English Language Teaching and Linguistics Studies
ISSN 2640-9836 (Print) ISSN 2640-9844 (Online)
Vol. 5, No. 3, 2023
www.scholink.org/ojs/index.php/eltls

Original Paper

A Critique of the Computer-Based English Speaking Test in Fujian (CEST-FJ)

Wanqing Li[1]

[1] Institute of Education, University College London, London, the United Kingdom

Received: June 12, 2023 Accepted: July 29, 2023 Online Published: August 01, 2023
doi:10.22158/eltls.v5n3p123 URL: http://dx.doi.org/10.22158/eltls.v5n3p123

Abstract

This essay applies Bachman and Palmer's (1996) Test Usefulness Framework to scrutinize the reliability, construct validity, and impact of the Computer-based English Speaking Test in Fujian (CEST-FJ). The CEST-FJ's inclusion in the National College Entrance Examination (NCEE) represents a positive step towards promoting communicative ability in English language learning in China. The computerized administration and scoring processes enhance the test's reliability and control over construct-irrelevant variance. However, limitations such as potential variations in difficulty, construct under-representation of communicative competence, challenges for pronunciation assessment in the context of English as an International Language (EIL), and the non-compulsory and low-value nature of the test have been identified. Additionally, the lack of transparency surrounding the CEST-FJ hampers assessment and accountability. This essay emphasizes the need for more theoretical and empirical research to justify or enhance the test in the future, particularly in light of its high-stakes nature and influence on test-takers' university admissions. Test designers are urged to prioritize transparency and provide more information to facilitate further research and ensure the effectiveness of the CEST-FJ.

Keywords

test usefulness, computer-based English speaking test in Fujian, language testing and assessment, construct validity, reliability, impact

1. Introduction

This essay analyses the oral test in the English section of the National College Entrance Examination (NCEE) in Fujian, China. The NCEE is a standardized test undertaken by final-year high school students in China. It is widely considered as one of the most important examinations in China and serves as the primary determinant for university and college admissions. The test comprises various subjects including but not limited to Chinese, mathematics, English, physics, and history.

Over an extended period of time, the NCEE solely evaluated the English language proficiency of high school students based on a written examination that tested reading, listening, and writing skills, owing to the neglect of communicative ability and technological constraints. However, recent years have witnessed 31 out of 34 provincial districts of China, including Fujian, a coastal province in the southeast of China, encompassed the oral test in the English section of the NCEE, reflecting an increased focus on communication skills and technological advancements.

It is worth noting that this study focuses exclusively on the oral English test in Fujian. This is due to the fact that the format and content of the NCEE English oral test can differ significantly across provinces. For example, Guangdong, a coastal province in southern China, combines the oral and listening tests in a computer-based format, and the oral test score is directly calculated into the total score of the NCEE (People's Government of Guangdong Province, 2018). In contrast, Fujian administers a separate oral test and incorporates the listening test into the written exam. Moreover, the oral test is not compulsory in Fujian, and its score does not contribute to total score of the NCEE (EEAFJ, 2022).

This essay aims to analyse the oral test in the English section of the NCEE in Fujian for several reasons. First, speaking is a fundamental component of foreign language proficiency (Alderson & Bachman, 2004). Hence, it is essential to investigate the effectiveness of the oral test in assessing students' speaking ability. Second, the inclusion of an oral test for the English part of the NCEE represents a significant change in the educational environment of China. Given the high-stakes nature of the test, this change could have a substantial impact on all stakeholders, including students, teachers, and society at large. Thus, it is important to explore the strengths and weaknesses of the oral test for further improvement. Finally, despite the growing body of research on the oral test in other provinces such as Guangdong and Shanghai, there is a lack of research on the oral test in the English section of the NCEE in Fujian. Therefore, using the Test Usefulness Framework proposed by Bachman and Palmer (1996), this essay can contribute to the limited literature on this topic and provide insight into the assessment of speaking skills in the NCEE in Fujian.

2. Description of the Computer-based English Speaking Test in Fujian (CEST-FJ)

The earliest official announcement regarding the implementation of the NCEE English oral test in Fujian was made in 2014, as stated on the official website of Education Examination Authority of Fujian (EEAFJ, 2014). Initially, the English oral test was conducted in-person with examiners from local universities or high schools (EEAFJ, 2016). However, due to the purpose of ensuring fairness and reducing human resources required to implement the in-person test, Fujian switched to a computer-based format from 2020 to date (EEAFJ, 2020). Specifically, Figure 1, developed by Galaczi (2010), illustrates that a test can be classified as human-delivered or computer delivered in terms of delivery method, and human-scored or computer-scored in terms of the scoring approach, thereby forming four quadrants representing four test types. In line with this model, the English oral test in

Fujian can be deemed to have transitioned from quadrant 1 to quadrant 3, from a human-delivered and human-scored to a computer-delivered and computer-scored test. Since numerous provinces in China have switched from in-person oral English test to computer-based oral English test, and the computer-based format is the current format for the English speaking test in Fujian, thus, this essay aims to scrutinize the Computer-based English Speaking Test in Fujian, and will henceforth refer to it as the CEST-FJ.

