International Journal of Language and Linguistics
2025, Vol. 13, No. 3, pp. 147-157
https://doi.org/10.11648/j.ijll.20251303.14

Research Article

From Validity to Equity: Revisiting CET-4 Through Messick, Kane, and Paraskeva's Frameworks

Zhi Ma[*]

International Affairs Office, Shanghai University of International Business and Economics, Shanghai, China

Abstract

The College English Test Band 4 (CET-4) is a high-stakes, standardised English proficiency assessment widely administered in Chinese universities. Initially developed to evaluate non-English majors' language competence, it has since evolved into a gatekeeping mechanism that influences graduation eligibility, access to scholarships, postgraduate admissions, and employment opportunities. As English education becomes increasingly globalised, concerns have emerged about the test's ability to fairly and validly assess diverse learners across China. In particular, the CET-4 has been criticised for its narrow focus on decontextualised tasks, its reliance on Western-centred content, and its failure to accommodate the sociolinguistic realities of rural, minority, and low-income students. This study critically reassesses the CET-4 through the lens of three theoretical frameworks: Messick's theory of validity, Kane's principles of fairness, and Paraskeva's epistemic pluralism. Drawing on these perspectives, it examines the test's alignment with international benchmarks, such as the Common European Framework of Reference for Languages (CEFR). It identifies key limitations in construct validity, cultural inclusivity, and equitable access. The analysis reveals that the CET-4 often produces construct-irrelevant variance, reinforces systemic educational inequalities, and reduces English learning to test-oriented preparation. The study proposes a multidimensional reform agenda that includes modular testing tailored to regional contexts, performance-based assessments that reflect real-world communication, and integrating culturally responsive content. It also recommends stakeholder engagement through participatory test design, implementing fairness audits, and adopting equity-based funding models such as Weighted Student Funding (WSF) to address systemic disparities. These reforms seek to uphold psychometric rigour and ethical responsibility, ensuring that large-scale assessments like the CET-4 better reflect the realities and needs of diverse test-takers. Ultimately, the paper argues that the CET-4 must transition from a rigid gatekeeping instrument to a more inclusive and context-sensitive evaluation platform that aligns with international standards while promoting educational equity and opportunity in the Chinese context.

Keywords

Validity, Fairness, Cultural Inclusivity, College English Test Band, Educational Equity, Assessment Reform

1. Introduction

English proficiency is regarded as a gateway to academic achievement and professional advancement in today's globalised world and plays an increasingly critical role in shaping futures [8]. In China, this reality is epitomised by the College

[*]Corresponding author: Zhi.ma@bath.edu (Zhi Ma)

Received: 12 April 2025; **Accepted:** 10 May 2025; **Published:** 16 June 2025

English Test (CET), a national high-stakes standardised assessment that can make or break students' prospects. The CET, overseen by the Ministry of Education, serves as a yardstick for English language proficiency through its two primary levels—CET-4 and CET-6—alongside the spoken English components, CET-SET4 and CET-SET6. CET-4, a mandatory hurdle for most non-English major undergraduates, assesses a range of skills, including listening, reading, writing, and speaking, to assess English proficiency comprehensively [1, 7].

Since its establishment in the 1980s, the CET has become integral to China's higher education system [29]. It is often perceived as more than just a test—its certificates open doors to academic milestones, competitive job markets, and career advancements, particularly in fields requiring international collaboration perceived not merely as a language test but as a credential that can unlock academic opportunities, facilitate entry into competitive job markets, and support career advancement, especially in fields with international dimensions [7, 30]. However, despite its intended role as an objective measure of proficiency, concerns have emerged regarding its inclusivity and fairness.

Critics argue that the CET reflects inner-circle English norms—such as those of the United States and the United Kingdom—that do not align with the multilingual and multicultural realities of many test-takers. Critics dub the CET both inner circle English norms, like the US or British English, which are usually more than its candidates' multilingual and multicultural lives. Students from marginalised, rural, or low-income backgrounds often encounter cultural content that feels distant from their lived experiences [27]. For students who come from marginalised communities, are from the country, or do not have much money, this can seem to come from someone else and be far away from their lives. In addition to cultural incongruence, inequalities in access to resources further disadvantage these students, raising concerns about the fairness and validity of the test [14, 27].

This paper will explore the validity and fairness of the CET, which says its technical soundness and implementation of its design bring out disparities in the process. It also reflects the biases and disparities incorporated in the test, which must be revised to evaluate the truth accurately. This paper examines the validity and fairness of the CET-4, focusing on how technical design and implementation contribute to disparities. It analyses cultural and systemic biases embedded in the test, arguing that substantial revisions are necessary for the CET to provide an equitable evaluation of English proficiency. Drawing on theoretical frameworks such as Messick's (1989) theory of validity [18], Kane's (2010) principles of fairness [11], and Paraskeva's (2011, 2018) epistemic pluralism [21, 22], this study challenges the notion that a singular cultural or regional standard should determine proficiency.

This type of research, rather than just a technical analysis, goes further into the ethical issues associated with the CET. The prejudices hidden in the test may serve as the linchpin for social stratification, compromising the possibility of even the brightest students from remote regions reaching their potential for academic and professional success. These inequities in the system call for radical reforms. Rather than offering a purely technical critique, this research explores the ethical implications of CET-4 as a high-stakes assessment. By identifying embedded prejudices contributing to educational stratification, it argues for reforms that would transform the test from a gatekeeping mechanism into a more inclusive tool for opportunity. Recommendations include modular testing, performance-based assessment, culturally inclusive content, and integrating stakeholder perspectives through fairness audits and targeted policy reforms [15].

2. Theoretical Foundations of Validity and Fairness

2.1. Messick's Approach to Validity: Beyond Technical Precision

As conceptualised by Messick (1989), validity extends beyond a one-dimensional definition focused solely on technical accuracy [18]. Validity, as discussed by Messick (1989), was discussed differently than a one-dimensional concept that stands for a high level of precision and technical correctness. He proposes a dual-layered model: the evidential basis of validity, which evaluates how well test content reflects the intended construct, and the consequential basis, which addresses the societal and ethical implications of test use. This holistic perspective asserts that even a technically sound test may be invalid if it produces unjust consequences.

