Liu *et al. Language Testing in Asia* (2023) 13:49
https://doi.org/10.1186/s40468-023-00263-z

Language Testing in Asia

REVIEW Open Access

Test review: a general academic English summative test

Yunlong Liu[1], Yaqiong Cui[1] and Hua Yu[1*]

*Correspondence:
yuhuatj@ucas.ac.cn

[1] Department of Foreign Languages, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract

Many Chinese universities are implementing an educational reform to transform their general English courses into academic English courses. Accordingly, how to assess students' English ability should also be reformed. In this test review, we introduce a school-based general academic English summative test developed by English instructors. An argument-based approach was adopted to analyze the test validity by obtaining students' test data and their reflective responses to the test. This review can provide a practical reference for the development of a valid general academic English summative test by including practices of course instructors and voices of students, two important test stakeholders, in test design.

Keywords: General academic English, Language assessment, Summative test development, Validity argument

Introduction

Considering Chinese graduate students' increasing need to publish in top international academic journals, a growing number of scholars have advocated curriculum reform aiming to transform general English courses into academic English courses (Liu et al., 2020), which necessitates the development of high-quality tests to assess students' academic English abilities. A high-quality academic English test, which is based on test development principles and course syllabuses, can be a useful tool for monitoring learners' progress over time (Douglas, 2014) and promoting students' language development (Wolf, 2020). Many international high-stakes tests, such as TOEFL and IELTS, have been widely used for admission purposes due to their ability to diagnose students' language proficiency to communicate in academic settings (Malone & Montee, 2014). However, these tests do not apply to contexts where low-stakes tests are more suitable, such as a curriculum reform, as scores of low-stakes tests can be used to diagnose students' learning problems and assess their progress, so as to make necessary adjustments in teaching contents (Chapelle & Voss, 2013) and advance the reform progress. Therefore, a well-designed low-stakes test that positively influences learning and teaching outcomes within the classroom (Alderson & Wall, 1993) is highly necessary. Considering the variety of testing contexts and test uses, different approaches to validation should be considered (Norris, 2008). According to Chapelle and Voss (2013), an argument-based approach aims to use evidence to justify the use of test

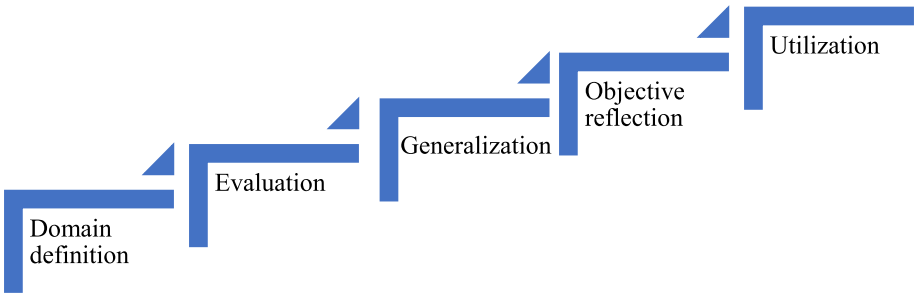

Fig. 1 Steps in the validity argument for a low-stakes test

scores which indicate test takers' ability and involves domain definition, evaluation, generalization, objective reflection, and utilization (see Fig. 1). Each step needs supportive evidence (e.g., scores) to analyze the test validity. The approach, which concerns the necessity to make a link between the test and classroom objective, has been proven feasible to validate low-stakes tests (Chapelle & Voss, 2013).