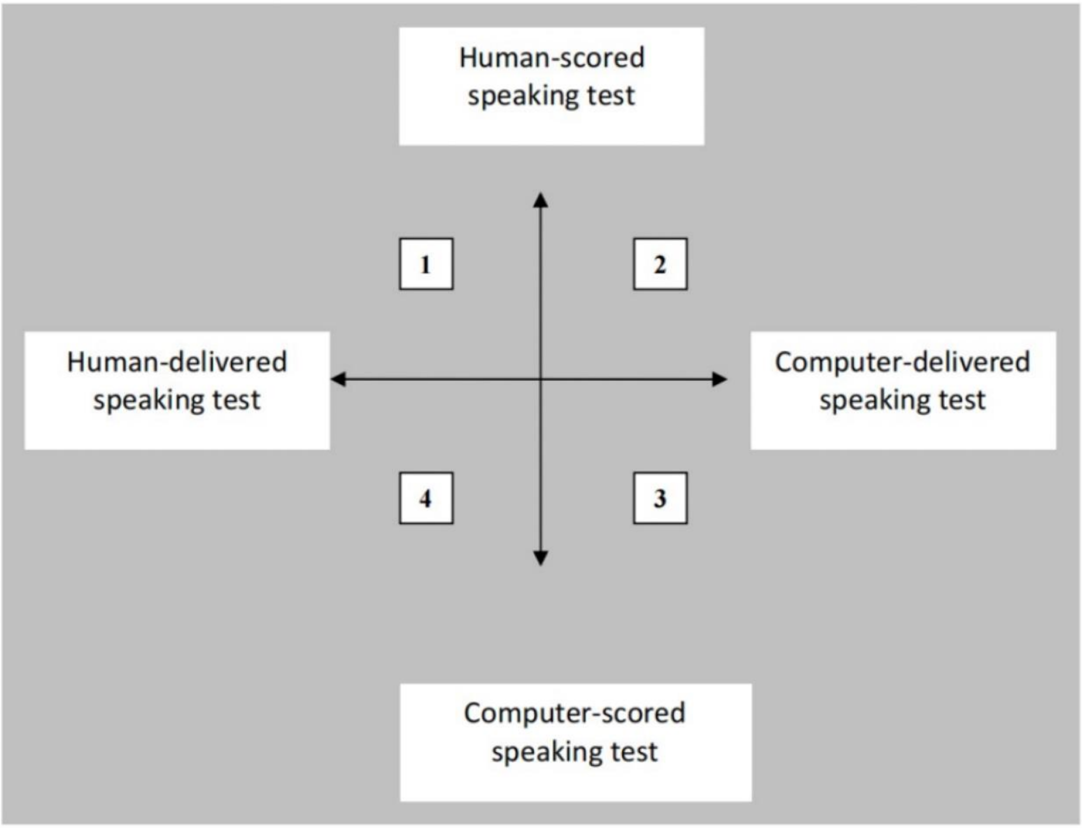

Human-scored speaking test
1
2
Human-delivered speaking test
Computer-delivered speaking test
4
3
Computer-scored speaking test

Figure 1. Delivery and Scoring Possibilities in Speaking Assessment (p. 3)

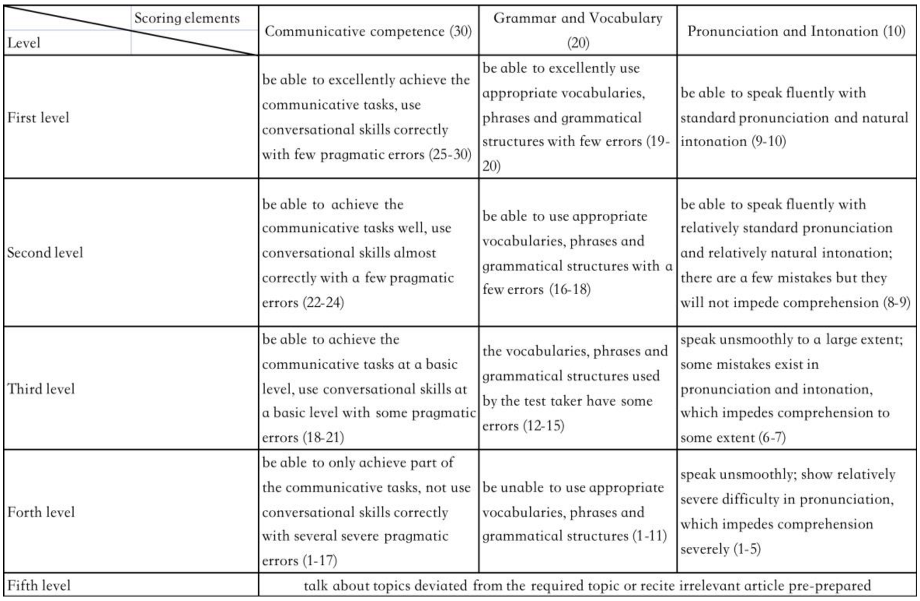
Although the CEST-FJ is not compulsory in Fujian, and its score does not contribute to the CEST-FJ total score, its purpose is to provide a relatively objective reference basis for students’ foreign language reading aloud and speaking ability for majors with special requirements for foreign languages in colleges and universities (EEAFJ, 2016). Thus, certain universities, especially the first-tier universities require applicants to take the CEST-FJ and achieve at least a passing grade if they wish to apply for language-related or commerce-related majors (EEAFJ, 2022). For instance, admission to the French Language and Literature major at Peking University, one of the most outstanding universities in China, in 2022 in Fujian requires applicants to pass the CEST-FJ (Fujian College Entrance Examination Information Platform, n.d.; EEAFJ, 2022).

The official website of the authority provides only the names and scoring criteria of the two components comprising the test, without offering any detailed information on their contents or sample test papers (EEAFJ, 2022). Furthermore, there is a dearth of literature discussing this test. Consequently, to describe the specific content of the CEST-FJ and provide sample exam questions, the present study relies on non-public teaching materials provided by an incumbent senior-three English teacher in Fujian. However, this lack of accessible and transparent information from the authoritative institute could lead to undesirable consequences. This issue will be examined in greater detail in the conclusion sections of this essay.

The CEST-FJ consists of two parts, both of which are completed by the examinee facing the computer. The first part is an ‘answering questions’ section. The computer presents on screen an English question with the audio playing, and the examinee has one minute to formulate an English response. This ask-and-answer cycle is repeated three times. The entire section contains three questions. The second part of the exam is the ‘passage reading aloud’ section, where a brief passage is presented on the computer screen, and the examinee is required to read it out within a specified time frame.

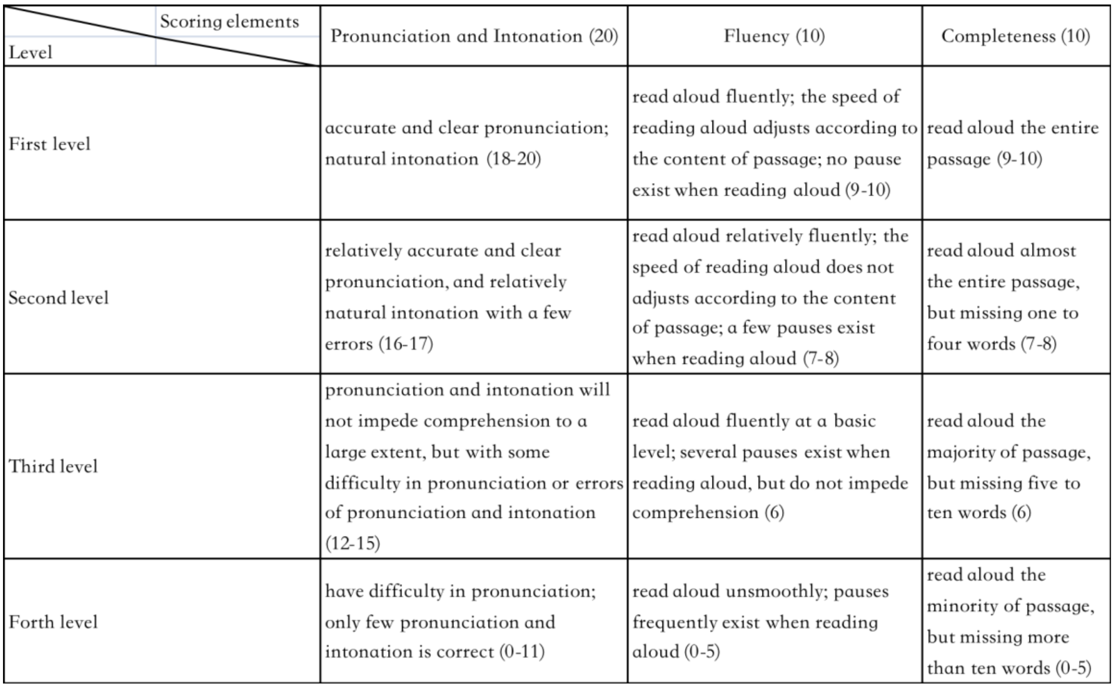
The CEST-FJ is marked entirely by computer according to uniform scoring criteria. The scoring criteria for each of the two test components are outlined in the accompanying table1 and table2. In the question answering section, scores are based on three categories: communicative competence (30%), grammar and vocabulary (20%), and pronunciation and intonation (10%). Meanwhile, the score for the passage reading section is determined by pronunciation and intonation (20%), fluency (10%), and completeness (10%). The two sections use a five-rank and a four-rank grading system respectively based on the examinee’s performance in each aspect, with scores assigned to corresponding grade ranks. The candidate’s final score is calculated by summing all scores obtained.

