VNU Journal of Foreign Studies

Journal homepage: https://jfs.ulis.vnu.edu.vn/

EXAMINERS' PERCEPTIONS OF A COMPUTER-ASSISTED ENGLISH LANGUAGE TEST

Nguyen Thi Phuong Thao*

Office of Quality Assurance, VNU University of Languages and International Studies, No. 2 Pham Van Dong, Cau Giay, Hanoi, Vietnam

Received 13 December 2025
Revised 15 March 2026; Accepted 21 July 2026

Abstract: Computer-assisted language testing (CALT) is of great development in the last ten years thanks to the technological advances in the 4.0 era and due to the COVID-19. Increasingly, test takers - including young learners - are becoming accustomed to completing both standardized and classroom-based assessments online or with digital support. Widely recognized platforms such as IELTS, APTIS, and Duolingo exemplify this shift toward digital testing environments. Simultaneously, educators are gaining experience in transitioning from traditional paper-based assessments to computer-based formats. This evolution raises important questions about how examiners perceive and assess language performance in digital contexts. It is noteworthy to listen to examiners' voices in regard to evaluating performances on computers or with computer support (Forrester, 2020; Yulianto & Mujtahid, 2021). This study investigates Vietnamese examiners' perceptions on a specific computer-based English proficiency test. The qualitative data collected from focus-group discussions with six examiners provided insightful information on their perceptions and experiences on the test. Their insights offer a nuanced understanding of how they experienced digital interfaces, and how test takers are assessed with such an assessment mode. The findings from this research will contribute to a deeper understanding of the human dimension in CALT, highlighting both opportunities and challenges in digital assessment, especially the role of examiner agency and reflective practice in shaping reliable assessment.

Keywords: computer-assisted language testing, examiners' perceptions, speaking test

* Corresponding author.
Email address: phuongthaonguyen310@gmail.com
https://doi.org/10.63023/2525-2445/jfs.ulis.5691

GÓC NHÌN CỦA GIÁM KHẢO VỀ MỘT BÀI THI ĐÁNH GIÁ NĂNG LỰC TIẾNG ANH TRÊN MÁY TÍNH

Nguyễn Thị Phương Thảo

Phòng Quản trị chất lượng, Trường Đại học Ngoại ngữ, Đại học Quốc gia Hà Nội,
Số 2 Phạm Văn Đồng, Cầu Giấy, Hà Nội, Việt Nam

Nhận bài ngày 13 tháng 12 năm 2025
Chỉnh sửa ngày 15 tháng 3 năm 2026; Chấp nhận đăng ngày 21 tháng 7 năm 2026

Tóm tắt: Kiểm tra ngôn ngữ hỗ trợ bằng máy tính (CALT) đã có những bước phát triển vượt bậc trong mười năm qua, nhờ vào tiến bộ công nghệ trong thời đại 4.0 và sự ảnh hưởng từ đại dịch COVID-19. Ngày càng nhiều thí sinh, bao gồm cả học sinh nhỏ tuổi, đã quen với việc thực hiện các bài kiểm tra tiêu chuẩn hóa và kiểm tra trong lớp học thông qua hình thức trực tuyến hoặc có sự hỗ trợ của máy tính như tham gia các bài thi IELTS, APTIS, Duolingo, v.v. Đồng thời, các giám khảo cũng đang tích lũy kinh nghiệm trong việc chuyển đổi từ hình thức kiểm tra trên giấy truyền thống sang định dạng kiểm tra trên máy tính. Việc lắng nghe tiếng nói của giám khảo trong việc đánh giá bài làm trên máy tính hoặc có sự hỗ trợ của công nghệ là vô cùng cần thiết (Forrester, 2020; Yulianto & Mujtahid, 2021). Nghiên cứu này° tập trung khảo sát nhận thức và trải nghiệm của các giám khảo đối với một bài kiểm tra năng lực tiếng Anh trên máy tính tại Việt Nam. Dữ liệu định tính được thu thập thông qua thảo luận nhóm với sáu giám khảo, nhằm cung cấp những thông tin sâu sắc về cảm nhận và trải nghiệm của họ với vai trò cán bộ chấm thi. Những phản hồi này góp phần làm sáng tỏ khía cạnh con người trong CALT, đồng thời chỉ ra những cơ hội và thách thức trong đánh giá số - đặc biệt là vai trò của giám khảo và thực hành phản hồi trong việc xây dựng hệ thống đánh giá đáng tin cậy.