Applying Messick's framework to the CET-4 exposes multiple limitations. Nevertheless, using this analytical model to the CET in China leads to several consequential problems from cultural prejudices and system-related inequalities. First, the CET often incorporates culturally biased content, particularly Western-centric idioms, references, and contexts. These elements are not necessarily familiar to urban pupils, especially if they may be unfamiliar to test-takers from rural or under-resourced areas, undermining the fairness of score interpretations [1, 7, 27]. For instance, listening sections can be composed of dialogues from Western social interactions about scenarios or customs not found in China. Though reading passages sometimes refer to Western historical facts, the same happens with literature not written in Chinese and not part of the mandatory school program, candidates were not involved in before joining their higher education listening sections may feature dialogues rooted in Western social customs, and reading passages may reference literary or historical material not taught in Chinese curricula. Such content introduces construct-irrelevant variance, causing scores to reflect cultural familiarity rather than genuine English proficiency [24].

Second, disparities in educational resources further compromise the validity of CET-4 outcomes. Systemic inequities in literary resources undermine the legitimacy of the CET even further. Students from rural schools need help from their urban counterparts. They also face many shortcomings, including frequently lacking access to qualified English teachers, up-to-date materials, and immersive English environments [14]. These performance inconsistencies are elevated by economic circumstances, making access to private tuition and test preparation inaccessible to some.

While Messick's framework highlights the importance of consequences, it insufficiently addresses how structural inequalities affect validity. Messick's theory inadequately addresses external variables. It rests on the flawed assumption that this style of testing guarantees equal training opportunities for all test takers, on the assumption that test-takers begin with equal access to learning opportunities. Performance gaps may result from unequal opportunities rather than actual differences in proficiency. The CET risks perpetuating rather than correcting educational disparities without considering these external factors. Hence, although Messick's theory enriches our understanding of validity, it requires expansion to encompass sociocultural realities that shape assessment outcomes in the Chinese context.

2.2. Kane's Fairness Framework: Bridging Equality and Opportunity

Kane (2010) views fairness not as an optional complement to validity but as a central criterion of assessment quality [11]. Kane (2010) claims that fairness is not simply a complementary concept of validity but a crucial part of the fairness criteria as it implies that a particular assessment cannot be valid if it systematically disadvantages a group of test-takers. He distinguishes between two core dimensions: procedural fairness, which concerns the consistent and impartial administration of tests, and equitable fairness, which addresses whether all test-takers have meaningful access to the preparation and resources needed to succeed.

This framework offers a valuable lens through which to examine the CET-4. Procedural fairness is generally upheld in the test's administration, which follows standardised processes nationwide. His perspective is based on fairness in two dimensions: Procedural and equitable fairness, which systematise ability and chances in any action. The fair procedure is focused primarily on consistency and impartiality in how tests are taken and graded, keeping in mind that all candidates should be subject to the same testing conditions. However, equal access to the test and its resources was the core issue of equity of opportunity, we should also ensure that all examinees get a fair chance by being able to access the test and its preparation resources. Significant concerns arise when considering equitable fairness. Many students from rural and marginalised backgrounds face substantial barriers to accessing quality English instruction, experienced teachers, and supportive learning environments [14].

These disparities have material consequences. In China, candidates for the CET from rural areas often lack access to quality English education, experienced teachers, and supplementary resources compared to their urban counterparts. For example, rural students may have limited exposure to authentic English input and fewer opportunities for test preparation, particularly for listening and speaking tasks that benefit from interactive practice. As a result, their performance may reflect environmental disadvantages rather than actual language ability [29].

While attentive to fairness, Kane's model does not fully account for these systemic issues. Additionally, Kane's criticism may seem misguided. It suggests fairness is simply about treating everyone equally during testing. Its focus on treating all test-takers the same during assessment overlooks the unequal starting points from which students approach the test. More profound inequities in education access and social capital remain unaddressed when fairness is equated with uniform procedures.

Thus, although Kane's framework strengthens the ethical grounding of assessment, its application to the CET-4 reveals limitations. Achieving fairness requires standardised administration and structural reforms that address socioeconomic disparities and ensure all students can engage with the test on equitable terms.

2.3. Connecting Validity and Fairness Through Inclusive Design

Validity and fairness are inherently interlinked in the context of educational assessment. Fairness is not only an ethical tactic but a fundamental component of validity, and validity and fairness are closely related in the creation and evaluation of assessments. According to Messick's (1989) theory of consequential validity, a test's social impact is just as significant as its technical soundness [18], such as its capacity to measure the desired constructs. Even if a test accurately measures the intended construct, it may still be considered invalid if its use has adverse consequences for specific groups, such as underrepresented or disadvantaged students.

Kane's (2010) framework complements this view by introducing procedural and equitable fairness as essential considerations for ethical assessment design [11]. Procedural fairness emphasises consistent and unbiased administration, while equitable fairness ensures that all examinees have access to the resources and preparation needed to perform successfully. Kane's approach to Messick's consequential validity demonstrates how social injustices, including unequal access to trained instructors or study materials, compromise validity and fairness. These theories underscore that validity cannot be separated from fairness when applied together.

Paraskeva's (2011, 2018) concept of epistemic pluralism adds a further dimension by advocating for the inclusion of diverse cultural and linguistic perspectives in assessment design [21, 22]. His theoretical approach emphasises that assessments should reflect the realities and experiences of all test-takers rather than favouring a narrow, culturally specific viewpoint. He argues that traditional assessment models often reflect dominant paradigms and fail to account for marginalised groups' lived experiences and knowledge systems. In the context of the CET-4, this implies a need to diversify test content beyond Western-centric norms to reflect all learners' realities better.

Incorporating epistemic diversity reduces construct-irrelevant variance by minimising the advantage held by students familiar with dominant cultural references. Embracing epistemic diversity makes fairness an essential aspect of validity. Incorporating varied perspectives reduces construct-irrelevant variance and ensures that the examination assesses genuine language competency rather than mere familiarity with dominant cultural norms. This enhances fairness and validity, aligning with the principle that assessments should reflect genuine language proficiency rather than background familiarity. For example, modifying listening and reading tasks to include scenarios familiar to rural or minority students would make the test more accessible without compromising rigour.