Table 1. Scoring Criteria for the Question Answering Section of the CEST-FJ (Translated from EEAFJ, 2022)



Scoring elements Level	Communicative competence (30)	Grammar and Vocabulary (20)	Pronunciation and Intonation (10)
First level	be able to excellently achieve the communicative tasks, use conversational skills correctly with few pragmatic errors (25-30)	be able to excellently use appropriate vocabularies, phrases and grammatical structures with few errors (19-20)	be able to speak fluently with standard pronunciation and natural intonation (9-10)
Second level	be able to achieve the communicative tasks well, use conversational skills almost correctly with a few pragmatic errors (22-24)	be able to use appropriate vocabularies, phrases and grammatical structures with a few errors (16-18)	be able to speak fluently with relatively standard pronunciation and relatively natural intonation; there are a few mistakes but they will not impede comprehension (8-9)
Third level	be able to achieve the communicative tasks at a basic level, use conversational skills at a basic level with some pragmatic errors (18-21)	the vocabularies, phrases and grammatical structures used by the test taker have some errors (12-15)	speak unsmoothly to a large extent; some mistakes exist in pronunciation and intonation, which impedes comprehension to some extent (6-7)
Forth level	be able to only achieve part of the communicative tasks, not use conversational skills correctly with several severe pragmatic errors (1-17)	be unable to use appropriate vocabularies, phrases and grammatical structures (1-11)	speak unsmoothly; show relatively severe difficulty in pronunciation, which impedes comprehension severely (1-5)
Fifth level	talk about topics deviated from the required topic or recite irrelevant article pre-prepared		

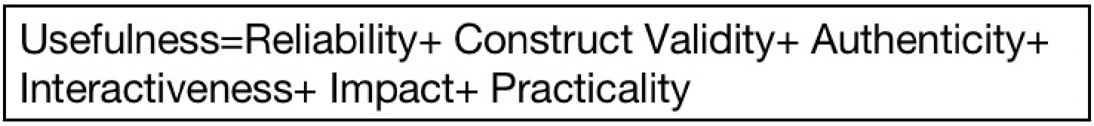
Table 2. Scoring Criteria for the Passage Reading Aloud Section of the CEST-FJ (Translated from EEAFJ, 2022)



Scoring elements Level	Pronunciation and Intonation (20)	Fluency (10)	Completeness (10)
First level	accurate and clear pronunciation; natural intonation (18-20)	read aloud fluently; the speed of reading aloud adjusts according to the content of passage; no pause exist when reading aloud (9-10)	read aloud the entire passage (9-10)
Second level	relatively accurate and clear pronunciation, and relatively natural intonation with a few errors (16-17)	read aloud relatively fluently; the speed of reading aloud does not adjusts according to the content of passage; a few pauses exist when reading aloud (7-8)	read aloud almost the entire passage, but missing one to four words (7-8)
Third level	pronunciation and intonation will not impede comprehension to a large extent, but with some difficulty in pronunciation or errors of pronunciation and intonation (12-15)	read aloud fluently at a basic level; several pauses exist when reading aloud, but do not impede comprehension (6)	read aloud the majority of passage, but missing five to ten words (6)
Forth level	have difficulty in pronunciation; only few pronunciation and intonation is correct (0-11)	read aloud unsmoothly; pauses frequently exist when reading aloud (0-5)	read aloud the minority of passage, but missing more than ten words (0-5)

3. The Test Usefulness Framework of Bachman and Palmer (1996)

Bachman and Palmer (1996) proposed a framework, emphasizing the importance of multiple qualities in determining the usefulness of a test. Specifically, test usefulness is understood as a function of six interrelated qualities showed in Figure 2: reliability, construct validity, authenticity, interactiveness, impact, and practicality.



Usefulness=Reliability+ Construct Validity+ Authenticity+ Interactiveness+ Impact+ Practicality

Figure 2. The Test Usefulness (Bachman & Palmer, 1996, p. 18)

The Bachman and Palmer (1996) Test Usefulness Framework is a well-established and widely used framework to evaluate language tests. This framework has been adopted in numerous studies to assess the quality of various language tests, including Quaid's (2018) evaluation of the usefulness of the International English Language Testing System's (IELTS) speaking test in East Asia. Since this framework provides a comprehensive checklist of qualities that need to be investigated and the types of evidence required to support claims about the quality of the test (Im et al., 2019), it is appropriate for tests such as the CEST-FJ, which lack sufficient research, to analyse the most essential elements in this framework to provide a foundation for future research.

For the purpose of this essay, I have chosen to analyse and evaluate three qualities of the CEST-FJ: reliability, construct validity, and impact. Bachman and Palmer (1996) stated that tests must possess reliability and construct validity as they are fundamental measurement properties that provide the principal justification for using test scores to draw inferences or make decisions. Moreover, for high-stakes tests such as the CEST-FJ, reliability and construct validity are typically the top priority for test developers and evaluators (Bachman & Palmer, 1996). In addition to reliability and construct validity, impact is also a crucial quality to consider when evaluating the usefulness of a language test. In the context of the high-stakes nature of the NCEE, the inclusion of an English speaking test could have significant implications for teaching and society. Therefore, the impact of the test on various stakeholders, such as teachers, students, and society, should be carefully examined.

Another reason why I chose these three aspects is that they are closely interrelated. Moss (1994) argued that reliability was just one crucial element of construct validity. Messick (1996) broadened the definition of construct validity by incorporating washback (i.e., impact on teaching and learning) into the consequential aspect of construct validity. This demonstrates that reliability and impact are closely linked to construct validity.

In conclusion, this essay will analyse the qualities of reliability, construct validity, and impact of the CEST-FJ, as these three qualities are crucial to consider when evaluating the usefulness of a language test, especially in the high-stakes context of the NCEE. However, it is worth emphasizing that evaluating a language test comprehensively requires considering all six qualities (Bachman & Palmer, 1996). Therefore, while this essay has only covered three aspects due to word count limitations, it is important to recognize the interplay and importance of all six qualities in evaluating language tests.

