TYPE Original Research
PUBLISHED 10 January 2024
DOI 10.3389/feduc.2023.1275518

Check for updates

OPEN ACCESS

EDITED BY
Janet Clinton,
The University of Melbourne, Australia

REVIEWED BY
Xiaodong Zhang,
Beijing Foreign Studies University, China
Fidel Çakmak,
Alanya Alaaddin Keykubat University, Türkiye

*CORRESPONDENCE
Haomin Zhang
✉ hmzhang@english.ecnu.edu.cn

RECEIVED 28 August 2023
ACCEPTED 01 December 2023
PUBLISHED 10 January 2024

CITATION
Ma X and Zhang H (2024) Predictability of Duolingo English mock test for Chinese college-level EFLs: using assessment use argument.
Front. Educ. 8:1275518.
doi: 10.3389/feduc.2023.1275518

Predictability of Duolingo English mock test for Chinese college-level EFLs: using assessment use argument

Xinyi Ma and Haomin Zhang*

School of Foreign Languages, East China Normal University, Shanghai, China

Online international language tests have been gaining popularity since the COVID-19 pandemic. This study explored the predictive ability of the mock test of Duolingo English Test (DET) for Chinese students. College-level EFLs ($N = 42$) participated in this study and completed the official online mock test. Through quantitative analysis, the findings demonstrated that mock scores were significantly correlated with TOEFL or IELTS results. However, given that the mock test scores are significantly lower than expected, the consequential validity needs further analysis. Using Assessment Use Argument in qualitative analysis, the study established the claim, warrants, evidence, and potential rebuttals regarding the consequential validity of Duolingo English mock test. With further analysis on test specification, the characterization of Target Language Use domain of DET reading test items were investigated. The results indicate that although the DET mock test is relatively more accessible, it needs modification to become more suitable for Chinese EFLs. Suggestions and limitations on Chinese students using DET and its mock test were also discussed.

KEYWORDS

Duolingo English test, assessment use argument, consequential validity, target language use domain, Chinese EFLs

1 Introduction

The outbreak of the COVID-19 pandemic led to the suspension of many TOEFL and IELTS offline test centers in China, causing a significant number of students failing to provide their English test results in a timely manner and missing opportunities to study abroad. To address this issue, some higher institutions have accepted Duolingo English Test (DET) results to compensate for other standardized language tests. The DET is a more affordable online test that utilizes computerized adaptation to test-takers' language proficiency. The score report can be obtained within 48 h, making it an attractive option for Chinese students applying to study abroad.

Although the DET was launched in 2014 and revised in 2019, it seemed to only gain recognition in 2020 because of the pandemic and the popularity of online platforms. However, there are controversies centering around its consequential validity. While the DET may be convenient and cost-effective, some institutions do not recognize its validity as TOEFL and IELTS, which are widely considered authoritative standardized tests for international students seeking admission to overseas universities (Bézy and Settles, 2015). Therefore, it is important for students to research and verify the recognition of DET scores by their desired institutions before taking the test.

This study aims to provide a comprehensive analysis of the DET scores of Chinese candidates and evaluate the test's validity in assessing English language proficiency for university admission. To achieve this goal, the study utilizes Bachman's Assessment Use Argument (AUA) model, a test development and evaluation framework commonly used in the evaluation of TOEFL. Specifically, the study evaluates the overall design of the official online mock test of the DET, and focuses on analyzing the construct definition, test characteristics, the relationship to Target Language Use (TLU) domain, and interpretation of the test-taker's response performance.

2 Literature review

2.1 Language test validity

The study on language test validity can be traced back to the early 20th century, when researchers sought to determine whether test results could accurately reflect language proficiency of the test-takers (Cronbach and Meehl, 1955). Initially, the focus was on the consistency and reliability of tests, with limited research on their validity. As language testing evolved and its applications expanded, the study of test validity has gained attention. Test validity now refers to the extent to which evaluation results provide appropriate, meaningful, and useful inferences for evaluation purposes (Gronlund, 1998). It can be understood as the quality or acceptability of the test (Chapelle, 1999). Kane (1992, 2001) argues that validity arguments should be based on critical analysis of sufficient evidence to support or refute hypotheses or proposed explanations.

Since the 1950s, the Toulmin Model has contributed to the framework for demonstrating the validity of language testing. It aims to help people analyze and establish effective arguments, including data, claims, reasons, warrants, rebuttals, and backups, emphasizing logic, empirical and rational elements. With this model, one can better understand and analyze the structure of arguments, discover fallacies and weaknesses in them, and build more effective arguments (Toulmin, 2003).

Combining this model, Kane further elaborates the concepts, types, and methods to evaluate validity, and points out factors of test purpose, content, and interpretation of test results, emphasizing the clarity, consistency and plausibility of arguments and conceptions (Kane, 2012).

Also combining the Toulmin Model, especially the six key elements (claim, grounds, warrant, backing, qualifier, and rebuttals), Bachman proposes Assessment Use Argument (AUA) that links assessment performance to assessment use, applicable to a variety of situations and purposes of language testing practice in the 21st century (Bachman, 2005; Bachman and Palmer, 2010). Suitable for the trial stage of test development and use, the model can make four types of claims, namely consequence, interpretation, decision, and consistency, each of which has corresponding warrants, backings, and potential rebuttals for the exam, as is shown in Figure 1.

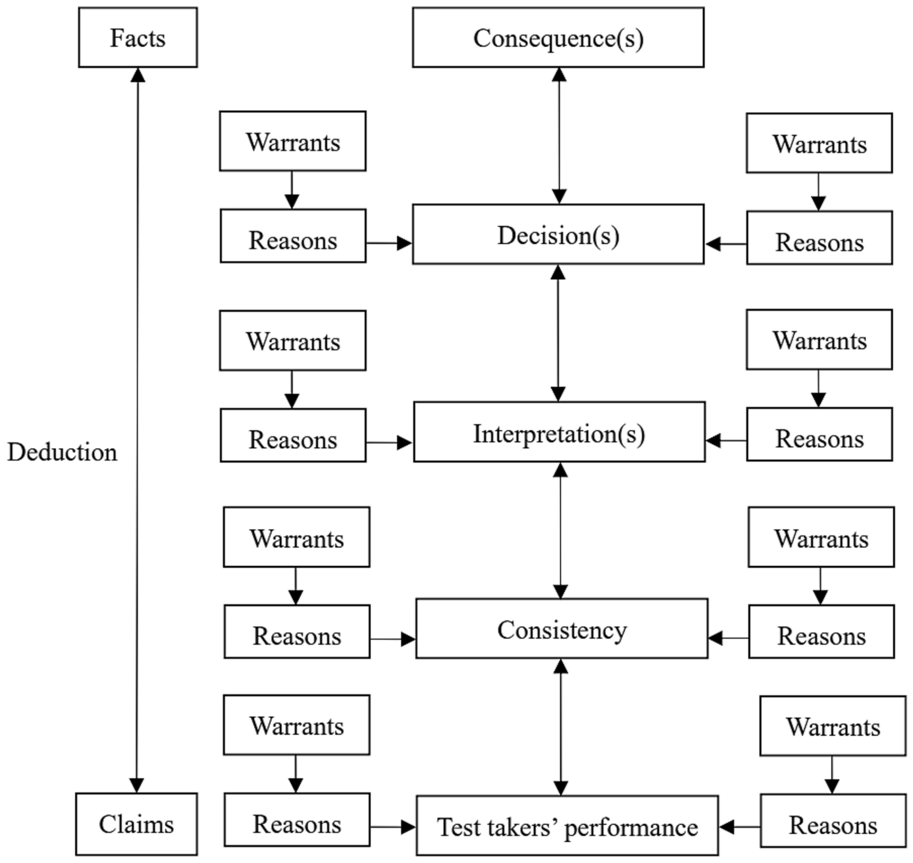

FIGURE 1
Assessment use argument based on Bachman and Palmer (2010).