Creating inclusive assessments also involves participatory design processes that engage diverse stakeholders. Assessment design is made equitable by promoting users' participation and raising awareness of the necessity of the keen involvement of marginalised social groups in providing assessment content. Involving teachers, students, and community representatives in test development can surface biases and lead to more equitable outcomes. This approach recognises that fairness is not just about uniform procedures but contextual relevance and social responsibility.

Ultimately, aligning Messick's, Kane's, and Paraskeva's frameworks reveals that assessments like the CET-4 must evolve beyond technical precision. Several factors dealing with the validity issue entail device and fairness; the task can be even trickier. They must also address the ethical implications of test content and administration. By embedding cultural inclusivity, community voice, and equity-oriented reforms, assessments can serve as tools for opportunity rather than instruments of exclusion.

2.3.1. Promoting Diversity in Test Content

Promoting diversity in assessment content is critical for improving validity and fairness. Paraskeva's arguments of epistemic pluralism match the evidential dimension of validity through test content that reflects and incorporates all test-takers' realities. Paraskeva's epistemic pluralism supports this approach by arguing for assessment materials that reflect the realities of all learners, not only those from dominant sociocultural backgrounds [21, 22].

This involves adapting test items to include contexts, themes, and scenarios that resonate with students across different regions, income levels, and cultural backgrounds. This requires transformation to the local or cultural sense of language and its invented context in the CET context, shifting the focus beyond Western-centric models. For example, listening tasks referencing local community settings or daily student life in rural China may better engage test-takers than content drawn exclusively from Western or urban scenarios. Such contextualisation reduces construct-irrelevant variance and ensures that assessments accurately reflect language proficiency rather than cultural familiarity.

Developers can ensure that language is being tested by designing test content that reflects broader linguistic and cultural diversity, rather than background knowledge of dominant cultural practices. An illustration of this statement would be listening to the sections referring to school experiences, which are more appropriate to the lives of regional students than to state ones that highlight construct-irrelevant variance and content validity. By this process, the language is thoroughly tested, not the knowledge of manners.

2.3.2. Creating Inclusive and Fair Assessment Practices

Developing inclusive and fair assessment practices requires active stakeholder engagement and culturally responsive design. Assessment design is made equitable by promoting users' participation and raising awareness of the necessity of the keen involvement of marginalised social groups in providing assessment content. This means involving representatives from marginalised communities, such as rural educators, ethnic minorities, and low-income groups, in test development to ensure that diverse perspectives are reflected in test items. Without these voices, test developers risk embedding unconscious bias that disadvantages underrepresented learners. It will not be possible to eliminate biases that negatively impact the specific groups of test developers without the communities' perspectives, which exemplify these biases. This participatory approach strengthens the fairness and validity of assessments by ensuring that they align with all learners' social and cultural contexts. Ultimately, fairness is not just about uniform procedures, but about whether the assessment genuinely reflects and respects the diversity of those it evaluates.

2.3.3. Decolonisation and Consequential Validity

Decolonising assessment design highlights the ethical imperative to address cultural dominance and systemic exclusion in high stakes testing. When tests like the CET-4 draw heavily from dominant cultural norms—often Western or urban—they risk reinforcing social inequalities and marginalising students whose experiences diverge from these standards. As a high-stakes exam that significantly influences students' educational and career opportunities, the CET-4 has far-reaching consequences. When test content privileges particular cultural narratives, students from non-dominant groups face barriers to fair participation and achievement.

Decolonising the CET involves integrating locally relevant cultural content and developing scoring rubrics that are sensitive to contextual variation. This broadens the test's relevance and fairness and enhances its consequential validity, ensuring that its social outcomes align with principles of equity and justice.

2.3.4. Addressing Systemic Challenges to Fairness

Implementing Paraskeva's theory to strengthen the relationship between validity and fairness in the CET-4 presents complex challenges that require thoughtful solutions [21, 22]. A central tension lies in balancing the need for standardisation with including diverse cultural perspectives. While psychometric reliability ensures consistency across test-takers, it can conflict with efforts to reflect regional and cultural variation [11]. Overcoming this challenge demands an approach to test design that upholds technical rigour while acknowledging the diverse realities of learners.

One key solution is to involve marginalised voices—such as rural educators, ethnic minorities, and low-income students—in curriculum and test development. This participatory process helps ensure test content reflects a broader range of lived experiences, increasing relevance and legitimacy [17, 28]. Rather than viewing fairness solely in procedural terms, inclusive design recognises the social contexts that shape learning and testing conditions.

A comprehensive strategy is necessary to tackle these challenges, aligning Paraskeva's framework with actionable methods to enhance validity and fairness [21, 22]. One viable solution is implementing a modular test design, which harmonises standardisation and diversity by customising test sections to resonate with regional or cultural contexts. This method maintains the assessment's reliability while ensuring the content remains relevant for all test-takers [16].

In parallel, equity-focused policies are essential. These include targeted investment in teacher training, instructional materials, and preparatory resources for under-resourced schools. Such reforms help reduce structural disparities and enable students from marginalised backgrounds to engage with the CET-4 more equitably [14, 17].

3. Evaluating the CET-4: Insights and Critiques

Evaluating the CET-4 requires critically examining its role as a language proficiency measure and a gatekeeping tool within China's educational system. The written CET-4 lasts 125 minutes and has long been a key assessment tool in measuring the English proficiency of university students in China. Intended initially to assess non-English majors' language skills, it is now widely used as a graduation requirement and a criterion for postgraduate study and job access [7, 30].

It is one of the graduation criteria for most universities regarding bachelor's degrees. However, critiques from empirical research and theoretical frameworks reveal limitations in the test's construct and content validity. This section integrates empirical insights and compares the CET-4 with international frameworks, particularly the Common European Framework of Reference for Languages (CEFR), to assess its alignment, equity, and authenticity. It explores the test's conformity to international standards, identifies issues of equity and authenticity, and offers recommendations for reform.