The knowledge involved in the theoretical model of communicative language ability proposed by Bachman and Palmer (2010), which consists of language knowledge and strategic competency, should be measured in a test. While language knowledge measures test takers' vocabulary, syntax, cohesion, pragmatic, and sociolinguistic knowledge, strategic competence refers to test takers' ability to use metacognitive, cognitive, and affective strategies to finish different tasks. These strategies can enhance the effectiveness of communication or compensate for breakdowns in communication (Swain, 1989). Considering that the corresponding assessments for academic English courses are in urgent need, many Chinese universities have started to develop academic English summative tests, some of which are more institutional (He et al., 2021), while some are more focused on classroom achievements, like the exam introduced in this paper, which is developed by individual classroom teachers. Unfortunately, the studies on validating these tests are scarce. Among the few validation studies, Zhou and Yoshitomi (2019) found that although students showed positive attitudes toward the validity of the TOEIC Speaking test, such positive attitudes had little influence on their test performance. Razavipour et al. (2020) found that test takers' positive perceptions of test contents were correlated with more intensive test preparation which, however, could not be attributed to their perceptions of test uses and test value. Both studies focused on test takers' views on high-stakes English proficiency tests. Therefore, when gathering evidence to support the validity of a summative test of general academic English, it is important to include test takers' views on the test (Ahmadi Safa & Sheykholmoluki, 2023), as such information will enable us to understand the problems that students face, inspire course instructors to make corresponding changes in their teaching plans, and improve the instructors' language assessment literacy.

Development of a summative test of general academic English

The academic reading and writing course

The academic English reading and writing course, a selective course in the master program of a research university in Beijing, China, is offered to improve graduate students' ability to read and write research papers in English. The course focuses on four parts,

Table 1 Course contents and academic skills

Vocabulary knowledge	Grammar knowledge	Academic reading skills	Academic writing skills
• Acquiring widely-used academic words	• Analyzing sentence members	• Identifying topic sentences	• Summarizing
• Acquiring widely-used collocations	• Analyzing complex sentences	• Identifying structure	• Paraphrasing
• Using logical connectors properly	• Understanding the use of active and passive voices	• Making inferences	• Inserting in-text citations
• Using phrases properly	• Sentence structures	• Analyzing the logic between sentences	• Synthesizing sources
• Identifying reporting verbs	• Using Verb tenses correctly	• Identifying research gaps	• Using cohesive devices
		• Understanding cohesive devices	• Using Nominalization
		• Guessing the general meaning of unfamiliar words	• Using hedges
			• Evaluating previous studies

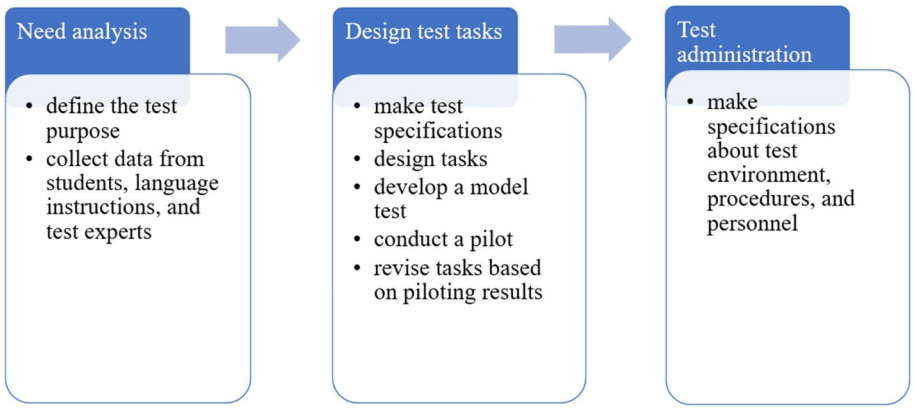

Fig. 2 Test development process

including the introduction to writing a research paper (e.g., IMRD structure), academic vocabulary, grammar knowledge, and academic reading skills. The specific contents and skills taught in class are shown in Table 1.

Development process and test description

The test development process (Fig. 2) is informed by Douglas (2014). Al Lawati (2023) argued that test specifications are crucial for developing good-quality tests. With different ways to design test specifications, test developers can choose what is suitable according to their situated context (Davidson & Lynch, 2008). In this study, to ensure the quality of test item writing, we followed components of test specifications proposed by Bachman and Palmer (1996), including target language use (TLU) domain, definition of construct to be measured, task characteristics, characteristics of the input and the expected response, and characteristics of test takers (Weir, 2005).