According to the AUA model, to support the claims regarding consequences, one needs evidence for two types of validity, namely face validity and outcome validity (Bachman and Palmer, 2010). The former refers to whether the test is considered fair, relevant to the test-taker's language ability, and helpful to the test-taker's language learning; the latter refers to the impact of the test and the impact of the test on teaching and learning (Abeywickrama and Brown, 2010).

As one of the tools in test assessment, AUA model is widely applied in education, psychology, and linguistics to record student progress, measure L2 learners' performance, and/or diagnose improper learning patterns (Bachman and Palmer, 1996). It justifies arguments for test assessment, evaluates language curriculum development and guides L2 instructions in the classroom. Most importantly, it can be used in both international language tests, such as the Pearson Test of English Academic, TOEFL and IELTS, or Chinese language tests, such as College English Tests and various in-class second language tests (Wang et al., 2012). Various studies have evaluated language tests from different perspectives: vocabulary acquisition can be assessed in classroom settings (Waluyo, 2018) and online mobile applications (Schmitt et al., 2020). It is shown that more rigorous vocabulary tests can effectively enhance L2 learners' acquisition. To analyze speaking proficiency, some studies showed that there were different learning opportunities in different test frameworks by comparing the performance of Spanish learners in two conversation test items (Pardo-Ballester, 2020). To analyze the effectiveness of the listening grading test, researchers grouped learners according to their language proficiency level and tested it (Pardo-Ballester, 2010). In assessing and predicting speaking and listening proficiency, a new assessment method with a variety of descriptive, graphical, and inferential statistical techniques was used to investigate the competencies for different speaking skills and predict learners' intermediate listening skills (Mozgalina, 2015).

2.2 Assessment for Duolingo English test

2.2.1 Consequential validity

The assessment of a candidate's language proficiency should accurately reflect their ability to effectively convey ideas through language in specific environment with the targeted language use (Isbell and Kremmel, 2020). While traditional English language tests typically measure a candidate's reading, listening, speaking, and writing abilities, these skills are not used in isolation in practical applications. The effective use of multiple language skills is integral to natural and effective communication. The DET classifies a candidate's reading and writing abilities as Literacy, while Comprehension is determined by their listening and reading skills. Production is assessed through a combination of a candidate's writing and speaking abilities, while Conversation measures their competence in both listening and speaking.

The DET offers a convenient and accessible option for assessing English proficiency, as it does not require an appointment and can be taken online from any location. The test utilizes computerized adaptation to assign questions of appropriate difficulty based on each candidate's previous responses, ensuring a tailored and efficient testing experience. The large question pool and adaptive format effectively prevent exam fraud and leaks, as repeated test questions only appear after a candidate has taken more than 1,000 real exams. However, it may introduce biases when students from various cultural and educational background have different definitions regarding proficiency levels.

The formal test comprises task types like Interactive Reading, Read and Complete (C-test), Yes/No Vocabulary, Dictation, Writing Sample, Speaking Sample, and so on. The average completion time for each type of question is approximately 16 min, with the longest test time being 20 min and overall test duration being approximately 2 h. Its official online mock test, on the other hand, has shorter test lengths of approximately 6–10 min per question type and a total duration of less than 1 h. The DET score report is sent free of charge within two business days of completing the test. Generally speaking, the test places emphasis on natural and fluent language use, effective expression, and the simultaneous use of multiple language skills. Its goal is to provide a comprehensive assessment of language proficiency that is reflective of real-world language use (Cardwell et al., 2022).

The DET is increasingly recognized by foreign universities as a reliable reference standard for admission, and recent studies have demonstrated its academic value in assessing language ability. Scholars have extrapolated from the Common European Framework of Reference (CEFR) and found a significant correlation between Duolingo English test scores and TOEFL and IELTS scores, thereby confirming its academic reliability (Council of Europe, 2001; Verhelst et al., 2009; Bézy and Settles, 2015). Furthermore, logistic regression analysis has shown that the effectiveness of computer-adaptive test questions in the Duolingo English Test is not affected by individual differences among candidates, allowing for accurate measurement of English proficiency across diverse student populations (Maris, 2020). These findings underscore the growing importance of the Duolingo English Test in academic settings and highlight its potential as a valid and accessible measure of language ability.

Despite the growing recognition of the DET, some schools and scholars maintain reservations regarding its validity and practicality. In Wagner's (2020) critical commentary on the test, he concluded that using DET scores as a reference indicator for college admission is not recommended due to several disadvantages, such as the lack of correlation between test content and the context and objectives of university learning, which makes it difficult to evaluate candidates' pragmatic, discourse processing, and interaction abilities. Wagner also noted that the test has the potential to have a negative impact on test-takers and their learning systems. Additionally, Isbell and Kremmel (2020) argue that the content of DET primarily focuses on the psycholinguistic perspective of language knowledge and processing, and its correlation with other academic test scores remains inadequate. These critiques suggest that AUA model, which provides a framework for analyzing the validity and practicality of tests, has not been fully applied to DET. The AUA offers an explicit logical structure for evaluating the links in the argument and makes explicit the relationship between the validity and utilization arguments (Llosa, 2008; Mann and Marshall, 2010; DeBarger et al., 2016). Despite limitations in accessing the content of the actual test due to copyright issues, this paper aims to evaluate the consequential validity and specific test performance in the official mock test with AUA model.

Language ability plays a crucial role in L2 learners' academic success and future enrollment opportunities. As such, it is essential to gain a deeper understanding of how college admissions management systems make decisions regarding English proficiency tests. Higher education institutions are complex organizations, and students' selection of a foreign school is one of the most significant decisions they will make in their lives. However, high-stakes decisions are often made under conditions where the available information is insufficient. If methods of proving English proficiency are unreliable or ineffective, it can negatively impact both institutions and students. Therefore, the validity of English proficiency test results is essential to the decision-making process of college admissions management systems. Accurate assessment of students' language abilities is critical to predict their potential academic performance and overall life skills.