Reading and listening sections often draw on culturally specific references, including Western idioms and unfamiliar settings, which may disadvantage students from rural or under-resourced backgrounds [24, 27]. Such construct-irrelevant variance undermines validity, as scores may reflect cultural familiarity rather than actual language proficiency.

The test's high stakes have also fuelled a growing test-preparation industry, where students focus on memorising formats and fixed expressions instead of developing communicative competence [9, 13]. In many underfunded institutions, teaching to the test becomes the only viable strategy, narrowing the curriculum and further distancing instruction from the CET-4's intended goals. For the CET-4 to remain relevant and credible, it must evolve beyond a one-size-fits-all framework. This includes diversifying content sources, adapting delivery modes, and calibrating performance benchmarks for learners' varied contexts. Without meaningful reform, the CET-4 risks reinforcing the same educational barriers it was designed to overcome.

3.1. Challenges in Validity and Fairness

One key concern in evaluating the CET-4 is the extent to which it validly and fairly reflects real-world language proficiency. Messick's (1989) theory of validity stresses that assessments should align with the communicative tasks learners are expected to perform beyond the test setting [18].

Empirical studies raise questions about the CET-4's content validity. Ying and Liying (2008) argue that its heavy reliance on standardised item formats fails to capture the practical language skills needed in authentic contexts. Zhao (2022) supports this critique, observing that although the test shows strong criterion validity (r = 0.754), it inadequately represents the multifaceted nature of English use in real-world scenarios [31].

Paraskeva's (2011, 2018) concept of epistemic plurality provides a valuable lens for critiquing the CET-4's Western-centric orientation, which tends to marginalise test-takers from culturally and linguistically diverse backgrounds [22]. Wang (2023) underscores this issue, arguing that CET-4 writing tasks are overly standardised and limited in their ability to assess higher-order thinking or authentic communication [26]. These findings suggest that the CET-4 reinforces systemic inequities by privileging learners with greater exposure to Western linguistic and cultural norms, typically those in urban environments.

3.2. A Global Lens: Aligning the CET-4 with CEFR Standards

A comparison between the CET-4 and the Common European Framework of Reference for Languages (CEFR) reveals distinct priorities in language assessment. The CET-4 prioritises decontextualised tasks such as vocabulary cloze and sentence translation, focusing on linguistic precision and grammatical knowledge. In contrast, the CEFR adopts an action-oriented approach, emphasising integrated and authentic language use. It promotes interactive activities, such as simulated dialogues and real-life scenarios, that assess learners' ability to communicate effectively in practical contexts [5]. By mapping CET-4 task types to CEFR descriptors, educators and policymakers can better evaluate alignment and identify areas where the CET-4 could adopt more communicative, learner-centred elements.

3.2.1. Writing Section

The CET-4 writing tasks require candidates to produce 120-word essays on argumentative or descriptive topics, such as university libraries or public services. This aligns well with CEFR B1-B2 descriptors [5]:

B1: Produce simple connected text on topics that are familiar or of personal interest.

B2: Produce clear, detailed text on various subjects and explain a viewpoint, giving pros and cons.

The authentic CET-4 papers from June 2024 (see Appendix 1) exhibit deliberate subject selection, often grounded in students' campus experiences and points of view. This initiative anchors the writing assignments in recognisable situations, promoting accessibility and inclusion. The CET-4 writing component better connects with CEFR's communicative objectives by including scenario-based questions, such as composing professional emails or addressing workplace situations [6]. These assignments would augment the authenticity and practicality of the evaluation, endowing students with transferable abilities relevant outside academic settings.

As Montenegro and Jankowski (2017) proposed, integrating culturally responsive assessment methodologies may enhance the CET-4's conformity with CEFR principles [19]. Writing prompts representing multiple cultural contexts would provide a more precise assessment of applicants' communication skills in authentic environments, promoting inclusion. This aligns with Bryan and Lewis's (2019) assertion that evaluations must consider participants' cultural contexts to guarantee equity and pertinence [3].

3.2.2. Listening Section

CET-4 listening tasks include news items, dialogues, and passages in standard British or American accents, aligning partially with CEFR B1-B2 descriptors [5]:

B1: Understand the main points of clear standard speech on familiar matters, such as work or school-related topics.

B2: Understand extended speech and follow complex arguments in familiar contexts.

The CET-4 June 2024 exam (Paper 1, Section A, see Appendix 1) includes a news story about a transit project, which assesses understanding of information. This corresponds with CEFR B1 competencies, although it needs activities that promote active engagement, such as summarisation or problem-solving based on the audio. CEFR B2 descriptors promote interaction elements that are inadequately reflected in the CET-4.

Wang (2014) emphasises that the CET-4 listening materials, despite their diverse subjects, often do not represent the tonal and linguistic variety of authentic English [25]. The absence of exposure to global English accents constrains the test-takers' preparedness for many communication contexts. Integrating interactive challenges and diverse accents will enhance students' readiness for real-world listening requirements and connect the CET-4 with the CEFR's emphasis on communication.

3.2.3. Reading Section

The CET-4 reading section includes vocabulary cloze exercises, skimming tasks, and comprehension questions. These align with CEFR B1-B2 descriptors but often prioritise factual recall over higher-order critical evaluation [5]:

B1: Can identify the main idea of standard texts on familiar topics.

B2: Can understand texts that consist mainly of high-frequency language and can infer meanings.

Han (2021) criticises the CET-4 for limiting test-takers' exposure to real-world text genres, such as multimedia or business reports, by relying on standardised problems with slight variations in genre. China's Standards of English Language Ability (CSE) and the CEFR highlight the importance of integrating various types of texts to accurately depict the complex nature of real-world English use [23]. Incorporating professional papers and informal interactions within the reading part would improve construct validity and accord with the CEFR's action-oriented paradigm.

3.2.4. Translation Section

The CET-4 translation task requires candidates to render a Chinese paragraph into English, prioritising linguistic accuracy over adaptive mediation. This aligns with CEFR B2-C1 descriptors [5]:

B2: Can translate text clearly and convey key ideas.