Target language use (TLU) domain

The purpose of the test is to measure whether test takers have acquired the academic skills taught in class (see Table 1) and whether they can use those skills in academic settings. One of the indicators that can assess the usefulness of a test is authenticity which refers to the degree of correspondence between the characteristics of language learning tasks and those of test tasks (Bachman & Palmer, 1996). The tasks involved in this test are highly similar to what test takers may encounter in academic settings. For instance, while reading academic papers, they need to comprehend long and complex sentences, make inferences, understand logical relationships via cohesive devices, and identify the structure of a paragraph, and these skills are tested in our test.

Definition of construct to be measured

This test mainly measures test takers' academic English proficiency. The definition of construct to be measured is introduced based on the theoretical model of communicative language ability (Bachman & Palmer, 2010), which consists of language knowledge and strategic competency.

Language knowledge measures test takers' vocabulary, syntax, cohesion, pragmatic, and sociolinguistic knowledge (Bachman & Palmer, 1996). At the lexical level, this test measures whether test takers have a good knowledge of the forms, meanings, and usages of words that are frequently used in academic settings. The syntax knowledge measured in the test entails structures commonly used in paper writing, including participial modifiers, dependent clauses, passive structures, verb forms, and infinitives. Test takers' ability to identify and use cohesive devices is also measured.

Pragmatic knowledge is about how texts used in the test align with the communicative goals of language users and the characteristics of the language use setting. In the Academic Reading section, test takers' abilities to skim, scan, and deduce implicit information (Alderson, 2000), as well as to interpret, evaluate, and synthesize viewpoints in reading materials, are assessed. The Academic Writing section (e.g., synthesizing sources), on the other hand, measures their abilities to select relevant sources, summarize and paraphrase original texts, insert in-text citations, and arrange sources logically, which are essential skills for writing academic papers. This section also measures test takers' register awareness (e.g., academic genre and language formality), one of the components involved in sociolinguistic knowledge (Bachman & Palmer, 1996).

Strategic competence in Bachman and Palmer's framework (1996) refers to test takers' ability to use metacognitive, cognitive, and affective strategies to finish different tasks. Using reading strategies appropriately while comprehending academic materials and completing the comprehension questions is considered a key academic ability. In this test, test takers' abilities to make full use of cognitive strategies to solve problems, such as scanning for the main idea, guessing unfamiliar words from context, and locating details, are mainly evaluated.

Table 2 Demographic information of test developers

Test developers (pseudonym)	Gender	Age	Education background	Years of academic English instruction	Tasks developed
Kate	Female	37	Ph.D. in applied linguistics	4	a. Vocabulary b. Academic Reading comprehension
Noah	Male	35	Ph.D. in applied linguistics	7	a. Academic writing
Jenny	Female	50	M.A. in linguistics	20	a. Grammar knowledge b. Academic Reading comprehension

Table 3 Test sections overview

Test sections	Tasks	Time limit	Questions	Points
Use of academic English	One: Cloze	30 min	1–10	15 points
	Two: Grammar knowledge		11–20	15 points
Academic English reading	One: Multiple choice questions	40 min	21–28	20 points
	Two: Complete a text using sentences given		29–32	10 points
Academic English writing	One: Sentence integration	50 min	33–36	20 points
	Two: Synthesizing sources		37	20 points

Task characteristics

Characteristics of setting

Characteristics of the setting include physical characteristics, participants, and time of task. The test was administered in regular classrooms, each with a capacity of 60 people, for 150 minutes. Adequate distance between seats was maintained. Test takers were carefully monitored by the three-course instructors who also developed the summative test (Table 2).

Characteristics of test rubrics

The test language is English, the test takers' target language. Aural aids are not available. Test procedures include handing out test papers 1 and 2, together with answer sheets, clarifying test instructions, finishing the test, and collecting materials at the end of the test.