4. Reliability

Reliability refers to the consistency of test scores, both among equivalent tests and among test takers (Bachman & Palmer, 1996). In other words, if a test is reliable, it should produce the same results every time it is administered to the same group of people, and it should also produce similar results when administered to different groups of people who have similar characteristics. Given the characteristics of the CEST-FJ, its reliability is analysed from four aspects. The first two aspects are 1) interviewer variability and 2) rater variability, two primary types of variation that can pose a threat to the reliability of speaking assessments highlighted by Galaczi (2010). The other two aspects are parallel-form reliability and internal consistency, two commonly used reliability types proposed by Yan and Fan (2021).

To begin with, compared to the in-person version before the reform of the test format, the CEST-FJ performs better on reducing the interviewer variability. Interviewer variability refers to differences in how interviewers deliver the test (Galaczi, 2010). Human interviewers may introduce unwanted consequences that affect students' performance due to the existence of interviewer variance. For instance, Ma and Ma (2019) examined the in-person oral English test section of the NCEE in Ningxia Autonomous Region in China, discovering that different human interviewers in the test had varying accents and provided different instructions, all of which affected test takers' performance. In contrast, the interviews in the CEST-FJ are conducted by computers, thereby removing the potential variability caused by human interviewers. Besides, the government of Fujian has constructed standardized computer examination rooms with uniform equipment, delivering standardized test prompts (EEAFJ, 2018). This addressed the issue of interviewer variability from both the equipment and the test content perspectives. Thus, given the concerns regarding interviewer variability caused by human interviewers, it could be more reliable for the CEST-FJ to use computers to deliver the test, thus eliminating the potential for such variability.

Similar to reducing interviewer variability, the CEST-FJ also fights against rater variability by eliminating the rater effects caused by human raters. Rater variability refers to differences in how different raters score the test (Galaczi, 2010). Rater effects, as defined by Scullen et al. (2000), refer to systematic variations in performance ratings associated with the rater and not the actual performance of the person being rated. Factors such as rater severity or leniency, restriction to subsets of rating scale categories, and rater fatigue can contribute to rater effects (Myford & Wolfe, 2003, 2004). Davis (2021) has highlighted, rater severity persists even among pools of well-trained raters. Thus, the use of machine scores is particularly advantageous in avoiding rater variability. In the case of the CEST-FJ, all students are evaluated using the same scoring system and the same criteria by the computer, eliminating the rater variability. This is in contrast to the in-person oral English test in Ningxia, where raters were only trained for two to three hours and had to evaluate 22 students in 50 minutes (Ma & Ma, 2019), leading to difficulty in removing rater effects. Consequently, compared to human raters, the use of computer raters in the CEST-FJ offers a more reliable and unbiased approach to scoring oral English tests, which provides consistency and fairness to all students.

In addition to the reduction of interviewer and rater variability, maintaining parallel-form reliability is also critical in test design. Parallel-form reliability measures the consistency of scores obtained from different but comparable forms of the same test (Yan & Fan, 2021). In other words, parallel-form reliability is a way to assess how consistent the scores are when individuals take two or more forms of the same test that are designed to be equivalent in terms of content, format, difficulty and other factors. Using different test forms inappropriately can affect the test's reliability by leading to varying levels of difficulty for test-takers. For example, Ma and Ma (2019) found that unfamiliar idioms and expressions, e.g., "cool your heels in jail", in one of the passages of the 2019 oral English test in Ningxia could impact reliability. The CEST-FJ employs multiple test forms to prevent cheating, but it is important to ensure that the difficulty of questions in each form is comparable. Indicated in the unpublished teaching materials (see Appendix 2) obtained from a senior-three high school English teacher in Fujian, the questions' difficulty levels are generally consistent and relate to high school students' daily life, such as "Can you tell us something about your hobby?" and "How do you spend your Spring Festival?".

However, differences in background and experiences of test takers may impact the difficulty of certain questions. For example, the question "Which place do you like to travel to most? Why?" may be easy for students from urban, affluent families but challenging for those from rural, economically disadvantaged backgrounds who have never traveled. Thus, it is crucial to address the potential impact of background differences on the parallel-form reliability of the test in the design of the CEST-FJ.

It is worth highlighting that the Education Examination Authority of Fujian (EEAFJ) has not provided a comprehensive introduction to ensuring internal consistency through test design, nor have they published any relevant research data. Internal consistency refers to the degree of consistency among different items or questions in a test that are intended to measure the same construct or attribute (Yan & Fan, 2021). However, some other provinces of China have conducted preliminary studies on the oral English test in the NCEE and have introduced methods to ensure internal consistency. For instance, Cao (2019) and Liu and Zhang (2020) have explored the computer-based English listening and speaking test of the NCEE in Guangdong Province using the many-facets Rasch measurement theory, and demonstrated that the machine score has more advantages than the expert score in terms of internal reliability. These findings provide valuable evidence for the reliability of computer-based listening and speaking examinations in Guangdong Province. Therefore, further research should be conducted to validate the internal consistency of the CEST-FJ.

In summary, the CEST-FJ exhibits a significant degree of reliability owing to its computerized administration and scoring processes, as well as consistent test forms. However, the test's difficulty may vary due to differences in the backgrounds and experiences of the test-takers, necessitating careful consideration in test design. Furthermore, the lack of information on validating internal consistency represents a potential limitation of the CEST-FJ, requiring further empirical investigation in the future.

5. Construct Validity

In addition to reliability, construct validity is an essential quality of tests. According to Bachman and Palmer (1996), construct refers to the underlying ability that a given test or test task is based on and the ability to interpret scores derived from this task. In their definition, construct validity concerns 'the meaningfulness and appropriateness of the interpretation that [are] made on the basis of test score' (p. 21). Therefore, to evaluate the construct validity of a language test, the fundamental question that needs to be answered is whether the test score reflects the intended language ability. As for the CEST-FJ, its construct validity is threatened from four aspects: construct under-representation, construct-irrelevant variance, the unmatched scoring criteria and challenge brought by English as an International Language (EIL).

According to the American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (2014), construct under-representation denotes the degree to which a test fails to capture crucial components of the construct being evaluated. In other words, construct under-representation occurs when a test does not adequately measure all of the important components of the construct being assessed. As stipulated in the official document published by the EEAFJ (2016), the purpose of the CEST-FJ is to furnish an objective and reliable reference point for assessing the English reading aloud and speaking ability of students vying for admission to majors with high requirement for English in colleges and universities, e.g., English Language and Literature, International Commence, etc.

Although the CEST-FJ appears to evaluate test-takers' reading aloud ability through the reading aloud section and examine their speaking ability through the answering questions section, it missed certain important components in the speaking ability it aims to measure. For example, English majors are required to be able to communicate on simple daily topics prior to matriculation (Foreign Language Teaching Steering Committee of Colleges and Universities, 2000). Although the topics included in the exam indeed cover daily topics of high school students, the test only assesses the test-taker's ability to answer questions. Communication, however, is interactive and requires active participation from both parties instead of one-way output. In contrast, the Cambridge B1 Preliminary for Schools Exam's speaking section serves as an effective example to evaluate mutual communication as opposed to one-way statements, as two test takers are paired together and asked to engage in a discussion on a given topic, such as deciding on the most interesting activity for students on a trip to do. Thus, the speaking construct aimed to be measured through the CEST-FJ is underrepresented, and the test in future could benefit from incorporating artificial intelligence to enable test-takers to engage in interactive communication with the computer or with each other. This would more effectively assess speaking ability, including skills such as questioning, answering, and negotiating, rather than solely focusing on answering questions.