2.2.2 Specific test items analysis: reading tests

Test items are the fundamental unit of any language proficiency test. Bachman argues that the design of test tasks should be based on AUA model to ensure that the tasks are reasonable and practical. To assess the rationality and usefulness of a test item, it is important to consider the Target Language Use (TLU) domain and the normativity of the task (Bachman and Palmer, 2010). This paper selected reading test items from the DET for analysis due to several reasons. First, the reading questions are generated by GPT-3 and use a standardized structure and principles, with original materials obtained from publicly available sources such as textbooks and free novels (Park et al., 2022). Therefore, the answers to these questions can be verified through public webpages. Meanwhile, the website has a strict monitor system to ensure that the test-takers cannot look up on the internet while testing, but they can find most of the standard answers afterwards. Second, the test questions are mainly developed by computers, with little manual review or evaluation of the construction of a demonstration model in the design process (Settles et al., 2020). Third, other test items do not have precise and complete answers available on the internet, making them less suitable for specific performance analysis. The choice of the reading test items in this study provides a valuable opportunity to evaluate the validity and reliability of DET in a targeted and rigorous manner.

The constructs of L2 reading often include conceptualized reading based on cognitive processes (Alderson, 2000; Khalifa and Weir, 2009), reading for purpose (Britt et al., 2018), and texts in the domain of TLU (Green et al., 2010). DET combines the first two perspectives and envisages the idea of reading based on the purpose of the test-taker's reading and the cognitive processes used in reading (Chapelle, 1999), all of which are relevant to the academic context.

Various reading models are associated with different response patterns to questions. In traditional reading tests, multiple-choice patterns have been found to be highly correlated with partial reading ideas and can predict reading proficiency more effectively (Riley and Lee, 1996; Alderson, 2000; Grabe and Jiang, 2014). Nevertheless, in DET, digital-first assessment employs technology to construct multiple response patterns simultaneously, which is challenging to achieve in paper-based tests.

For instance, *Interactive Reading* in DET utilizes the highlighting strategy, requiring test-takers to identify and mark sentences that can properly respond to the test questions. Commonly used by college students in language proficiency tests, this strategy can promote the test-takers to recall the content while reading. Research shows that highlighting strategy and behaviors can indicate reading ability and comprehension levels (Blanchard and Mikkelson, 1987; Winchell et al., 2020) as well as benefiting knowledge acquisition from reading. In essence, what students highlight in a test reveals their knowledge and understanding of the reading materials. Such response pattern is not only innovative but also help to establish the concept of L2 reading more effectively (Bachman and Palmer, 1996; Qian and Pan, 2013).

Another type of reading test items in DET, *C-test*, can measure a test-taker's reading ability standard as well (Khodadady, 2014). Research statistics suggest that C-test is closely related to many other authoritative language proficiency tests, such as TOEFL, in detecting language and especially spelling skills.

At present, few scholars have studied the correlation between the Duolingo English mock test and TOEFL or IELTS scores. Moreover, AUA model is not fully applied to evaluate the validity of the test results, the TLU domain of the reading test items, or the test task specification.

Based on the works above, this research may answer two questions:

1. Does Duolingo English mock test serve as a valid and reliable English academic proficiency test for Chinese college EFL students?
2. To what extent does it predict English proficiency among Chinese college EFL students?

3 Method

3.1 Data processing

This study utilizes a quantitative methodology to address the first research question. Specifically, the project recruits test-takers to complete the online Duolingo English mock test, and employs the score conversion rules provided on the official website to convert TOEFL, IELTS, and mock test scores into a unified standard. The study then uses unitary regression analysis and paired-sample t-tests to explore the correlation between the two types of test results.

In total, 42 Chinese mainland college students who had not taken the DET or its mock tests in multiple countries were recruited for the study. Ultimately, 35 test-takers were able to complete the mock test, having taken the TOEFL or IELTS offline test within the past two years and possessing valid transcripts. The participants underwent two Duolingo English Test mock exams, which were conducted one week apart in a standardized testing environment using online private meeting rooms in the Tencent Meetings app. Given that the test-takers were totally unfamiliar with DET test items, only the results of the second tests were recorded and used in following data processing to ensure credibility. Furthermore, to compare the academic validity of the Duolingo English mock test in predicting the English language level of Chinese college students, the entire process of the mock test was recorded with the consent of the test-taker. The mock test scores and TOEFL and IELTS scores of the participants were analyzed in SPSS software and paired-sample t-tests to determine whether there was a significant correlation between the offline tests, TOEFL and IELTS, and the online Duolingo English mock tests.

3.2 AUA model

Given the past coronavirus pandemic, this study seeks to determine whether Chinese college-level students should prepare for and then take DET to assess and prove their English proficiency before furthering their study abroad. To accomplish this goal, the study draws upon the framework proposed by Bachman and Palmer (2010), who argue that test developers should incorporate detailed AUA models for each intended use of testing. By fully utilizing the rationales in AUA to clarify the intent of the test through assessments, developers can more accurately interpret test-taker performance, measure language ability, and enable other stakeholders to make decisions based on scores.

Therefore, the second research question in this study is addressed through a qualitative approach. Drawing upon questions and controversies raised in existing literature, this project utilizes AUA model to evaluate the design and specification of Duolingo English mock test. The model includes hypotheses and corresponding rebuttals to analyze various aspects of the test, such as consequential validity, test-takers' performance and its educational impact.

Using AUA, the present study puts forth claims, warrants, and evidence to evaluate the Duolingo English mock test in terms of its suitability for measuring the English proficiency of Chinese college-level students and whether it meets the standards of academic language testing. To evaluate the validity of test results, the study utilizes the consequential validity in AUA model, putting forth corresponding claims, warrants, and potential rebuttals to assess the impact of test results. Originally, the model is supposed to include decisions, interpretations and consistency analysis regarding the test. Nevertheless, since this study does not have access to decision making processes in higher education institution or test items in authentic DET, consequential validity and specification on reading tests will be the focus. The adapted AUA model is shown in Figure 2.

4 Quantitative analysis

4.1 Logistic regression analysis

To answer the first research question, this project explores the relationship between Duolingo English Test mock test scores and TOEFL and IELTS scores. To this end, results of TOEFL, IELTS and Duolingo English mock test from 42 test-takers were analyzed as follows.

First, IELTS scores were converted to TOEFL scores based on official scoring rules (https://www.englishtest.duolingo.com/institutions/scores). Unlike the result of a real DET, the mock test only showed approximately estimated score bands with different ranges between the maximum and minimum. Thus, the scores of 7 test-takers were removed because the highest and lowest scores differed by more than 80 out of 160 points. For the remaining 35 mock test scores, the median of each score band was calculated and compared with TOEFL scores.