Structure and time allotment

Table 3 summarizes test sections, task types, time limit, number of questions for each section, and score distribution.

Scoring methods

The Use of Academic English and Academic English Reading sections, containing multiple-choice and blank-filling items, are automatically scored by an online testing

Table 4 Scoring standards for task 2 of the academic writing section

Evaluation items	The test-taker can	Scale
Cohesion	produce clear language, showing controlled use of organizational patterns, connectors, and cohesive devices	0—strongly disagree 1—disagree 2—somewhat disagree 3—somewhat agree 4—agree 5—strongly agree
Coherence	produce a well-organized, coherent, and logical synthesis of sources	
Language quality	skillfully use sentence variety and precise vocabulary to convey meaning effectively; demonstrate superior facility with sentence structure	
Evaluation	can appropriately use reporting verbs/ verb tenses/opinions makers to express attitudes towards a source	0—strongly disagree 1—disagree 2—agree 3—strongly agree
In-text citations	add in-text citations properly	0-Strongly disagree 1-Agree 2-Strongly agree

platform. The Academic Writing section is scored by the course instructors who have reached an agreement on the scoring standards.

Holistic scoring and analytic scoring are widely used to rate test takers' compositions, but the latter is widely used in foreign language proficiency tests because of its higher reliability (Nakamura, 2002). Therefore, analytic scoring was adopted for the Academic Writing section. Based on the course objectives, we used five evaluation items, including cohesion, coherence, language quality, in-text citations, and use of evaluative language (Table 4). Two raters separately judged test takers' performance on each evaluation item based on the scale (Table 4), and their average score was the final score. The Pearson correlation coefficient of the two raters for the writing tasks was 0.309 ($p = 0.001$), indicating the scoring methods are practical and reliable.

Characteristics of the input and the expected response

Given the test objective to measure graduate student's ability to use English in academic situations, the texts used are from research papers in various fields whose topics are largely comprehensible to test takers.

The Use of Academic English section assesses test takers' ability to use academic words and grammar knowledge, such as how to use verb tenses in different parts of a research paper. The Academic English Reading section assesses how well test takers can comprehend excerpts of academic papers from various research areas. This section includes three reading passages, approximately 450 words each, with 4 questions per passage. Task 1 measures test takers' abilities to understand the main idea, details, and complex sentences, make inferences, and identify the structure of a paragraph. These abilities are necessary for them to comprehend research papers. Task 2, which asks test takers to complete a passage with scrambled sentences, mainly evaluates their critical thinking ability and knowledge of cohesion. The Academic English Writing section mainly measures test takers' ability to write for academic purposes. Task 1 requires test takers to integrate two or three short and simple sentences into a longer and more logical sentence by using various cohesive devices. For task 2, test takers are required to comprehend several reading paragraphs on the same topic, and synthesize them into one coherent

Table 5 Knowledge on moves in different parts of a research paper

Knowledge on moves	The number of the question
Identifying moves	Q26

Table 6 Knowledge of vocabulary measured in the test

Vocabulary knowledge	The numbers of questions
Using widely-used academic words	Q2, Q5, Q8, Q9
Using widely-used collocations	Q3, Q4, Q6
Using logical connectors properly	Q27
Using phrases properly	Q1, Q7, Q10
Identifying reporting verbs	Q37

and logical paragraph of about 150 words. This task models the procedures involved in writing a literature review, a common academic task assigned by graduate supervisors.

Characteristics of test takers

Weir (2005) proposed the framework for introducing the characteristics of test takers, including their physical, psychological, and experimental characteristics. Our test takers were 118 graduate students (majoring in management, economics, food, biology, and engineering) from three intact classes, with an average age of 22. They were in good physical condition when taking the test. They were motivated when studying the course owing to their need to publish academic papers in English. Although they are considered as intermediate-advanced English learners, most of them still reported that they felt nervous when preparing for the exam.