In addition to construct under-representation, construct-irrelevant variance is another threat to construct validity. Messick (1984) argued that educational achievement tests may contain elements beyond the intended psychological constructs of knowledge and skills, which may impact test scores. These extraneous factors, referred to as construct-irrelevant variance, pertain to any variance in scores that is not attributable to the intended construct being assessed (AERA, APA, & NME, 2014). The CEST-FJ successfully controls for some construct-irrelevant variance from four perspectives. Firstly, using computers as interviewers helps to mitigate issues caused by unclear pronunciation from examiners that could negatively impact examinees' understanding of questions. Secondly, displaying questions on the screen along with verbal presentation aims to reduce errors attributable to poor listening skills. Thirdly, selecting topics familiar to high school students circumvents issues arising from lack of topical knowledge. Fourthly, examinees are unlikely to face significant challenges related to computer familiarity as they are not required to operate the computer during the test, the program operates automatically.

However, despite these efforts, the CEST-FJ test suffers from the issue of construct-irrelevant variance from two additional perspectives. First, as mentioned previously, topics such as travel may disadvantage candidates from underprivileged family backgrounds who lack travel experience, leading to potentially inaccurate interpretations of their scores. Second, the absence of a human interlocutor during the test is uncertain in its effect on examinees. Qian (2009) reported that participants in a university setting in Hong Kong preferred face-to-face testing over person-to-machine testing due to the lack of interaction in the latter. However, it should be noted that some students in this study expressed a preference for computer-based testing as it can alleviate test anxiety. Therefore, using computer as interviewer may influence students' affection differently, leading to unreliable measurement of their performance.

The issue of the discrepancy between the constructs measured and the scoring criteria employed is another threat to correctly interpret test-takers' scores. The CEST-FJ scoring criteria predominantly focuses on evaluating the test-takers' communicative competence, which accounts for 40% of the total scores. According to Zhang and Isaacs (2022), communicative competence is composed of two core components: interactional competence and pragmatic competence. Interactional competence refers to the ability to collaboratively construct and accomplish communicative objectives with others, while pragmatic competence pertains to the ability to express intentions in a socially and culturally appropriate manner (Taguchi, 2019). However, the CEST-FJ itself is sufficient for measuring neither interactive skills nor pragmatic competence. The test's unilateral nature involves only answering questions without any interactive component, leading to its deficiency to measure interactive skills. However, interpersonal communication typically involves a bidirectional exchange of information whereby an individual acts as both a transmitter and a recipient of an auditory signal that requires encoding and subsequent decoding by both parties to facilitate effective real-time communication (De Bot et al., 2007). As such, it fails to provide a suitable context for assessing the pragmatic ability of test-takers. Consequently, communication skills on the scoring scale can not be comprehensively tested on the CEST-FJ. This drawback diminishes the ability of the test scores to interpret the constructs of test-takers effectively, resulting in the reduction of construct validity.

The construct validity of the CEST-FJ faces a final challenge posed by EIL. As McKay (2002) has pointed out, English has evolved into a language of wider communication that transcends national or cultural boundaries, serving both local and global needs. This perspective challenges the notion of native speaker superiority and shifts the focus of English language teaching and learning from attaining nativelikeness to achieving communication goals. However, the shift also poses challenges to language assessment. For example, the CEST-FJ scoring criteria include "standard pronunciation" and "natural intonation" (EEAFJ, 2022). Given the context of EIL, it is essential for the CEST-FJ to specify the criteria it employs. In particular, the CEST-FJ should clarify whether it deducts points for non-English or non-American pronunciation, or whether it penalizes the use of Chinglish, a translanguaging phenomenon that combines Chinese and English elements. Such clarification would help ensure that the test-taker's score is not influenced by regional or cultural bias, which could result in a misinterpretation of the score.

In conclusion, the CEST-FJ exhibits several limitations concerning construct validity, which include construct under-representation, construct-irrelevant variance, unmatched scoring criteria, and challenges posed by EIL. Although the test has made efforts to control for some construct-irrelevant variance, the diversity of test-takers' family background and their attitudes towards speaking to a computer may lead to residual construct-irrelevant variance. Moreover, the test has failed to comprehensively measure the speaking ability it aims to assess and the communicative ability stated in its scoring criteria. The EIL context also presents challenges to the test, necessitating clarification of its scoring criteria to avoid regional or cultural bias in test-taker scores. Therefore, to ensure the effectiveness of the test, further examination and validation of the CEST-FJ's construct validity are required.

6. Impact

The impact of the test, which encompasses the positive and negative consequences of testing, both on micro or macro level, depending on its influence on stakeholders such as test takers and teachers, or the educational system and society as a whole, respectively (Bachman & Palmer, 1996). The following paragraphs will analyze the positive and negative impact of the CEST-FJ on individuals and the society.

In the case of the CEST-FJ, it is necessary to analyze its impact on individuals, especially on teachers and students, who are most directly affected by the test. One of the positive impacts of the CEST-FJ is that it is likely to encourage teachers to pay more attention to speaking skills. Previous studies have shown that high schools tend to focus more on teaching knowledge of a language with the goal of achieving high scores in the high-stake written test, neglecting the importance of speaking skills (Cortazzi & Jin, 1996; Qi, 2005). However, Hou (2018), through sending questionnaires to 372 high school English teachers in Shanghai and interviewing 24 of them, found that a large number of high school English teachers increased the time for oral practice in their classes after the introduction of the oral English test of the NCEE in Shanghai. Thus, the CEST-FJ has the potential to positively push teachers to change the current situation and increase their attention to oral English teaching and practice.

The positive impact of the CEST-FJ on teachers and students could be limited due to its non-compulsory nature and low score value. As a result, teachers and students may not pay sufficient attention to the test, which hinders the promotion of communicative ability in English teaching. For instance, Zhang and Bournot-Trites (2021) conducted a study, finding that students in Guangdong province, where the oral English test is compulsory for all senior three high school students, were more likely to acknowledge that preparing for the test enhanced their English speaking competence. Similarly, Xu's (2016, 2021) investigation revealed that some high school teachers, particularly those in senior three, neglected oral English teaching due to the low value placed on it in the oral English test of Shanghai college entrance examination. This neglect of oral English teaching can have a negative impact on English language teaching. Therefore, the nature of the CEST-FJ being optional and having a low score-value may weaken its intended positive influence on teachers and students.