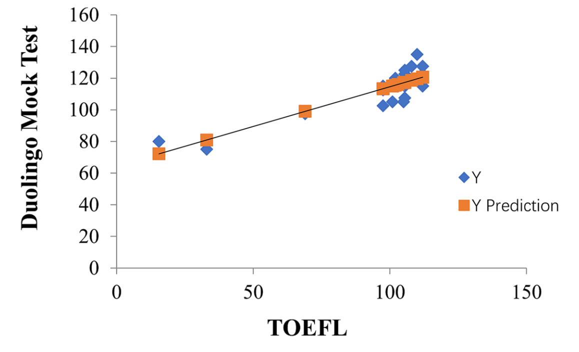

FIGURE 2
Adapted AUA model for consequential validity.

According to the unitary linear regression analysis in SPSS, the regression coefficient value is about 0.502. Further correlation coefficient significance test shows $p < 0.001$, proving that the mock test scores are significantly positively correlated with TOEFL and IELTS scores, as shown in Figure 3.

This result showed that the Duolingo English mock test score could be used as a reference for evaluating TOEFL and IELTS scores. It is worth noting that this study only considered the data of a small number of test-takers, so the sample size needs to be further expanded to better understand the relationship between the Duolingo English mock test results and TOEFL and IELTS scores. In addition, it should be taken into consideration that adding other testing tools to assess English proficiency to assess the candidate's English ability more comprehensively.

4.2 Paired-samples *t*-test

In order to further explore whether the Duolingo English mock test is beneficial to promote language learning for test-takers, this project continued to use the matching sample t-test to determine the impact of using the test on student performance before and after. The results showed that after using the Duolingo English mock test, the test-takers' performance showed a difference at a significance level of 0.01. Further comparing the mean, it was found that students using the test scored lower than TOEFL and IELTS scores, as detailed in Table 1.

This result suggests that the Duolingo English mock test has limited effect on improving student achievement. To better promote the language learning of the test-takers, we recommend combining a variety of teaching methods and tools in actual teaching to achieve the best teaching results. At the same time, English tests should be carefully selected according to the actual situation of test-takers, and strengthen the interpretation and analysis of test results to help students better improve their English.

5 Qualitative analysis: AUA

Based on the above quantitative analysis, it is reasonable to deduce that the Duolingo English mock test can distinguish the levels of college-level test-takers with their language proficiency to a certain extent. However, its overall design and test items seem to hinder the test-takers from performing at their highest level. Therefore, in order to answer the second research question, this following part combines the AUA model to analyze consequential validity of mock test results, the TLU domain and characterization of reading test items.

5.1 Consequence

Claim: Duolingo English mock test results will influence the decision of test-takers, especially Chinese college-level EFL who will

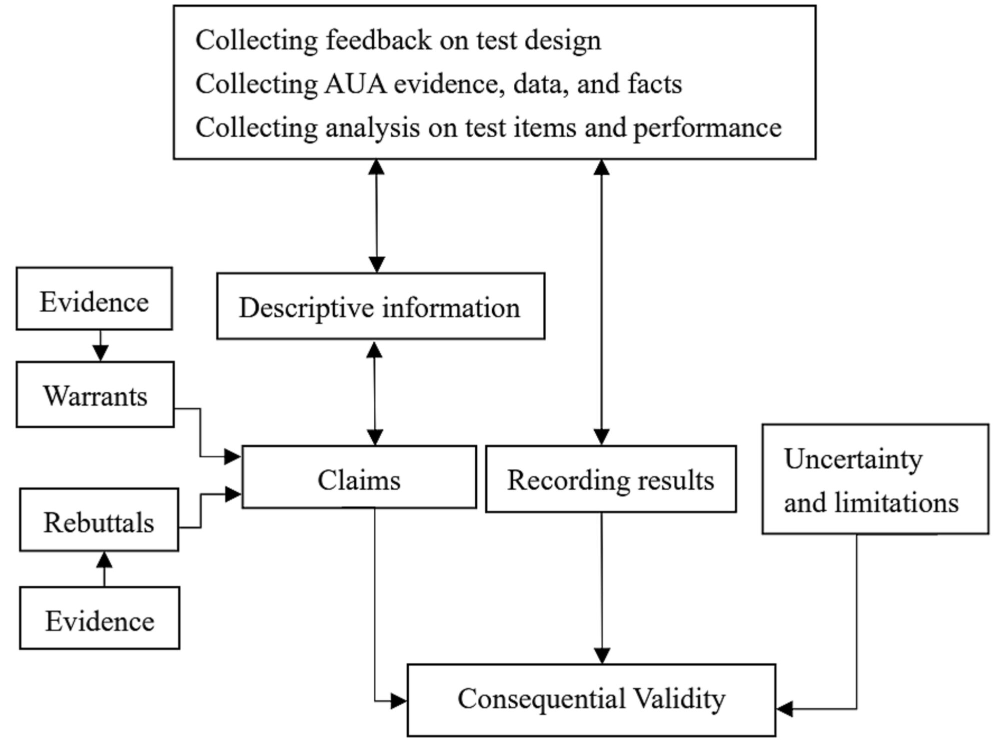
Collecting feedback on test design
Collecting AUA evidence, data, and facts
Collecting analysis on test items and performance
Evidence
Descriptive information
Warrants
Claims
Recording results
Uncertainty and limitations
Rebuttals
Evidence
Consequential Validity

FIGURE 3
Linear correlation between Duolingo Mock test and TOEFL scores.

TABLE 1 Paired-sample *T*-test Before and after Duolingo English Mock test.

	t	*df*	Sig. (2-tailed)
Duolingo Mock Test – TOEFL	−8.802	34	0.000**

decide whether to continue to prepare for the Duolingo English test based on the mock test results, or to use the results to apply to foreign universities, etc.

Stakeholders: test-takers (Chinese college-level EFL), teachers, schools, and other test users.

Warrant 1: Test-takers prefer using Duolingo English mock test result to evaluate their language proficiency and apply to foreign universities for studying abroad.

Evidence: After conducting a mock test, the test-takers will compare the results with those of established standardized tests, such as TOEFL and IELTS, to make informed decisions based on the unique characteristics of each test. The Duolingo English test offers a free official mock test with a user-friendly interface that can be completed online at any time. The computer-adaptive question writing and scoring functionality addresses the limitations of other standardized language tests, which require manual review for speaking and writing components (Wagner, 2020). Therefore, the Duolingo English mock test provides candidates with a total score segment, while other tests only offer scores for reading and listening. Additionally, the Duolingo test registration fee is significantly lower, costing only $49 (approximately 350 CNY) per test, compared to over $281 (more than 2,000 CNY) per test for TOEFL or IELTS registration. This cost difference makes the Duolingo English test a more attractive option for students seeking a language score for studying abroad. Overall, these factors make the Duolingo English test a viable alternative to traditional standardized tests, providing students with a convenient and cost-effective means of demonstrating their language proficiency.

Potential rebuttals 1: The current study highlights two key concerns regarding the Duolingo English mock test. On the one hand, the results cannot accurately evaluate language proficiency. On the other hand, test-takers may be negatively influenced by the score rages.