Test takers took the summative test in November 2022 after they finished the selective English course for 36 h (i.e., 9 weeks). To familiarize test takers with the test tasks and to lower their anxiety, a sample test was given to them two weeks before the test date, and the objectives, grading rubrics, and answers to each task were clearly explained by the instructors.

Validation arguments for the summative test

The three instructors designed and developed the summative test from January to November, 2022. From December 2022 to February 2023, relevant evidence has been collected to support the test validity. Instructors used the argument-based approach (Fig. 1) proposed by Chapelle and Voss (2013) to validate the test.

Domain definition

The first step is domain definition, which refers to the relevance of the test tasks to the course objectives. Given that the test tasks are designed based on the course content, they are closely related. Specifically, the tasks in the test can cover most of the knowledge and skills taught in the course (Tables 5, 6, 7, 8, and 9).

Table 7 Grammar knowledge measured in the test

Grammar knowledge	The numbers of questions
Analyzing sentence members	Q14, Q16, Q20
Understanding the use of active and passive voices	Q17, Q18,
Sentence structures	Q33, Q34, Q35, Q36
Analyzing complex sentences	Q21, Q22, Q25, Q28,
Using Verb tenses correctly	Q11-Q13, Q15, Q19

Table 8 Academic reading skills measured in the test

Academic reading strategies	The numbers of questions
Identifying topic sentences	Null
Identifying structure	Q24
Making inferences	Null
Analyzing the logic between sentences	Q33–Q36
Identifying research gaps	Null
Identifying and understanding cohesive devices	Q23, Q29, Q30, Q31, Q32
Guessing the general meaning of unfamiliar words	Null

Table 9 Academic writing skills measured in the test

Academic writing strategies	The numbers of questions
Summarizing	Q37
Paraphrasing	Q37
Inserting in-text citations	Q37
Synthesizing sources	Q37
Using cohesive devices properly	Q37
Using Nominalization properly	Q37
Using hedges properly	Q37
Evaluating previous studies properly	Q37

Evaluation

The second step, evaluation, requires test developers to make specific scoring requirements and standards. All the tasks in the first two sections of the test are multiple-choice questions or gap-filling tasks, so automatic scoring from an online educational platform was used, which ensures scoring consistency. The Pearson correlation coefficient of the two raters for the writing tasks was 0.309 ($p = 0.001$), suggesting that the consistency between the two raters is acceptable.

Generalization

Generalization, the third step, means whether test takers can receive consistent evaluation if they receive the same test on different occasions or are evaluated by different raters. This step can be supported by the evidence that the descriptions of test-takers' performance are clear and consistent. The descriptors are written based on those used

by TOEFL iBT to evaluate its test takers' performance. Given its focus on test takers' academic language abilities, which aligns with the purpose of our test, the descriptors used by TOEFL iBT are of reference significance. Some new descriptors that evaluate students' course learning outcomes are added, such as "They can identify how cohesive devices are used in an academic paper, but sometimes may fail to understand the implicit logical relationship between ideas." The Cronbach's α of the first two sections was 0.618, an acceptable value, supporting the reliability of the first two sections of the test. The discrimination index was 0.28, an unsatisfying value, which can be attributed to the homogeneity in test takers' English proficiency. As they are all intermediate-advanced learners and skilled at understanding academic papers in English, most of them could perform well in the academic vocabulary and reading sections. The difficulty index of the first two sections was 0.76, indicating that test takers may have finished the two sections effortlessly, which may also explain the unsatisfactory discrimination index. These indexes indicate the need to design more challenging tasks that can better differentiate the performance of test takers.

Objective reflection

The correlation between students' scores on the summative test and their course grades can be used to demonstrate the next step, objective reflection. The Pearson correlation coefficient was 0.640 ($p = 0.009$), indicating the consistency between their scores on this test and their comprehensive performance in this course.

Utilization

Utilization, similar to the consequential validity included in Weir's (2005) framework that concerns the effect of a test on an individual within society and washback in the classroom, requires test developers to make decisions based on the score of a low-stakes test. Regarding the effect on test takers, test developers need to determine whether a test-taker can pass the course and earn the corresponding credits.