Moreover, the CEST-FJ has the potential to bring about negative effects to teachers and students. As discussed earlier, the issue of construct validity has been raised, as the test only focuses on the ability to respond in communication and ignores the ability to ask questions, which is an essential skill in English learning. However, questioning skills play a significant role in language learning as they enable students to better comprehend new knowledge (Almeida, 2012) and incorporate it into their existing knowledge (Robinson & Song, 2019). This neglect of questioning skills in test design has been shown to negatively impact teaching, as seen in the TOEFL iBT speaking test. Brooks and Swain (2014) highlighted that this test only prioritizes students' responses and disregards their questioning ability, leading to teachers neglecting the development of questioning skills in favor of improving test scores. As Kane and Case (2004) has discovered that content that is not included on the test tends to be neglected or under-emphasized, it is likely that the same issue may arise in the context of the CEST-FJ, and both teachers and students may overlook the importance of improving questioning skills during test preparation, leading to undesirable outcomes.

The impact of the CEST-FJ is not limited to individuals but also extends to society. On the one hand, implementation of this test may have positive implications for society by enhancing the oral communication skills of Chinese English learners, which would prepare them to effectively communicate on an international platform. In the analysis of Chinese government policy documents, Pan (2015) has identified the dual-functional nature of English education in China, which is aimed at improving individual language proficiency while also promoting national development and participation in international communication. Hence, the implementation of high-stakes oral tests such as the CEST-FJ could help raise the significance of oral communication skills among Chinese English learners and support China's involvement in international communication and development.

However, the implementation of the CEST-FJ may also result in negative consequences, particularly in terms of exacerbating educational inequality. As previously discussed regarding reliability, certain test questions may put students from economically disadvantaged families at a disadvantage. Furthermore, the inclusion of oral English test in the college entrance examination or using it as a criterion for university admission could further widen educational inequality. Despite the CEST-FJ's attempts to ensure fairness by standardizing the testing environment and employing controlled question types, Goto Butler et al. (2022) suggested that equal remedial measures do not necessarily result in fair testing, owing to the significant disparity in the accessibility and availability of English language exam preparation resources. Studies have demonstrated that oral proficiency is strongly influenced by students' regional and socioeconomic backgrounds (e.g., Butler & Iino, 2021). Specifically, students from affluent families possess greater access to oral training resources, while students from low-income families may be disadvantaged in terms of accessing such resources, potentially precluding their admission to prestigious universities due to insufficient oral English skills. Therefore, it is crucial for government officials, researchers, and educational institutions to collaboratively consider and mitigate the potential exacerbation of educational inequality that may arise from the implementation of the CEST-FJ.

In conclusion, the impact of the CEST-FJ oral English test is complex, and it can have both positive and negative consequences on various stakeholders, including teachers, students, and society. On the one hand, the test has the potential to positively influence teachers to increase their attention to oral English teaching and practice, which could enhance the oral communication skills of Chinese English learners and support China's participation in international communication and development. On the other hand, the non-compulsory nature and low score value of the test could hinder its intended positive impact on teachers and students, while its neglect of questioning skills and potential exacerbation of educational inequality could have negative effects on English language teaching and society. Therefore, it is essential to critically examine the impact of the CEST-FJ and work collaboratively to mitigate any potential negative consequences.

7. Conclusion

In conclusion, the analysis of the CEST-FJ reveals several strengths and limitations. As speaking is much difficult to assess (Fan & Yan, 2020; Ginther, 2012), it has long been ignored in the NCEE in China. The inclusion of the CEST-FJ, a speaking test, in the high-stakes the NCEE represents a positive step towards promoting communicative ability in English language learning in China. The computerized administration and scoring processes also enhance the test's reliability and control over construct-irrelevant variance. However, several limitations have also been identified, including potential variations in difficulty due to test-takers' backgrounds and experiences, construct under-representation of communicative competence, challenge for pronunciation assessment in the context of EIL, and the potential negative impact due to the non-compulsory and low value nature of the test.

Another significant limitation of the CEST-FJ is its lack of transparency. The official authority of the test has not made available any sample test papers or provided specific details regarding the test content, making it difficult for others to assess the consequences of the test and hold the test designers accountable (Weideman, 2006). This lack of transparency may contribute to the limited research on the CEST-FJ despite its decade-long operation. Additionally, as high-stakes testing is known to have a profound impact on the educational system (Wall, 2005), the CEST-FJ's influence on test-takers' university admissions and subsequent consequences further underscores the significance of this issue (Madaus, 1988). Therefore, given the high-stakes nature of the CEST-FJ, it is crucial for test designers to prioritize transparency and provide more information to serve as a foundation for future research.

Overall, this essay applied Bachman and Palmer's (1996) Test Usefulness Framework to scrutinize the reliability, construct validity and impact of the CEST-FJ, contributing to the limited existing literature on this test. As noted by Bachman and Palmer (1996), test design is one of the few factors that we can control. Therefore, it is imperative for test designers and researchers to conduct a more comprehensive examination of the CEST-FJ, including more theoretical and empirical research to justify or enhance the test in the future, to ensure its effectiveness.

References

Alderson, J., & Bachman, L. (2004). Series editors' preface. In S. Luoma (Author), *Assessing Speaking* (pp. Ix-Xi). Cambridge University Press. https://doi.org/10.1017/CBO9780511733017

Almeida, P. A. (2012). Can I ask a question? The importance of classroom questioning. *Procedia-Social and Behavioral Sciences*, *31*, 634-638. https://doi.org/10.1016/j.sbspro.2011.12.116

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Brooks, L., & Swain, M. (2014). Contextualizing Performances: Comparing Performances During TOEFL iBT ^{TM} and Real-Life Academic Speaking Activities. *Language Assessment Quarterly*, *11*(4), 353-373. https://doi.org/10.1080/15434303.2014.947532

Butler, Y. G., & Iino, M. (2021). Fairness in College Entrance Exams in Japan and the Planned Use of External Tests in English. In Lanteigne, B., Coombe, C., & Brown, J. D. (Eds), *Challenges in Language Testing Around the World*. Springer. https://doi.org/10.1007/978-981-33-4232-3_5

Cao, L. (2019). Comparison of Automatic and Expert Teachers' Rating of Computerized English Listening-Speaking Test. *English Language Teaching*, *13*(1), 18. https://doi.org/10.5539/elt.v13n1p18

Cortazzi, M., & Jin, L. (1996). English teaching and learning in China. *Journal for Language Teaching*, *29*(2), 61-80.