C1: Can adapt language style for specific audiences and contexts.

For example, translating a section on Chinese culture is part of the June 2024 translation work (see papers 1 and 2 in Appendix 1). This demonstrates linguistic clarity but highlights fidelity to the original text, which limits options for adaptive mediation, a crucial aspect of CEFR C1. Chunyuan et al. (2024) argue that translation tasks can be improved by incorporating culture-specific elements and promoting methods for content adaptation for diverse audiences [4]. This would enhance equity and practical relevance, aligning the CET-4 with the CEFR's focus on mediation.

4. Ethical Concerns in CET-4 Design

4.1. Mitigating High-stakes Pressure on Students

The high-stakes nature of the CET-4 intensifies its ethical challenges and negatively influences teaching and learning practices. Its role in graduation decisions, university admissions, and employment opportunities places excessive pressure on students and educators [7, 30].

Zhao (2022) and Bai (2020) argue that this intense focus encourages test-specific strategies that prioritise rote learning over the holistic development of language skills [2, 31]. Similarly, Wang (2023) critiques the CET-4 writing section for fostering formulaic responses, limiting opportunities for critical thinking and real-world communication [26]. This results in an adverse washback effect, where authentic language learning is sidelined in favour of test performance.

Messick's (1989) validity framework emphasises the importance of aligning assessment outcomes with broader educational objectives. From this perspective, the CET-4's reinforcement of surface-level learning signals a misalignment between assessment and intended pedagogical goals [18]. Paraskeva's theory of participatory fairness further critiques the lack of inclusion of affected communities, particularly students and educators, in shaping the test [21, 22].

To address these concerns, policymakers should involve stakeholders such as teachers, students, and community representatives in the test development process. Such participatory design would help mitigate adverse washback effects and restore the CET-4's ethical legitimacy by positioning it as a tool that supports meaningful language learning.

4.2. Addressing Cultural Preference for Fairer Outcomes

Cultural bias is a key ethical concern in the CET-4, particularly in its reliance on Western-centric norms within reading passages and vocabulary tasks. Zhao (2022) and Liu et al. (2023) show how such content marginalises students from rural or minority backgrounds, whose cultural and linguistic realities remain underrepresented in test materials [15, 31]. Wang (2023) adds that while CET-4 writing tasks meet technical standards, they often fail to engage students' cultural identities or lived experiences, reducing the test's authenticity and relevance [26].

Messick's construct validity framework stresses the importance of assessing intended constructs in an equitable way across diverse populations [18]. Building on this, Paraskeva's (2018) theory of decolonisation advocates for test designs that incorporate locally relevant content, challenging hegemonic cultural assumptions [22].

Integrating culturally responsive content would reduce construct-irrelevant variance and improve inclusivity, allowing students to see their realities reflected in assessment tasks. By embracing these approaches, the CET-4 could evolve into a fairer, more representative assessment that values the diversity of its test-takers.

5. Reforms for an Inclusive CET

5.1. Redistributing Resources to Bridge Inequities

Persistent disparities in resource allocation between urban and rural schools present a major obstacle to fair implementation of the CET-4. While previous studies advocate for resource redistribution, implementing such reform requires a structured and evidence-based strategy. One promising model is Weighted Student Funding (WSF), which allocates resources based on individual student needs [20]. By linking funding to socioeconomic indicators, linguistic barriers, and geographic challenges, WSF ensures that schools serving marginalised populations receive equitable support.

In addition to financial measures, professional development focused on culturally responsive pedagogy can empower teachers to address students' varied needs better. Together, these dual interventions—needs-based funding and teacher development—offer a comprehensive approach to narrowing the socioeconomic divide in language education.

5.2. Introducing Modular and Regionalised Testing

A modular approach to CET-4 design offers a practical solution to accommodate the cultural and geographical diversity of test-takers while maintaining validity and standardisation. This model includes a central, standardised component to assess core skills, such as reading, writing, and listening, while incorporating optional modules tailored to specific cultural or regional contexts. For example, rural modules might focus on agricultural contexts, while urban modules could assess communication in commercial environments [13].

This approach reduces construct-irrelevant variance caused by unfamiliar cultural references [18]. It aligns with Paraskeva's epistemic pluralism by integrating region-specific knowledge and acknowledging diverse lived experiences [21, 22]. Modular testing enhances fairness and consequential validity by recognising cultural variation and promoting authenticity.

Successful implementation requires a phased, iterative approach. Pilot tests should be conducted across diverse regions, using mixed-method evaluation strategies. Quantitative tools such as Differential Item Functioning analysis (DIF) can detect scoring disparities, while qualitative feedback from students and educators can ensure contextual relevance and clarity [13].

Adaptive testing technology could further personalise real-time modules, improving accuracy and learner engagement [6]. Modular testing can complement reforms like performance-based assessment, emphasising real-world language use. Scenario-based writing and oral tasks—aligned with CEFR's communicative objectives—evaluate practical proficiency rather than rote recall [6].

Together, modular and performance-based testing offer a dual strategy for improving the validity and fairness of the CET-4. While modular testing supports cultural inclusivity and regional relevance, performance-based tasks reinforce real-world applicability. Challenges remain, such as maintaining psychometric comparability and ensuring sufficient resources for implementation. However, adopting a multi-pronged approach that integrates both formats can transform the CET-4 into a more equitable and context-sensitive assessment framework.

5.3. Collaborative Design with Stakeholder Input

Fairness audits represent a robust methodological tool for improving the CET-4's cultural inclusivity and equity. These audits systematically evaluate test content, administration procedures, and outcomes to identify potential sources of bias or structural disadvantage. The process often begins with DIF analysis, a statistical method used to detect whether specific test items yield different outcomes for different demographic groups, such as urban versus rural students.

Complementing statistical analysis, engaging diverse stakeholder panels—including teachers, students, and cultural experts—ensures test content reflects the lived experiences of a broad range of test-takers. This participatory design process surfaces potential biases and integrates feedback from underrepresented communities to refine test materials.