Regarding the washback effect, after knowing that the average score of task 2 of the Use of Academic English section was 6.62 points out of 15, for example, the course instructors pay more attention to helping students better understand how to use verb tenses in a research paper. For test takers, by reading the score report, they can understand their study problems, which may exert a positive influence on their future academic English study. For instance, a test taker whose score in the Academic Writing section is 23 can read the following performance descriptors:

> *"Test takers who receive an Academic Writing section score at the Below-Intermediate level are able to produce a simple literature review in English, but its language quality is not very good. They can express some ideas and synthesize sources on a topic, but insufficient explanations can lead to limited development of ideas. Important ideas from the sources are misinterpreted because of complex sentences in the sources. Minor language errors can occur in the literature review in which cohesive devices are sometimes misused."*

These descriptors can help the test taker understand his or her problems in writing a literature review, ensuring the utilization of the test.

Table 10 Test taker views about the summative test

Statements:	*N*	*Mean* rating	SD
The duration of the test is acceptable	108	3.963	1.058
The difficulty of task 2 of Sect. 1 is reasonable	108	4.019	0.976
The difficulty of the Academic Writing section is reasonable	108	3.843	1.043
The characteristics of the test tasks are in line with those of the reading tasks in real academic settings	108	3.741	1.053
The characteristics of the test tasks are in line with those of the writing tasks in real academic settings	108	3.731	1.09
The test is effective in instructing me on how to improve my academic English ability in the future	108	3.87	1.077

Test takers' views on the summative test

General views on the summative test

A hundred and eight students (out of 118 test takers) completed a questionnaire designed by ETS to investigate test takers' perceptions of TOEFL (Sato & Ikeda, 2015), but the items in our questionnaire were adapted to be more relevant to the academic skills taught in this course. The Cronbach's α of the questionnaire is 0.88, indicating that the questionnaire is reliable and that the conclusions based on it are convincing. The questionnaire, written in Chinese, consists of two parts. Part 1 is about our test takers' personal information, such as their gender, age, and experiences of learning academic English. Part 2, consisting of 42 statements, is about test takers' comments on the administration of the test (statements 1–4), their perceptions of the difficulty level of the test and its constituent sections (statements 5–16), and their views on whether all the sections of the test can measure their academic English ability (statements 17–42). Their academic ability is measured by specific skills outlined in the course objectives. Test takers rated on a 5-point scale the extent to which they agree with each statement, with 1 indicating "totally disagree" and 5 meaning "totally agree".

The questionnaire data were analyzed using SPSS 26.0. Given the absence of a definitive standard on how to determine the midpoint for this kind of data (McIver and Carmines, 1981), we decided that a mean agreement of 4.14 or higher suggests that most test takers agreed with the statement. This method was used in Malone and Montee's study (2013) which aimed to investigate test takers' perceptions of whether TOEFL iBT can measure their academic English ability.

In order to further understand their comments on the course and the test, several students were interviewed through face-to-face interview or online communication platform, WeChat.[1]

Most test takers spoke highly of the administration of the summative test, agreeing that the test environment was quiet, the rubrics of each task were clearly stated, and the organization of the test was good. Also, they held that the test tasks corresponded to the academic reading and writing skills taught in class.

Table 10 lists the items with a mean agreement lower than 4.14, suggesting that the test takers generally held a negative attitude toward the statement. They considered that

[1] WeChat is an instant messaging application developed by Tencent, a technology giant in China.

Table 11 Test taker views about the academic ability measured

Statements: the test can accurately measure my ability to	*N*	*Mean* rating	SD
use widely-used academic words	108	3.87	0.968
use widely-used collocations	108	4.037	0.819
use phrases properly	108	4.037	0.842
identify reporting verbs	108	4.12	0.84
use hedges properly	108	4.065	0.812
insert in-text citations in a correct form	108	3.889	0.98

the test time was not enough (3.963). Some test takers claimed that the test was difficult, especially task 2 of the Use of Academic English section and the Academic Writing section, with the average values being 4.019 and 3.843, respectively. According to our test takers, the authenticity of the test should be improved, as a small proportion of them reported that the characteristics of the test tasks were not in line with those of the reading and writing tasks in real academic settings, with the average values being 3.741 and 3.731, respectively. We also found that the washback effect of the test can be improved, as some of them believed that the summative test might not be effective in instructing them on how to improve their academic English ability in the future (3.87).