Davis, L. (2021). Rater and interlocutor training. In G. Fulcher, & L. Harding, *The Routledge Handbook of Language Testing* (2nd ed., pp. 322-338). Routledge. https://doi.org/10.4324/9781003220756-25

De Bot, K., Lowie, W., & Verspoor, M. (2007). A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition*, *10*(01), 7. https://doi.org/10.1017/S1366728906002732

Education Examinations Authority of FuJian (EEAFJ). (2022, December 16). *Notice on Oral Examination of Foreign Languages for 2023 College Admissions in Fujian Province*. Retrieved 1 April 2023, from https://www.eeafj.cn/gkptgkgsgg/20221216/12511.html

Education Examinations Authority of FuJian (EEAFJ). (2020, December 3). *Notice on Oral Examination of Foreign Languages for 2020 College Admissions in Fujian Province*. Retrieved 2 April 2023, from https://www.eeafj.cn/gkptgkgsgg/20191203/9766.html

Education Examinations Authority of FuJian (EEAFJ). (2018, December 28). *Notice on the construction of standardized test rooms for the oral English test of the College Entrance Examination in 2019 in Fujian Province*. Retrieved 5 April 2023, from https://www.eeafj.cn/xxhgsgg/20181228/8903.html

Education Examinations Authority of FuJian (EEAFJ). (2016, June 10). *Notice on the issuance of the Opinions on the Implementation of Oral Examinations of Foreign Languages for Admissions of Colleges and Universities in Fujian Province in 2016*. Retrieved 1 April 2023, from https://www.eeafj.cn/syzxyw/20160610/5748.html

Education Examinations Authority of FuJian (EEAFJ). (2014, June 9). *Opinions on the implementation of Oral English in 2014 College Entrance Examination in Fujian Province*. Retrieved 2 April 2023, from https://www.eeafj.cn/gkptgkgsgg/20140609/4068.html

Fan, J., & Yan, X. (2020). Assessing Speaking Proficiency: A Narrative Review of Speaking Assessment Research within the Argument-Based Validation Framework. *Frontiers in Psychology*, *11*, 330. https://doi.org/10.3389/fpsyg.2020.00330

Foreign Language Teaching Steering Committee of Colleges and Universities. (2000). *English teaching syllabus for English Majors in Colleges and Universities*. Foreign Language Teaching and Research Press.

Fujian College Entrance Examination Information Platform. (n.d.). *Peking University 2022 Fujian enrollment plan*. Retrieved 2, April 2023, from http://www.fjgkedu.com/college/162/116/246257.html

Galaczi, E. D. (2010). Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In Ara újo, L (Ed.), *Proceedings of the Computer-based Assessment (CBA) of Foreign Language Speaking Skills* (pp. 29-51). Brussels, Belgium, European Union.

Ginther, A. (2012). Oller, John W., Jr. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics*. Blackwell Publishing Ltd. https://doi.org/10.1002/9781405198431.wbeal0882

Goto Butler, Y., Lee, J., & Peng, X. (2022). Failed policy attempts for measuring English speaking abilities in college entrance exams: Cases from China, Japan, and South Korea. *English Today*, *38*(4), 271-277. https://doi.org/10.1017/S0266078420000346

Hou, Y. (2018). A Study on the Washback Effect of the Reform of SHMET Listening and Speaking Test. *Technology Enhanced Foreign Languages*, *183*, 23-29.

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: Their limitations and future directions for validation research. *Language Testing in Asia*, *9*(1), 14. https://doi.org/10.1186/s40468-019-0089-4

Kane, M., & Case, S. M. (2004). The Reliability and Validity of Weighted Composite Scores. *Applied Measurement in Education*, *17*(3), 221-240. https://doi.org/10.1207/s15324818ame1703_1

Liu, J., & Zhang, B. (2020). Multi-level Rasch Model Analysis of Computer-assisted Automated Scoring of English Listening and Speaking Tests. *2020 International Conference on Computer*

Engineering and Application (ICCEA), 632-636. https://doi.org/10.1109/ICCEA50009.2020.00138

Ma, J., & Ma, L. (2019). An Analysis of Computer-based Oral English Test in College Entrance Examination -- A Case study of Ningxia Autonomous Region. *Journal of Ningxia Normal University*, *40*(5), 78-81.

Madaus, G. F. (1988). The Distortion of Teaching and Testing: High-Stakes Testing and Instruction. *Peabody Journal of Education*, *65*(3), 29-46. http://www.jstor.org/stable/1492818

McKay, S. L. (2002). *Teaching English as an international language: Rethinking goals and approaches*. Oxford University Press.

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241-256. https://doi.org/10.1177/026553229601300302

Messick, S. (1984). The Psychology of Educational Measurement. *ETS Research Report Series*, *1984*(1), i-55. https://doi.org/10.1002/j.2330-8516.1984.tb00046.x

Moss, P. A. (1994). Can There Be Validity Without Reliability? *Educational Researcher*, *23*(2), 5-12. https://doi.org/10.3102/0013189X023002005

Myford, C., & Wolfe, E. (2004). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Myford, C., & Wolfe, E. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of applied measurement*, *4*(4), 386-422.

Pan, L. (2015). *English as a Global Language in China: Deconstructing the Ideological Discourses of English in Language Education*. Springer. https://doi.org/10.1007/978-3-319-10392-1

People's Government of Guangdong Province. (2018, October 31). *Notice of the examination syllabus of the English Listening and Speaking Test in the College Entrance Examination in Guangdong in 2019*. Retrieved 22 April 2023, from http://www.gd.gov.cn/gdywdt/bmdt/content/mpost_160122.html

Qi, L. X. (2005). Stakeholders' conflicting aims undermine the washback function of a high-stakes test. *Language Testing*, *22*(2), 142-173. https://doi.org/10.1191/0265532205lt300oa

Qian, D. D. (2009). Comparing Direct and Semi-Direct Modes for Speaking Assessment: Affective Effects on Test Takers. *Language Assessment Quarterly*, *6*(2), 113-125. https://doi.org/10.1080/15434300902800059

Quaid, E. D. (2018). Reviewing the IELTS speaking test in East Asia: Theoretical and practice-based insights. *Language Testing in Asia*, *8*(1), 2. https://doi.org/10.1186/s40468-018-0056-5

Robinson, S. E., & Song, J. J. (2019). Student academic performance system: Quantitative approaches to evaluating and monitoring student progress. *International Journal of Quantitative Research in Education*, *4*(4), 332-353. https://doi.org/10.1504/IJQRE.2019.100170

Scullen, S. E., Mount, M. K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology*, *85*(6), 956-970.

https://doi.org/10.1037/0021-9010.85.6.956

Taguchi, N. (2019). Second Language Acquisition and Pragmatics: An Overview. In *The Routledge Handbook of Second Language Acquisition and Pragmatics* (pp. 1-14). Routledge. https://doi.org/10.4324/9781351164085-1

Wall, D. (2005). *The impact of high-stakes testing on classroom teaching: A case study using insights from testing and innovation theory*. Cambridge University Press.

Weideman, A. (2006). Transparency and accountability in applied linguistics. *Southern African Linguistics and Applied Language Studies*, *24*(1), 71-86. https://doi.org/10.2989/16073610609486407

Xu, W. (2021). The Practice of oral Test in large-scale high-stakes Examination: Taking Shanghai College Entrance Examination English Listening and Speaking Test as an example. *Foreign Language Testing and Teaching*, *1*, 21-27.