For instance, stakeholders may highlight how incorporating localised listening tasks or diverse English accents improves accessibility and authenticity for speakers of different varieties of English. By combining quantitative diagnostics with iterative feedback, fairness audits offer a practical pathway for aligning test design with principles of equity and validity [12].

5.4. Incorporating Real-world Performance Tasks

The CET-4's heavy focus on reading comprehension and vocabulary-based tasks has been criticised for narrowly defining English competence. Such a design often neglects key communicative skills like speaking, listening, and applied communication, undermining the test's objective of assessing practical English ability [6, 10]. Transitioning to performance-based assessments offers an opportunity to address these limitations and provide a more holistic evaluation of learners' abilities.

Scenario-based writing prompts are a key example. Rather than requiring abstract argumentative essays, such prompts could ask students to draft persuasive letters or respond to workplace communications—tasks that mirror real-world writing demands. Similarly, oral assessments—facilitated by recording technology—could evaluate speaking proficiency through simulated interviews or short presentations, reflecting authentic language use across settings [9]. For instance, test-takers could be asked to participate in simulated interviews or deliver short presentations, tasks that reflect authentic language use in diverse contexts.

Performance-based tasks advance equity by focusing on practical language ability rather than abstract knowledge or cultural familiarity with dominant norms. This shift reduces construct-irrelevant variance and ensures test outcomes reflect genuine proficiency rather than preparation strategies or cultural alignment. It also aligns with Paraskeva's epistemic pluralism, advocating for including varied knowledge systems and lived experiences [21, 22].

Additionally, these tasks support Messick's (1989) concept of consequential validity by fostering meaningful engagement with language learning [18]. Performance-based assessments promote transferable skills that enhance students' academic and career readiness. In this way, the CET-4 can evolve from a gatekeeping mechanism into a tool that expands opportunity and supports inclusive, real-world competence.

6. Conclusion

This paper has examined the multifaceted challenges of designing and implementing the CET-4, focusing on the interplay between validity, fairness, and cultural inclusivity in language assessment. Grounded in theoretical frameworks such as Messick's theory of validity [18], Kane's fairness principles [11], and Paraskeva's epistemic pluralism and itinerant curriculum theory [21, 22], the analysis emphasises the ethical implications of assessment design across diverse sociocultural contexts.

Findings underscore the importance of extending validation beyond technical precision to include social consequences and systemic impacts. Kane's fairness model and Messick's consequential validity reveal how CET-4's reliance on Western-centric norms undermines both its inclusiveness and credibility. Paraskeva's focus on stakeholder participation further reinforces the value of inclusive, context-responsive test design.

The CET-4's misalignment with CEFR benchmarks—particularly in assessing practical communication—exposes weaknesses in its current task formats. Test items often emphasise abstract, culture-specific knowledge, disadvantaging rural, minority, or low-income students unfamiliar with such content. The conflict with the CEFR's task-oriented rationale highlights the urgent need for modifications to align international standards with the needs of local students.

To address systemic inequities, a twofold strategy is required: first, equalising access to resources, and second, embedding cultural relevance into learning objectives and test design. Strategies such as modular testing, performance-based assessment, and inclusive design principles offer practical ways to integrate fairness and validity [6, 15, 31]. Combining regional contextualisation with a universal modular framework enhances both cultural relevance and comparability. Performance tasks like writing and speaking better reflect real-world competence, while participatory approaches—such as involving marginalised narratives—challenge systemic bias and promote equity.

These reforms align with Messick's consequential validity and Kane's fairness principles, creating a testing system that upholds both technical soundness and ethical responsibility [18, 11]. Key questions remain: How can psychometric standards coexist with cultural justice? Can modular testing and participatory design overcome existing gaps? How can the CET align with global frameworks like the CEFR while remaining locally relevant? Ongoing research and policy innovation will be vital in transforming the CET from a gatekeeping mechanism into an equitable platform that fosters opportunity and social justice.

Abbreviations

CET	College English Test
CET-4	College English Test Band 4
CET-6	College English Test Band 6
CEFR	Common European Framework of Reference for Languages
DIF	Differential Item Functioning
WSF	Weighted Student Funding

Author Contributions

Zhi Ma is the sole author. The author read and approved the final manuscript.

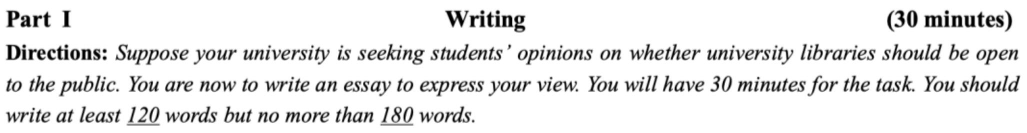
Conflicts of Interest

The authors report there are no competing interests to declare.

Appendix

Accurate Exam Items from CET-4 in June 2024, the resource is from https://zhenti.burningvocabulary.cn

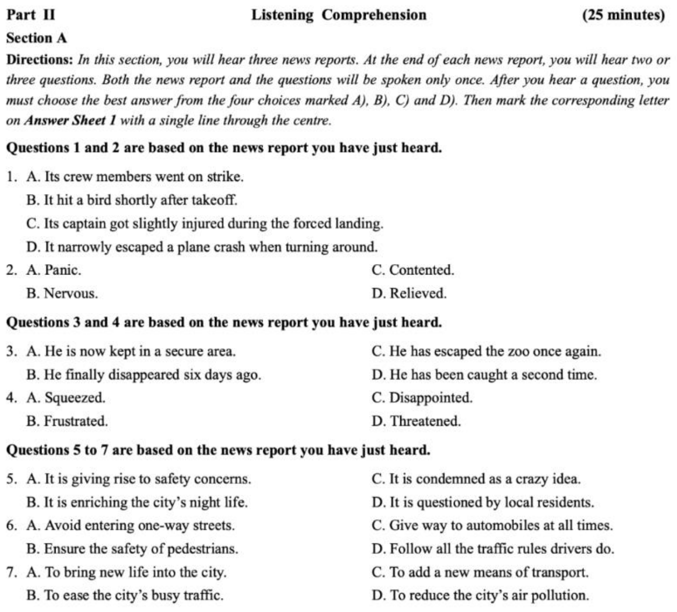
Paper 1



Part I **Writing** **(30 minutes)**

Directions: *Suppose your university is seeking students' opinions on whether university libraries should be open to the public. You are now to write an essay to express your view. You will have 30 minutes for the task. You should write at least 120 words but no more than 180 words.*

Figure 1. *Sample Writing of Paper 1.*



Part II **Listening Comprehension** **(25 minutes)**

Section A

Directions: *In this section, you will hear three news reports. At the end of each news report, you will hear two or three questions. Both the news report and the questions will be spoken only once. After you hear a question, you must choose the best answer from the four choices marked A), B), C) and D). Then mark the corresponding letter on* ***Answer Sheet 1*** *with a single line through the centre.*

Questions 1 and 2 are based on the news report you have just heard.