Evidence: Firstly, the test results lack targeted suggestions, and there is no direct evidence that Chinese students, or other stakeholders, benefit from taking the test. Although the free mock test provides an estimated score range, it fails to provide specific language proficiency feedback on different user groups, especially Chinese college-level EFL. Additionally, the mock test scores lack detailed information about the test-taker's language proficiency. As the speaking and writing components are initially judged by computers, the mock test scores can only fall within a specific score range, which varies by 15–60 points. Test-takers would have no idea about their specific performance in each type of test items. In contrast, TOEFL and IELTS offer accurate scores on reading and listening mock test sections, providing test-takers with more detailed feedback on their proficiency levels in these areas.

Secondly, the Duolingo English mock test scores may cause stress and anxiety, as the test questions are challenging, and the ranges of scores can vary significantly. This can lead to poor learning outcomes, for test-takers may focus too much on achieving high scores rather than the essential purpose of language learning.

Therefore, these concerns highlight the need for further research into the consequential validity of the Duolingo English mock test, including its ability to provide targeted feedback and its impact on test-takers' stress levels. Such research could aid in improving the test's design and its overall effectiveness in assessing language proficiency.

Warrant 2: Test-takers would use test items from Duolingo English mock test to improve language proficiency.

Evidence 2: The Duolingo English mock test questions provide a valuable tool for test-takers to self-assess their language proficiency, motivating effective learning and improving language scores. Firstly, the free practice test offers unlimited opportunities for practicing English literacy, comprehension, output, and communication. The mock test questions allow students to apply their listening, speaking, reading, and writing skills comprehensively, enhancing their language proficiency in a well-rounded manner.

Secondly, the Duolingo English mock test questions offer a challenging experience that mirrors the actual test. With a variety of question types, fast response speed, and high overall difficulty, the mock test stimulates test-takers' motivation to learn actively and improve their language levels. By providing an accurate representation of the actual test experience, the Duolingo English mock test helps students prepare more effectively for the real test.

Furthermore, the Duolingo English practice test questions allow students to assess their learning outcomes, identify areas of weakness, and adjust their learning strategies accordingly. The test results offer feedback that can guide teachers to teach more effectively and provide further support to students.

Therefore, the Duolingo English mock test questions provide a comprehensive and challenging tool for test-takers to self-assess their language proficiency, motivating effective learning, improving language scores, and enhancing overall language ability.

Potential Rebuttals 2: DET lacks test items on pragmatics or interactions.

Evidence: Some studies have already questioned whether DET does not have sufficient pragmatically-flavored items to prepare students for studying abroad on campus. First, compared to other standardized language tests, the content of the DET has little to do with the student's future university life and thus does not measure the student's pragmatic communication ability. It is true that DET does not specifically classifies its targeted group of test-takers, for instance, high school or college students, and people at work, so the lack of university life topics in the test questions cannot assess whether students have communicative skills and academic writing skills in future campus life. In contrast, the TOEFL and IELTS exams include many topics that are closely related to university life. For example, in the TOEFL listening test, each test-taker listens to two conversations related to university life and study, and students discuss with teachers, staff, and other students on campus, covering vivid topics such as dormitory life, campus facilities, part-time work, course selection, and thesis revision; The lectures on listening also involve the knowledge of biology, astronomy, geography, humanities, social sciences and other disciplines, which can not only test the language level of the test-taker, but also see the students' knowledge, adaptability and ability to accept and integrate new information, so as to more comprehensively evaluate and predict whether the test-taker has the basic language and communication skills of studying abroad.

Second, there is a lack of opportunities for interaction and communication in the DET compared to IELTS. Duolingo English mock tests and practical tests rely more on computer-based automatic questions, as do speaking and writing tests. During the exam, the test-taker can only passively accept the test questions and has no opportunity to communicate face-to-face or interact with others. In language learning, communication is a very important part, and only by communicating with others can we better master language knowledge and skills. In contrast, the third part of IELTS Speaking has direct communication between the examiner and the candidate. The examiner will ask questions from different perspectives based on what the candidate has previously expressed, effectively provoking thoughts and instantly evaluating oral language proficiency. This link requires candidates to be able to communicate face-to-face with people and adapt to changes, which is also difficult to achieve by computer automatic questions.

Warrant 3: Teachers and Chinese universities can use Duolingo English mock test items help improve students' English learning and evaluate their language proficiency according to the test results.

Evidence 3: The Duolingo English mock test is a cost-effective and efficient tool that can be utilized by teachers or schools to assess a test-taker's language ability. This test presents a fixed set of questions with varying levels of difficulty, and employs computer-adaptive questioning to match the difficulty level with each test-taker's proficiency. This objective evaluation standard can aid teachers or schools in assessing students' English proficiency more accurately.

Moreover, the responses of students to the Duolingo English Mock Test can provide valuable feedback to teachers or schools, allowing them to gain insight into students' English learning and deficiencies. This feedback can assist teachers in providing guidance and advice tailored to students' individual needs.

The results of the Duolingo English Mock Test also serve as a reference for teachers or schools to gain a better understanding of students' English proficiency and learning needs. This information can be used to design and adjust English courses more effectively, ultimately improving the quality and effectiveness of teaching.

Lastly, the results of the Duolingo English Mock Test can be utilized to identify outstanding students who excel in English academic performance. Such students can be rewarded and supported accordingly, motivating them to continue learning English and achieving academic success.

Overall, the Duolingo English mock test is a valuable tool that can aid teachers or schools in assessing, guiding, and motivating students in their English language learning journey.

Potential rebuttals 3: Upon further analysis of the above argument, it is evident that while the Duolingo English mock test is a valuable tool for assessing students' language knowledge and skills, it may fall short in comprehensively assessing their language application ability. Relying solely on test scores to evaluate a student's English proficiency may result in overlooking their actual language expression and communication skills, which are essential components of language learning.

Furthermore, the fixed nature of the Duolingo English mock test's questions and difficulty level may limit its effectiveness as a tool for gaging students' overall English language proficiency. If teachers solely focus on training students for the test and its specific question types, it may cause students to be unable to expand and deepen their English learning, and may place too much emphasis on test-taking strategies rather than genuine language acquisition.

Another potential limitation of the Duolingo English Mock Test is its reliance on computer-based questions. As a result, it may be challenging for teachers and schools to access new questions or create their own, making it difficult to test whether students can integrate their language knowledge and apply it to real-life scenarios.

Therefore, while the Duolingo English mock test is a useful tool for evaluating language knowledge and skills, it should be complemented with additional assessment methods to provide a more comprehensive evaluation of a student's English proficiency. Teachers and schools should strive to incorporate opportunities for students to practice and apply their language skills in real-life situations, rather than solely relying on standardized testing. Additionally, efforts should be made to diversify the question types used in language assessments, allowing for a more nuanced understanding of students' language abilities.