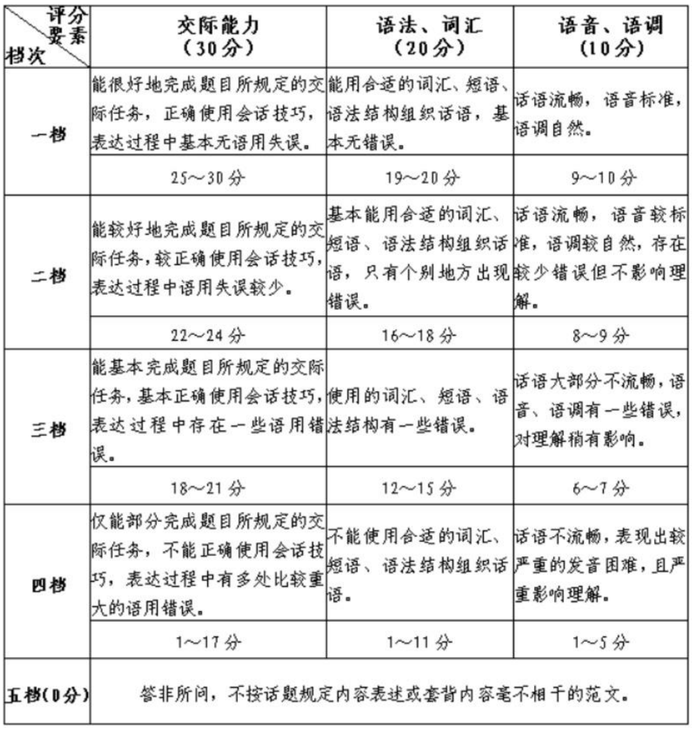
Xu, W. (2016). A discussion on College Entrance Examination English (Shanghai Volume) under the background of a new round of College Entrance Examination reform: Innovation, thinking and prospect. *Foreign Language Testing and Teaching*, *4*, 24-31.

Yan X., & Fan J. (2021). Reliability and dependability. In G. Fulcher & L. Harding, *The Routledge Handbook of Language Testing* (2nd ed., pp. 477-494). Routledge. https://doi.org/10.4324/9781003220756-37

Zhang, H., & Bournot-Trites, M. (2021). The long-term washback effects of the National Matriculation English Test on college English learning in China: Tertiary student perspectives. *Studies in Educational Evaluation*, *68*, 100977. https://doi.org/10.1016/j.stueduc.2021.100977

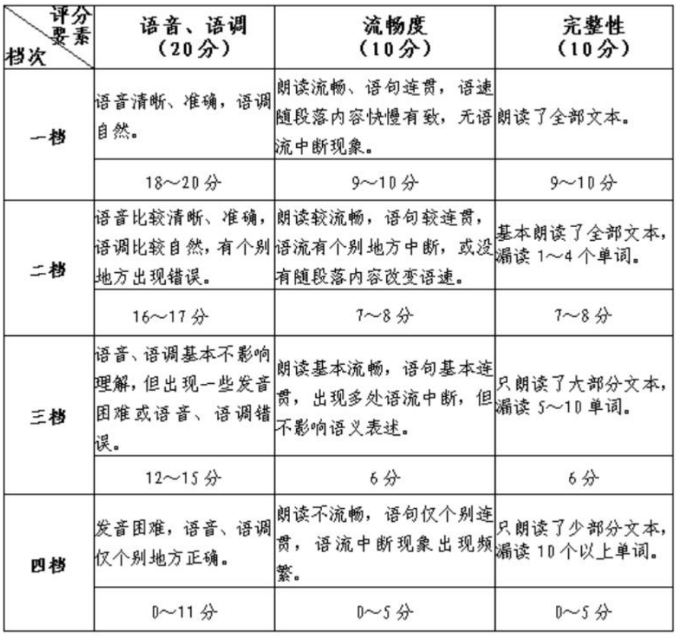
Zhang, S., & Isaacs, T. (2022). Can Interactions Happen across the Screens? In K. Sadeghi, *Technology-Assisted Language Assessment in Diverse Contexts* (1st ed., pp. 196-211). Routledge. https://doi.org/10.4324/9781003221463-16

Appendix 1. Scoring Criteria of the CEST-FJ in Chinese Scoring Criteria of the Question Answering Section



评分要素 档次	交际能力 （30分）	语法、词汇 （20分）	语音、语调 （10分）
一档	能很好地完成题目所规定的交际任务，正确使用会话技巧，表达过程中基本无语用失误。	能用合适的词汇、短语、语法结构组织话语，基本无错误。	话语流畅，语音标准，语调自然。
	25～30 分	19～20 分	9～10 分
二档	能较好地完成题目所规定的交际任务，较正确使用会话技巧，表达过程中语用失误较少。	基本能用合适的词汇、短语、语法结构组织话语，只有个别地方出现错误。	话语流畅，语音较标准，语调较自然，存在较少错误但不影响理解。
	22～24 分	16～18 分	8～9 分
三档	能基本完成题目所规定的交际任务，基本正确使用会话技巧，表达过程中存在一些语用错误。	使用的词汇、短语、语法结构有一些错误。	话语大部分不流畅，语音、语调有一些错误，对理解稍有影响。
	18～21 分	12～15 分	6～7 分
四档	仅能部分完成题目所规定的交际任务，不能正确使用会话技巧，表达过程中有多处比较重大的语用错误。	不能使用合适的词汇、短语、语法结构组织话语。	话语不流畅，表现出较严重的发音困难，且严重影响理解。
	1～17 分	1～11 分	1～5 分
五档(0分)	答非所问，不按话题规定内容表述或套背内容毫不相干的范文。		

Scoring Criteria for the Passage Reading Aloud Section



评分要素 档次	语音、语调 （20分）	流畅度 （10分）	完整性 （10分）
一档	语音清晰、准确，语调自然。	朗读流畅、语句连贯，语速随段落内容快慢有致，无语流中断现象。	朗读了全部文本。
	18～20 分	9～10 分	9～10 分
二档	语音比较清晰、准确，语调比较自然，有个别地方出现错误。	朗读较流畅，语句较连贯，语流有个别地方中断，或没有随段落内容改变语速。	基本朗读了全部文本，漏读 1～4 个单词。
	16～17 分	7～8 分	7～8 分
三档	语音、语调基本不影响理解，但出现一些发音困难或语音、语调错误。	朗读基本流畅，语句基本连贯，出现多处语流中断，但不影响语义表述。	只朗读了大部分文本，漏读 5～10 单词。
	12～15 分	6 分	6 分
四档	发音困难，语音、语调仅个别地方正确。	朗读不流畅，语句仅个别连贯，语流中断现象出现频繁。	只朗读了少部分文本，漏读 10 个以上单词。
	0～11 分	0～5 分	0～5 分

Appendix 2. Sample Tests for the "Answering Questions" part

Test form1

1 How do you spend your Spring Festival?

2 Can you say something about the people in your hometown?

3 Do you like reading English novel? Why or why not?

Test form 2

1 Which do you like best, Chinese food or western food? Why?

2 How do you like to celebrate your birthday?

3 Can you tell us something about your hobby?

Test form 3

1 What kind of school activities do you like best? Why?

2 Do you like watching TV during the summer vacation? Why?

3 Which place do you like to travel most? Why?