1. A. Its crew members went on strike.
   B. It hit a bird shortly after takeoff.
   C. Its captain got slightly injured during the forced landing.
   D. It narrowly escaped a plane crash when turning around.

2. A. Panic. C. Contented.
   B. Nervous. D. Relieved.

Questions 3 and 4 are based on the news report you have just heard.

3. A. He is now kept in a secure area. C. He has escaped the zoo once again.
   B. He finally disappeared six days ago. D. He has been caught a second time.

4. A. Squeezed. C. Disappointed.
   B. Frustrated. D. Threatened.

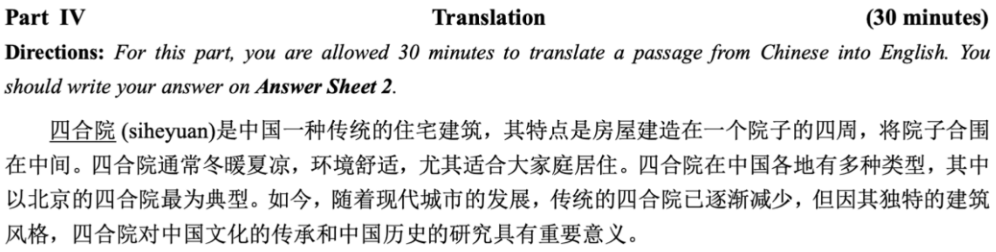
Questions 5 to 7 are based on the news report you have just heard.

5. A. It is giving rise to safety concerns. C. It is condemned as a crazy idea.
   B. It is enriching the city's night life. D. It is questioned by local residents.

6. A. Avoid entering one-way streets. C. Give way to automobiles at all times.
   B. Ensure the safety of pedestrians. D. Follow all the traffic rules drivers do.

7. A. To bring new life into the city. C. To add a new means of transport.
   B. To ease the city's busy traffic. D. To reduce the city's air pollution.

Figure 2. *Sample Listening of Paper 1.*



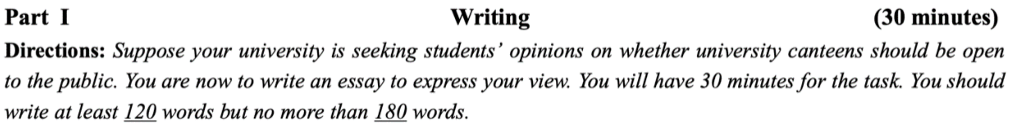
Part IV **Translation** **(30 minutes)**

Directions: *For this part, you are allowed 30 minutes to translate a passage from Chinese into English. You should write your answer on* ***Answer Sheet 2****.*

四合院 (siheyuan)是中国一种传统的住宅建筑，其特点是房屋建造在一个院子的四周，将院子合围在中间。四合院通常冬暖夏凉，环境舒适，尤其适合大家庭居住。四合院在中国各地有多种类型，其中以北京的四合院最为典型。如今，随着现代城市的发展，传统的四合院已逐渐减少，但因其独特的建筑风格，四合院对中国文化的传承和中国历史的研究具有重要意义。

Figure 3*. Sample Translation of Paper 1.*

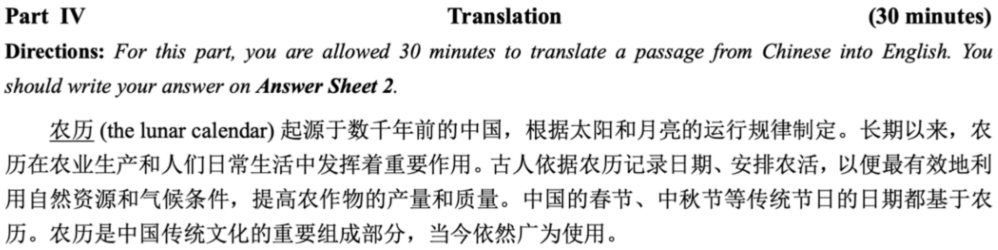
Paper 2



Part I **Writing** **(30 minutes)**

Directions: *Suppose your university is seeking students' opinions on whether university canteens should be open to the public. You are now to write an essay to express your view. You will have 30 minutes for the task. You should write at least 120 words but no more than 180 words.*

Figure 4*. Sample Writing of Paper 2.*



Part IV **Translation** **(30 minutes)**

Directions: *For this part, you are allowed 30 minutes to translate a passage from Chinese into English. You should write your answer on* ***Answer Sheet 2****.*

农历 (the lunar calendar) 起源于数千年前的中国，根据太阳和月亮的运行规律制定。长期以来，农历在农业生产和人们日常生活中发挥着重要作用。古人依据农历记录日期、安排农活，以便最有效地利用自然资源和气候条件，提高农作物的产量和质量。中国的春节、中秋节等传统节日的日期都基于农历。农历是中国传统文化的重要组成部分，当今依然广为使用。

Figure 5*. Sample Writing of Paper 2.*

References

[1] Adamson, B., & Xia, B. (2011). A case study of the College English Test and ethnic minority university students in China: negotiating the final hurdle. Multilingual Education, 1(1), 1. https://doi.org/10.1186/2191-5059-1-1

[2] Bai, Y. (2020). The relationship of test takers' learning motivation, attitudes towards the actual test use and test performance of the College English Test in China. Language Testing in Asia, 10(1), 10.