5.2 Reading tests

5.2.1 Construct definition

The reading test items in DET include two types, namely *C-test* and *Interactive Reading*. According to the requirements of the AUA model, in order to describe the characterization of the TLU domain when constructing the test items, it is necessary to first define the ideas involved, then determine the targeted groups of test-takers, and finally clarify the relationship between the input information from test material and the feedback test answer.

Upon reviewing the official definitions from DET, it is evident that the reading section consists of various question types that require a range of knowledge and skills. Specifically, the C-test item presents a paragraph with incomplete sentences, requiring test-takers to complete the missing words according to contextual and discursive information. This process demands a combination of vocabulary, morphological, and syntactic knowledge, making it a challenging task.

The Interactive Reading section comprises five different question types, namely Cloze Questions, Text Completion, Highlighting Sentences, Main-idea Questions, and Possible Title Question, each corresponding to a distinct reading construct and required knowledge skills. These constructs and skills are detailed in Table 2.

For instance, Cloze Question mobilizes test-takers' vocabulary, morphological, and syntactic knowledge. On the other hand, Text Completion and Main-idea Questions demand a sense of text structure and recognition of discourse organization. Possible Title Question also requires an understanding of the passage's general idea as a whole.

When selecting an article topic, test-takers must infer and summarize the most appropriate title based on textual information. Additionally, highlighting sentence questions require recalling and locating the relevant information to test the reading strategy. All of these question types necessitate adopting appropriate reading strategies, quickly identifying words, and searching for textual information.

5.2.2 Groups of test-takers

Upon analyzing the mock tests, it is apparent that the reading questions cater to a wide range of test-takers, covering various language characteristics, text structures, language domains, and cultural backgrounds.

The test questions consist of 100–150 words in each reading passage that assesses a variety of reading knowledge and skills. The complexity of grammar, vocabulary, and syntax varies according to the test-taker's level, and the questions also involve semantic characteristics such as synonyms and lexical collocations. The text structure varies between expository and narrative forms. The former usually has a topic sentence, supporting details and conclusion, while the latter is mostly composed in chronological order. Therefore, the structure of each passage is not too complicated. The language domains cover technical English, business English, and other formal use of language. Moreover, the test involves diverse cultural backgrounds, bringing a strong element of variability to the test.

The reading question types in the Duolingo English test are suitable for all types of English learners. For beginners, the computer-adaptive questions match their level, and the materials can help them expand their basic vocabulary, understand grammar, and sentence patterns, and improve overall reading comprehension. For intermediate and advanced learners, the reading test items can help them further improve their reading comprehension skills. The articles are sourced from online materials and keep pace with the times, allowing test-takers to expand their knowledge, practice their reading skills, and understand language and cultural knowledge.

In conclusion, DET reading section is a well-designed assessment tool that caters to a broad audience of English learners, providing them with the opportunity to expand their knowledge, practice their reading skills, and gain a deeper understanding of language and cultural knowledge.

5.2.3 Input information

Upon analyzing the information presented to the test-takers in the reading tests, as well as their corresponding answers, it is evident that there are differences between low-level and high-level test-takers in various aspects of language learning.

Regarding grammar, low-level test-takers tend to spend more time on vocabulary questions and have a higher accuracy rate in multiple-choice questions in Interactive Reading, but a lower accuracy rate in

TABLE 2 Definitions of reading constructs in DET.

CEFR	Test items	Reading purpose	Knowledge/Skills
Reading for information and arguments	Cloze	Learning and integrating Information	Vocabulary, Morphology, Syntax
	Text completion		Summary and structure awareness
	Main-idea questions	Understanding the gist	Understanding and inferring information
	Possible title		Evaluation and critical reading
Reading orientation	Highlighting Sentences	Search process and quick understanding	Fluency, reading speed, recalling information

fill-in-the-blank questions in C-test. On the other hand, high-level test-takers tend to spend less time on these vocabulary questions, with high accuracy in Interactive Reading, but relatively lower accuracy in spelling in C-test.

With regards to semantics, low-level test-takers typically struggle with filling in correct synonyms and word collocations, while high-level test-takers find it relatively easy. However, they may find individual word collocations more challenging, such as *set great score in certain qualities*, which are not commonly found in high school entrance examinations or Chinese College English Tests. Therefore, few test takers, regardless of their language proficiency, managed to fill in the word *score*.

Regarding text comprehension, all test-takers can correctly grasp the main idea and gist of the article. However, low-level test-takers may make errors when searching for information to highlight, often wrongly selecting redundant information. In contrast, high-level test-takers are more precise in highlighting for detailed information.

Regarding register, low-level test-takers may struggle with vocabulary in science and technology topics, like *warehouse* and *assemble*, while high-level test-takers can fully understand vocabulary of various topics.

Finally, in terms of culture, low-level test-takers may find foreign fiction and history reading more challenging, while high-level test-takers may struggle with words that are not commonly used in their daily lives, such as *racoon* and *yacht*.

In summary, both low-level and high-level test-takers have their strengths and weaknesses in different aspects of language learning. Therefore, they must focus on targeted learning and improvement to enhance their language proficiency. The overall description of the characterization of TLU domain in reading tests is shown below in Table 3.

6 Discussion

In conclusion, this study advocates for the use of the Duolingo English mock test by Chinese college-level students for assessing English language proficiency and promoting English learning, but the mock test may need modifications to meet the needs of Chinese EFL. The AUA model, which combines quantitative and qualitative data, provides advantages and disadvantages of using this test as a language assessment tool, as is shown in Figure 4.

The Duolingo English mock test offers a range of question types that can be practiced for free and unlimited times, enabling test takers to continue practicing and improving their language skills. Furthermore, the test's unified reference standard allows teachers and schools to objectively assess students' language proficiency and facilitate teaching and assessment.

However, there are some objections to the use of the Duolingo English mock test. Firstly, the test results only provide a score segment and lack specific scores for each subject or ability, which makes it difficult to determine a test taker's exact ability level. Secondly, the test questions may lack practicality and interactivity, and the content may not be conducive to promoting communication skills. Lastly, the test's design, which is completed by GPT-3, may not be adaptable, and the overall computer adaptive matching task may be challenging to promote without technical support.

Therefore, while the Duolingo English mock test offers significant advantages as an assessment tool, it may not be suitable for all contexts and needs to be used in conjunction with other assessment methods to provide a more comprehensive evaluation of a test taker's language proficiency. This exam has only been implemented for 6 years, and corresponding supplementary study materials and training coaching courses are still under further development. Here are some suggestions to hopefully help DET attract more test-takers:

First of all, to better address the problems of test content faced by low-proficiency level EFLs, the official guide can provide sample tests with reference answers that cover various topics and terminology of vocabulary. It will serve as a preparation guide for new test-takers. Additionally, the mock test website can visualize the progress made by its test-takers or offer some impetuses to redo and learn from each mock test. Here is the quotation from one of the low-proficiency level test-takers with a mock test score of 90–100:

TABLE 3 Characterization of TLU Domain of DET reading test items.