Interviews with the test takers further show their positive attitude towards the general design of the test and its tasks.

For example, as test taker A reported, "The test can comprehensively test what was learned from the course. The answers to the multiple-choice questions and gap-filling tasks are very definite. The tasks are very novel, and can measure our comprehensive academic English ability."

Test taker B also agreed that "The Academic Reading section covers the skills learned from the course. For instance, there is a question that requires us to identify the structure of a paragraph. What is more, the rubrics of each task are clear enough."

Test Taker C further recognized that "The design of the test is good, and I understand that test developers must have put a lot of effort into designing the questions."

These findings suggest that test developers should consider test takers' voices by conducting a survey. Furthermore, it is advisable for test developers to cooperate with course instructors, as they can make joint efforts to use test results more efficiently. Offering a score report seems insufficient to achieve the washback effect of a test, and a more detailed plan on how to improve test takers' academic English ability should be made.

Views about the specific academic English skills measured

Most students believed that the test covers the skills taught in class and that most of the specific academic English skills can be measured by the summative test because among the 27 specific academic skills, the average scores for 21 statements were higher than 4.14. The test takers held that six specific academic skills could not be measured by the test (see Table 11), including their abilities to use academic words (3.87), hedges (4.065), collocations (4.037) and phrases (4.037), to identify reporting verbs (4.12), and to insert in-text citations in a correct form (3.889).

Test takers' interviews further reveal that they are less likely to perceive skills indirectly assessed can be accurately measured in this test. For instance, test taker D reported, "It was difficult to judge whether my abilities to use phrases and collocations are accurately tested as these abilities may be tested in an indirect way."

Similarly, test taker E reported, "Without teachers' explanations, I couldn't judge whether these skills are tested since these skills are tested in an implicit manner. When answering the questions, I did not pay attention to phrases or collocations, I just wanted to finish the test."

The findings indicate the need to improve the face validity of a test. According to Sato and Ikeda (2015), test takers should clearly understand what test developers wish them to learn, so test developers should directly inform test takers of the specific skills measured by each item.

Conclusion

The current study is of practical significance in that it elaborates on the process of developing a summative test based on an academic reading and writing course. Some innovative tasks are also introduced, which can inspire test developers about how to design test tasks for a test that aims to measure test takers' academic English ability. In addition, the current study shows how to select evidence for each step in a framework designed by Chapelle and Voss (2013) and how to employ a questionnaire to collect test takers' views about whether their academic English ability can be measured by the test. However, the present study has some limitations. Firstly, it did not investigate how to achieve the washback effect of a summative test measuring test takers' academic English ability. Also, the reliability, difficulty, and discrimination of some tasks in the test can be further improved. Lastly, this study only collected data from test takers about their views on the test, but the data from other stakeholders, such as test experts and school administrators, should also be included. In particular, the design and revisions of test specifications based on comments from different stakeholders can further improve the operation of a test program (Davidson & Lynch, 2008); therefore, it is advisable for future studies to examine how test developers revise their test specifications. In addition, efforts are needed to improve the reliability of a local, teacher-developed test, and more attention should be given to the consequential validity of the test by collecting convincing evidence.

Abbreviations

ETS	Educational Testing Service
TOEFL iBT	Test of English as a foreign language, internet-based test
IELTS	International English Language Testing System
IMRD	Introduction, Methods, Results, and Discussion
LAL	Language assessment literacy

Acknowledgements

The authors are grateful to the anonymous reviewers and the editor for their helpful and constructive comments that greatly helped improve the manuscript.