[3] Bryan, M., & Lewis, A. (2019). Culturally Responsive Evaluation as a Form of Critical Qualitative Inquiry. Oxford Research Encyclopedia of Education. https://doi.org/10.1093/acrefore/9780190264093.013.545

[4] Chunyuan, N., Chwee Fang, N., Hazlina, A. H., & Mustapha, N. F. (2024). Translation Strategies of Chinese-English Culture-Specific Items: the Case of Translation Test from CET4 and CET6. AWEJ for Translation & Literary Studies, 8(1). https://doi.org/10.24093/awejtls/vol8no1.4

[5] Council of Europe. (2020). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages

[6] Fulcher, G., & Davidson, F. (2012). The Routledge handbook of language testing. Routledge New York, NY.

[7] Guo, A. (2006). The prospect of CET in China. Journal of Cambridge Studies, 1(2), 39-41.

[8] Han, F. (2021). Washback effects of the College English Teaching Band 4 Test in China and possible solutions. In B. Lanteigne, C. Coombe, & J. D. Brown (Eds.), Issues in language testing around the world: Insights for language test users. Springer. https://doi.org/10.1007/978-981-33-4232-3_4

[9] Harding, L. (2014). Communicative Language Testing: Current Issues and Future Research. Language Assessment Quarterly, 11, 186-197. https://doi.org/10.1080/15434303.2014.895829

[10] Heydarnejad, T., Tagavipour, F., Patra, I., & Farid Khafaga, A. (2022). The impacts of performance-based assessment on reading comprehension achievement, academic motivation, foreign language anxiety, and students' self-efficacy. Language Testing in Asia, 12(1), 51. https://doi.org/10.1186/s40468-022-00202-4

[11] Kane, M. (2010). Validity and fairness. Language Testing, 27(2), 177-182. https://doi.org/10.1177/0265532209349467

[12] Kyllonen, P., & Sevak, A. (2024). Charting the Future of Assessments. https://www.ets.org/Rebrand/pdf/FoA_Full_Report.pdf

[13] Leighton, J., & Gierl, M. (Eds.). (2007). Cognitive Diagnostic Assessment for Education: Theory and Applications. Cambridge University Press. https://doi.org/10.1017/CBO9780511611186

[14] Li, H., Meng, L., Mu, K., & Wang, S. (2024). English language requirement and educational inequality: Evidence from 16 million college applicants in China. Journal of Development Economics, 168, 103271. https://doi.org/10.1016/j.jdeveco.2024.103271

[15] Liu, H., Shi, X., Qiu, J., Shi, Y., Hao, Y., Zhu, L., Yan, C., & Li, H. (2023). Academic word coverage and language difficulty of reading passages in College English Test and Test of English for Academic Purposes in China [Original Research]. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1171227

[16] Mathew, R. (2004). Stakeholder Involvement in Language Assessment: Does it Improve Ethicality? Language Assessment Quarterly: An International Journal, 1, 123-135. https://doi.org/10.1080/15434303.2004.9671780

[17] McNamara, T., & Ryan, K. (2011). Fairness Versus Justice in Language Testing: The Place of English Literacy in the Australian Citizenship Test. Language Assessment Quarterly, 8(2), 161-178. https://doi.org/10.1080/15434303.2011.565438

[18] Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-104). American Council on education and Macmillan.

[19] Montenegro, E., & Jankowski, N. A. (2017). Equity and assessment: Moving towards culturally responsive assessment. Occasional Paper, 29(6), 10-11. https://learningoutcomesassessment.org/wp-content/uploads/2019/02/OccasionalPaper29.pdf

[20] OECD. (2017). The Funding of School Education. https://doi.org/10.1787/9789264276147-en

[21] Paraskeva, J. (2011). Conflicts in Curriculum Theory. Challenging Hegemonic Epistemologies. https://doi.org/10.1057/9780230119628

[22] Paraskeva, J. (2018). Against the scandal: itinerant curriculum theory as subaltern momentum. Qualitative Research Journal, 18(2), 128-143. https://doi.org/10.1108/QRJ-D-18-00004

[23] Peng, C., Liu, J., & Cai, H. (2022). Aligning China's Standards of English Language Ability with the Common European Framework of Reference for Languages. The Asia-Pacific Education Researcher, 31(6), 667-677. https://doi.org/10.1007/s40299-021-00617-2

[24] Schneider, E. W. (2020). Developmental patterns of English: Similar or different? The Routledge handbook of world Englishes, 408-421.

[25] Wang, C. (2014). Communicative validity of the new CET-4 listening comprehension test in China. Indonesian Journal of Applied Linguistics, 4, 111. https://doi.org/10.17509/ijal.v4i1.608

[26] Wang, Z. (2023). Language Assessment Instrument Analysis: Scoping and Investigating the Writing Assessment of College English Test Band 4. Journal of Education and Educational Research, 3, 33-36. https://doi.org/10.54097/jeer.v3i2.9012

[27] Yan, J., & Huizhong, Y. (2006). The English Proficiency of College and University Students in China: As Reflected in the CET. Language, Culture and Curriculum, 19(1), 21-36. https://doi.org/10.1080/07908310608668752

[28] Yao, D. (2023). Examining the subjective fairness of at-home and online tests: Taking Duolingo English Test as an example. PLOS ONE, 18(9), e0291629. https://doi.org/10.1371/journal.pone.0291629

[29] Ying, Z., & Liying, C. (2008). Test review: College English Test (CET) in China. Language Testing, 25(3), 408-417. https://doi.org/10.1177/0265532208092433

[30] Zhang, J. (2023). Universities scrapping English test requirements. Chinadaily.com.cn. https://govt.chinadaily.com.cn/s/202309/26/WS6514dec6498ed2d7b7e9c978/universities-scrapping-english-test-requirements.html

[31] Zhao, H. (2022). Quality Analysis of an English Test Designed against the Framework of China’s Standards of English Language Ability. Proceedings of the 2022 International Conference on County Economic Development, Rural Revitalization and Social Sciences (ICCRS 2022). https://doi.org/10.2991/aebmr.k.220402.015