	Characterization of TLU
Test-takers	EFL from all levels
Test construct	Reading knowledge and skills
Input form	Paragraphs of 100–150 words each
Features of TLU Domain in test items	Grammar: vocabulary and syntactic complexity vary with the level of the test taker Semantics: synonyms, antonyms, lexical collocations Context: expository and narrative text with simple structures Register: formal language, covering daily life, technical English, business English, etc. Culture: diverse and highly variable

Performance	Low-level test-takers	High-level test-takers
Grammar	Spending more time with low accuracy in C-test	Spending more time with low accuracy in C-test
Semantics	Struggling with lexical knowledge	Struggling with low-frequency collocations
Context	Redundancy in highlighting	Precisely highlighting
Register	Struggling with technological topics	Correctly understanding all topics
Culture	Struggling with fiction and history	Struggling with low-frequency words

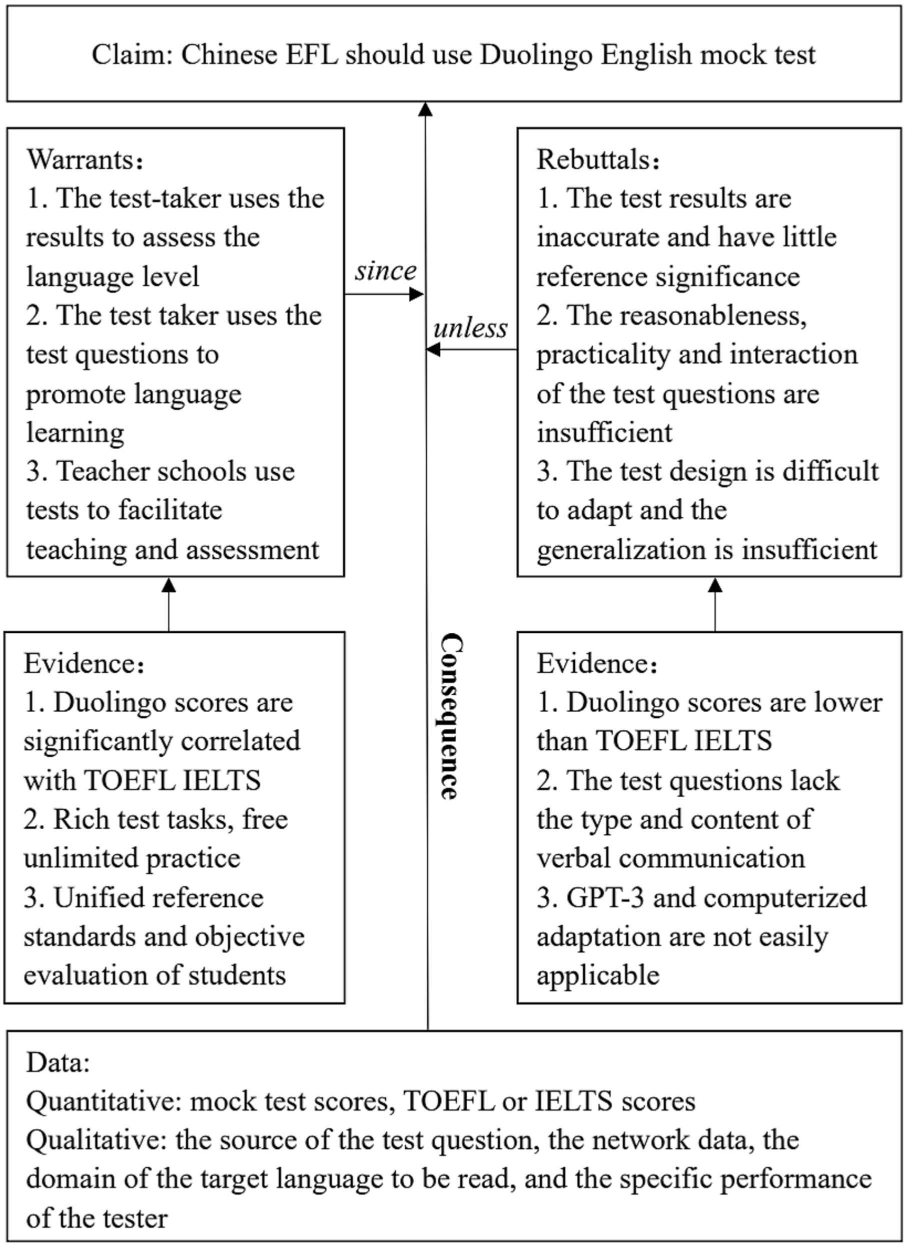
Claim: Chinese EFL should use Duolingo English mock test
Warrants：
1. The test-taker uses the results to assess the language level
2. The test taker uses the test questions to promote language learning
3. Teacher schools use tests to facilitate teaching and assessment
since
unless
Rebuttals：
1. The test results are inaccurate and have little reference significance
2. The reasonableness, practicality and interaction of the test questions are insufficient
3. The test design is difficult to adapt and the generalization is insufficient
Consequence
Evidence：
1. Duolingo scores are significantly correlated with TOEFL IELTS
2. Rich test tasks, free unlimited practice
3. Unified reference standards and objective evaluation of students
Evidence：
1. Duolingo scores are lower than TOEFL IELTS
2. The test questions lack the type and content of verbal communication
3. GPT-3 and computerized adaptation are not easily applicable
Data:
Quantitative: mock test scores, TOEFL or IELTS scores
Qualitative: the source of the test question, the network data, the domain of the target language to be read, and the specific performance of the tester

FIGURE 4
AUA model of DET consequential validity.

> *"…for the mock test, it would be better to narrow down the score range a bit. It would give test takers a more precise understanding of their scores. It could also include a percentage score, such as being in the top 30% or 10% among all test takers."*

Besides, the test form may need modification regarding the time for each test items. Compared to TOEFL/IELTS that present test-takers with longer passages or lectures, DET most tests feature mostly short test items of within 3 min. For Chinese students, it is demanding to quickly adapt to short test items because they are more likely to be customed to longer reading articles (each of which takes 5 min to finish reading) in both College Entrance Examinations and College English Tests. As is said by one high-proficiency level test-taker with a score of 110–120:

> *…Duolingo feels more like an instinctive response, without enough time to think. When I'm working on the previous question, the next question pops up before I even realize it, and sometimes I do not even know what I'm writing… Especially during the speaking exercises combined with listening, it would be helpful to have a pause before the audio plays so that I can gather my thoughts. Otherwise, the audio finishes before I can even process it.*

Nonetheless, there are some limitations in the research due to various factors. To be specific, there remain opportunities for further research to enhance its scope and depth, but the study was limited by certain factors, including individual difference, the impact of the pandemic on the number of college students studying abroad and preparing for TOEFL/IELTS, resulting in a small sample size for this study. Additionally, access to the complete question bank of the actual test is restricted due to copyright issues, and the evaluation of AUA itself was inadequate, with insufficient evidence collection and definition of quality attributes. Despite these limitations, the Duolingo English Test remains a compelling topic for continued research, especially given recent changes in question types and the development of GPT-4. Future research can focus on collecting more test results, enhancing the credibility of quantitative analysis, and incorporating interviews and questionnaires with stakeholders to provide more detailed qualitative evidence. Ultimately, this research aims to contribute to the optimization and popularization of this type of test and provide valuable insights for English learners and test result users.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

XM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – original draft. HZ: Conceptualization, Methodology, Supervision, Validation, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

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.

