to a single institution of higher learning in East China. While this focused case study approach undoubtedly yields profound and contextually rich insights, it concurrently raises concerns regarding the external validity and the breadth of the research's representative claims. The constrained sample selection may encumber the extrapolation of findings to a broader educational milieu. Future studies should adopt a more expansive sampling strategy.

As Yang (2015) said, examinations, teaching and the use of examination results constitute a complex system, and it is imperative to take a systematic approach to the relationship between the three pillars. All stakeholders of the in-house College English examinations should coordinate with each other in their efforts to ensure validity of testing, effectiveness of teaching and validity of test use. In that sense, test fairness or the lack of it is a factor that has far-reaching effects way beyond test itself.

Limitations and future recommendations

This study, while valuable, has several limitations. The application of Kunnan's Test Fairness Framework (TFF) lacks detailed explanation and empirical operationalization within this specific context. Additionally, focusing on a single institution in East China limits the external validity and generalizability of the findings. Methodologically, deeper exploration of factors like institutional policies, teacher training, and student preparation is needed.

Future research should adopt broader sampling to enhance representativeness and validity. Detailed explanations of theoretical frameworks like TFF should be provided, possibly through case studies or comparative analyses. A holistic approach involving all stakeholders, including external experts, can better address talent cultivation goals and students' language abilities. Coordinating efforts among stakeholders for valid testing, effective teaching, and proper test result usage is crucial. Reshaping exams into medium-or low-stakes assessments could reduce pressures and biases, improving overall fairness and effectiveness.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by Hefei University of Technology. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

LK: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. KM: Writing – review & editing. LL: Writing – review & editing. LJ: Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded by Anhui Provincial Philosophy and Social Sciences Planning Project grant number [No. AHSKY2022D223].

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

AERA (2014). Standards for educational and psychological testing. *5th* Edn. Washington, DC: American Educational Research Association.

Angoff, W. H. (1993). "Perspectives on differential item functioning methodology" in Differential item functioning. eds. P. W. Holland and H. Wainer (Mahwah, NJ: Lawrence Erlbaum Associates, Inc.), 3–23.

Brown, J. D. (1996). Testing in language programs. Upper Saddle River, NJ: Prentice hall.

Camilli, G. (2006). "Testing fairness" in Educational measurement. ed. R. Brennan. *4th* ed (Westport, CT: Praeger), 221–256.

Davies, A. (2010). Test Fairness: A Response. *Lang. Test.* 27, 171–176. doi: 10.1177/0265532209349466

Fan, J. (2014). Test fairness research: concepts, theories, and responsibilities. *For. Lang. Test. Teach.* 4, 26–34.

Fan, J., and Ji, P. (2013). Examining the validity of the Fudan English test (FET): test data analysis. *For. Lang. Test. Teach.* 2, 45–53.

Guo, Y., and Lin, G. (2016). Constructing a college English language testing system for Beijing Normal University: Theory & Practice. Beijing: Foreign Language Teaching and Research Press.

Hamid, M. O., Hardy, I., and Reyes, V. (2019). Test-takers' perspectives on a global test of English: questions of fairness, justice and validity. *Lang. Test. Asia* 9:16. doi: 10.1186/s40468-019-0092-9

Hamp-Lyons, L. (1997). "Ethics in language testing" in Encyclopedia of language and education. eds. C. Clapham and D. Corson (Dordrecht: Springer Netherlands), 323–333.

Jensen, A. R. (2006). Bias in mental testing. New York: Free press.

Jia, W., Chang, X., and Tang, X. (2013). A survey on school-based college English tests. *For. Lang. Test. Teach.* 2, 54–59.

Jin, Y. (2015). Constructing a comprehensive and diversified college English curriculum evaluation system: needs analysis and the way forward. *For. Lang. China* 12:5. doi: 10.13564/j.cnki.issn.1672-9382.2015.03.002

Jin, Y. (2020). College English testing and assessment: current practices and future developments. Foreign language. *World* 5, 2–9.

Kane, M. T. (2010). Validity and fairness. *Lang. Test.* 27, 177–182. doi: 10.1177/0265532209349467

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

Kunnan, A. J. (2004). "Test Fairness" in European language testing in a global context. eds. M. Milanovic and C. Weir (Cambridge University Press: Cambridge. UK), 27–48.

Kunnan, A. J.. Towards a model of test evaluation: using the test fairness and test context frameworks. In: Proceedings of the ALTE Berlin conference, Berlin, Germany. (2005). pp. 229–251.

Kunnan, A. J. (2008). "Towards a model of test evaluation: using the test fairness and wider context frameworks" in Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity - proceedings of the ALTE Berlin conference. eds. L. Taylor and C. Weir (Cambridge, UK: Cambridge University Press), 229–251.

Kunnan, A. J. (2014). "Fairness and justice in language assessment" in The companion to language assessment. ed. A. J. Kunnan (John Wiley & Sons, Inc.: New York. NY), 1098–1114.

McNamara, T., and Ryan, K. (2011). Fairness versus justice in language testing: the place of English literacy in the Australian citizenship test. *Lang. Assess. Q.* 8, 161–178. doi: 10.1080/15434303.2011.565438

Moghadam, M., and Nasirzadeh, F. (2020). The application of Kunnan's test fairness framework (TFF) on a Reading comprehension test. *Lang. Test. Asia* 10:1. doi: 10.1186/s40468-020-00105-2

Roever, C., and McNamara, T. (2006). Language testing: the social dimension. *Int. J. App. Linguist.* 16, 242–258. doi: 10.1111/j.1473-4192.2006.00117.x

Sun, H., Zhang, J., and Xiong, J. (2020). A survey on school-based college English tests in China and its implications. Foreign language. *World* 200, 63–71.

Wallace, M., and Qin, C. Y.. Language classroom assessment fairness: perceptions from students. (2024) Available at: https://www.tci-thaijo.org/index.php/learn (Accessed on 7 August 2024).

Xiaoming, H. (2010). How do we go about investigating test fairness? *Lang. Test.* 27, 147–170. doi: 10.1177/0265532209349465

Yang, H. (2015). Valid testing, effective teaching, and valid test use. *J. For. Lang.* 38, 2–26.