Authors' contributions

Yunlong Liu: conceptualization, data curation, investigation, formal analysis, software, visualization, writing original draft and revised draft, project administration. Yaqiong Cui: conceptualization, data curation, investigation, formal analysis, writing original draft and revised draft. Hua Yu: conceptualization, data curation, investigation, formal analysis, writing original draft and revised draft. All authors read and approved the final manuscript.

Funding
Beijing Social Science Foundation (No. 18YYB002) and the Fundamental Research Funds for the Central Universities (No. E1E41701).

Availability of data and materials
The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 12 July 2023 Accepted: 4 October 2023
Published online: 23 October 2023

References

Ahmadi Safa, M., & Sheykholmoluki, H. (2023). An impact study of the Iranian National University Entrance Exam from students and parents' perspectives. *Language Test in Asia,13*, 40. https://doi.org/10.1186/s40468-023-00254-0
Alderson, J. C. (2000). *Assessing Reading*. Cambridge University Press.
Alderson, J. C., & Wall, D. (1993). Does Washback Exist? *Applied Linguistics,14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115
Al Lawati, Z. A. (2023). Investigating the characteristics of language test specifications and item writer guidelines, and their effect on item development: A mixed-method case study. *Language Testing in Asia,13*(1), 21. https://doi.org/10.1186/s40468-023-00233-5
Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford University Press.
Carmines, E. G., & McIver, J. P. (1981). Analyzing models with unobserved variables: Analysis of covariance structures. In G. W. Bohrnstedt & E. F. Borgatta (Eds.), *Social Measurement: Current Issues* (pp. 65–115). Sage Publications Inc.
Chapelle, C. A., & Voss, E. (2013). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1081–1097). United Kingdom: Wiley Blackwell.
Davidson, F., & Lynch, B. K. (2008). *Testcraft: A teacher's guide to writing and using language test specifications*. Yale University Press.
Douglas, D. (2014). *Understanding language testing*. Routledge.
He, L.-Z., Yuan, J.-F., & Min, S.-C. (2021). Validation of the alignment between an in-house writing test based on textual features and China's Standards of English Language Ability. *Foreign Language Education,03*, 52–57. https://doi.org/10.16362/j.cnki.cn61-1023/h.2021.03.009
Liu, J. E., Lo, Y. Y., & Lin, A. M. Y. (2020). Translanguaging pedagogy in teaching English for academic purposes: researcher-teacher collaboration as a professional development model. *System*, 102276. https://doi.org/10.1016/j.system.2020.102276
Malone, M. E., & Montee, M. (2014). Stakeholders' beliefs about the TOEFL iBT® test as a measure of academic language ability. *ETS Research Report Series,2014*(2), 1–51. https://doi.org/10.1002/ets2.12039
Nakamura, Y.(2002). Effectiveness of paired rating in the assessment of English compositions. *JLTA Journal*, *5:* 61–71. https://doi.org/10.20622/jltaj.5.0_61
Norris, J. M. (2008). *Validity evaluation in language assessment*. Peter Lang.
Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: A potential impact of face validity on student learning. *Language Testing in Asia,5*, 10. https://doi.org/10.1186/s40468-015-0019-z
Swain, M. (1989). Manipulating and complementing content teaching to maximize second language learning. *TESOL Canada Journal, 6,* 68–83. https://doi.org/10.18806/tesl.v6i1.542
Weir, C. J. (2005). *Language testing and validation*. Hampshire: Palgrave Mcmillan.
Wolf, M. K. (2020). Assessing young language-minority students: Validation challenges and future research directions. *Language Assessment Quarterly,5*, 1–9. https://doi.org/10.1080/15434303.2020.1826488
Zhou, Y.-J., & Yoshitomi, A. (2019). Test-taker perception of and test performance on computer-delivered speaking tests: The mediational role of test-taking motivation. *Language Testing in Asia,9*(1), 10. https://doi.org/10.1186/s40468-019-0086-

Publisher's Note