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

Abeywickrama, P., and Brown, H. D. (2010). *Language assessment: Principles and classroom practices*. NY: Pearson Longman.

Alderson, J. C. (2000). *Assessing Reading*. Cambridge: Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Lang. Assess. Quar. Int. J.* 2, 1–34. doi: 10.1207/s15434311laq0201_1

Bachman, L. F., and Palmer, A.S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press

Bachman, L. F., and Palmer, A.S. (2010). *Language assessment in practice* (*2nd*). Oxford: Oxford University Press.

Bézy, M., and Settles, B. (2015). The Duolingo English test and East Africa: preliminary linking results with IELTS & CERF. Duolingo Research Report [DRR-15-01]. Duolingo.

Blanchard, J., and Mikkelson, V. (1987). Underlining performance outcomes in expository text. *J. Educ. Res.* 80, 197–201. doi: 10.1080/00220671.1987.10885751

Britt, M.A., Rouet, J.-F., and Durik, A.M. (2018). *Literacy beyond text comprehension: a theory of purposeful Reading*. London: Routledge.

Cardwell, R., LaFlair, G. T., and Settles, B. (2022). Duolingo English Test: technical manual. Duolingo Research Report

Chapelle, C. (1999). Validity in language assessment. *Annu. Rev. Appl. Linguist.* 19, 254–272. doi: 10.1017/S0267190599190135

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. *Psychol. Bull.* 52, 281–302. doi: 10.1037/h0040957

DeBarger, A. H., Penuel, W. R., Harris, C. J., and Kennedy, C. A. (2016). Building an assessment argument to design and use next generation science assessments in efficacy studies of curriculum interventions. *Am. J. Eval.* 37, 174–192. doi: 10.1177/1098214015581707

Grabe, W., and Jiang, X. (2014). "Assessing reading" in *The companion to language assessment*. ed. A. J. Kunnan (New York: John Wiley & Sons)

Green, A., Ünaldi, A., and Weir, C. (2010). Empiricism versus connoisseurship: establishing the appropriacy of texts in tests of academic reading. *Lang. Test.* 27, 191–211. doi: 10.1177/0265532209349471

Gronlund, N. E. (1998). *Assessment of student achievement*. Allyn & Bacon Publishing, Longwood Division, 160 Gould Street, Needham Heights, MA 02194-2310; tele.

Isbell, D. R., and Kremmel, B. (2020). Test review: current options in at-home language proficiency tests for making high-stakes decisions. *Lang. Test.* 37, 600–619. doi: 10.1177/0265532220943483

Kane, M. T. (1992). An argument-based approach to validity. *Psychol. Bull.* 112, 527–535. doi: 10.1037/0033-2909.112.3.527

Kane, M. T. (2001). Current concerns in validity theory. *J. Educ. Meas.* 38, 319–342.

Kane, M. T. (2012). Validating score interpretations and uses. *Lang. Test.* 29, 3–17. doi: 10.1177/0265532211417210

Khalifa, H., and Weir, C. J. (2009). *Examining Reading: research and practice in assessing second language Reading*. Cambridge: Cambridge University Press.

Khodadady, E. (2014). Construct validity of C-tests: a factorial approach. *J. Lang. Teach. Res.* 5, 1353–1362. doi: 10.4304/jltr.5.6.1353-1362

Llosa, L. (2008). Building and supporting a validity argument for a standards-based classroom assessment of English proficiency based on teacher judgments. *Educ. Meas. Issues Pract.* 27, 32–42. doi: 10.1111/j.1745-3992.2008.00126.x

Mann, W., and Marshall, C. R. (2010). Building an assessment use argument for sign language: the BSL nonsense sign repetition test. *Int. J. Biling. Educ. Biling.* 13, 243–258. doi: 10.1080/13670050903474127

Maris, G. (2020). The Duolingo English test: Psychometric considerations. Technical Report DRR-20-02, Duolingo.

Mozgalina, A. (2015). *Applying an argument-based approach for validating language proficiency assessments in second language acquisition research: the elicited imitation test for Russian*. Georgetown University ProQuest Dissertations Publishing.

Pardo-Ballester, C. (2010). The validity argument of a web-based Spanish listening exam: test usefulness evaluation. *Lang. Assess. Q.* 7, 137–159. doi: 10.1080/15434301003664188

Pardo-Ballester, C. (2020). An assessment use argument for Spanish for the professionals. INTED2020 Proceedings, Available at: https://library.iated.org/view/PARDOBALLESTER2020ANA

Park, Y., LaFlair, G. T., Attali, Y., Runge, A., and Goodwin, S. (2022). Interactive Reading—The Duolingo English test. Duolingo Research Report DRR-22-02. Duolingo

Qian, D. D., and Pan, M. (2013). Response formats. *Comp. Lang. Assess.* 2, 860–875. doi: 10.1002/9781118411360.wbcla090

Riley, G. L., and Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Lang. Test.* 13, 173–189. doi: 10.1177/026553229601300203

Schmitt, N., Nation, P., and Kremmel, B. (2020). Moving the field of vocabulary assessment forward: the need for more rigorous test development and validation. *Lang. Teach.* 53, 109–120. doi: 10.1017/S0261444819000326

Settles, B., Hagiwara, M., and LaFlair, G. T. (2020). Machine learning–driven language assessment. *Trans. Assoc. Comp. Linguist.* 8, 247–263. doi: 10.1162/tacl_a_00310

Toulmin, S. E. (2003). The uses of argument: Updated edition. doi: 10.1017/CBO9780511840005,

Verhelst, N.D., Avermaet, P.V., Takala, S., Figueras, N., and North, B. (2009). *Common European framework of reference for languages: Learning, teaching, assessment*.

Wagner, E. (2020). Duolingo English test, revised version July 2019. *Lang. Assess. Q.* 17, 300–315. doi: 10.1080/15434303.2020.1771343

Waluyo, B. (2018). Vocabulary acquisition through self-regulated learning on speaking and writing development. *Int. J. Lang. Teach. Educ.* 2, 286–302. doi: 10.22437/ijolte.v2i3.5747

Wang, H., Choi, I., Schmidgall, J., and Bachman, L. F. (2012). Review of Pearson test of English academic: building an assessment use argument. *Lang. Test.* 29, 603–619. doi: 10.1177/0265532212448619

Winchell, A., Lan, A., and Mozer, M. (2020). Highlights as an early predictor of student comprehension and interests. *Cogn. Sci.* 44:12901. doi: 10.1111/cogs.12901