Từ khóa: kiểm tra đánh giá trên máy tính, góc nhìn của giám khảo, bài thi nói

1. Introduction

Thanks to the surge in technology, education has witnessed critical changes in terms of the way it is operated. Regarding testing and assessment, technology has changed the way related activities are administered. More and more high-stake tests have been computerized, which brings the chance of test takers being administered in both paper-based mode and computer-based tests. Computer-assisted language testing (CALT) has been familiarized among test takers, examiners, test administrators and other groups of stakeholders in large-scale language testing contexts. In such a prominent change, understanding the perspectives of examiners who play the role in evaluating spoken responses has become crucial. Examiners' judgments play a central role in ensuring the validity and reliability of speaking test scores; however, their experiences and perceptions of CALT remain underexplored compared to traditional, face-to-face assessments.

Current studies on language testing have thoroughly examined technical aspects and test-taker performance in CALT; however, there has been less focus on how examiners adjust to and interpret this new format. Elements such as interface design, audio clarity, and the lack of immediate interaction with test takers may affect the scoring processes used by examiners, as

° This research was funded by VNU University of Languages and International Studies (VNU-ULIS) in the project No. N.26.02

well as their views on fairness and authenticity in assessment. Exploring examiners' viewpoints is thus crucial to shed light on the intricacies of human judgment within the digital testing landscape and to inform ongoing enhancements in test design, rater training, and quality management.

This study addresses a potential gap in language assessment literature by focusing on examiners' viewpoints toward CALT, particularly in speaking tests. The objective is to enhance understanding of their challenges, attitudes, and strategies. By foregrounding examiners ' voices, the research seeks to contribute to the development of more effective and computer-assisted speaking assessments in a growing field of language testing. To achieve the research aim, this study investigates how examiners perceived a specific computer-assisted speaking test in the context of Vietnam by answering the following questions:

1. What common themes characterize examiners' perceptions of computer-assisted speaking tests?

2. What divergent perceptions do examiners hold regarding computer-assisted speaking tests?

2. Literature Review

2.1. Computer-Assisted Speaking Tests

Computer-assisted speaking tests (CAST) belong to the system of CALT and focus on the oral assessment. CAST involves utilizing digital tools and technology to assess a person's oral language skills. This is categorized into various types based on design, purpose, and implementation, covering automated speaking tests, semi-automated tests, and human-rated computer-assisted tests (Hu et.al., 2025). This approach typically includes activities like replying to cues, engaging in mock dialogues, or documenting talks. Test takers are expected to interact with the computer screen or an artificial intelligence (AI) interlocutor. The assessment can be conducted promptly with automated scoring or with human examiners. According to Galaczi (2010), CAST is not a replacement for face-to-face approach in assessing oral competency, but "a new additional possibility." Instead of participating in the whole procedure as both interlocutors and examiners, humans can take part in certain stages, normally at the rating phase in which automated scoring is proved not to be able to replace humans totally. Raters significantly influence the validity and reliability of speaking test scores through their judgement of oral language performance. The transition from face-to-face to CAST environments introduces additional complexities, including the absence of real-time interaction, reliance on recorded speech, and the technological interface, which may affect raters' scoring processes and confidence (Dermo, 2009; Yulianto & Mujtahin, 2021).

2.2. Stakeholders' Perspectives on Computer-Assisted Speaking Tests

In the field of CAST, stakeholders often include different groups of people who directly or indirectly get involved in the process namely test takers, examiners as interlocutors and raters, test administrators, and others. Even though this study focuses on examiners' perceptions, the literature review is going to present how other groups of stakeholders have shared their beliefs and experiences regarding CAST based on typical features of the test qualities which Bachman and Palmer (1996), and Kunnan (2004) proposed. According to Bachman and Palmer (1996), test usefulness includes six qualities namely reliability, construct validity, authenticity, interactiveness, impact and practicality. Later, Kunnan (2004) supported the test fairness framework which covered validity, absence of bias, access, administration, and social consequences. In the context of this study, the qualities are selected as below to meet the research objectives and setting.

2.2.1. Test Validity

As we all know, validity refers to whether the test can measure what it aims to measure (Galaczi, 2020). The oral competence in English, in other words, face validity is believed to maintain in the computer-based tests according to test takers participating in a survey conducted by Pizarro and Laborda (2017). They completed the Aptis computer-delivered speaking tests and provided responses to an open questionnaire. The quantitative results showed participants held overall positive views on computer-test validity as a valid and proper measure of oral competence. This study shared common features with Malone (2014) investigating stakeholders' beliefs about the TOEFL iBT's speaking tests. Surveyed participants including students, instructors and administrators reported rather high confidence in the test tasks as valid indicators of English language ability for academic contexts. Besides advantages, there are some concerning issues related to the test validity. Zhan and Wan (2016) presented an investigation of test takers' attitudes and beliefs of a high-stake computer-based English-speaking test in China. The data showed that students had a distinct understanding of what the test measured. However, while the test designers focused on communicative competence, two thirds of the students believed the test required and measured unintended skills such as translation, summary writing and note taking. It meant that there might exist constructs assessed beyond the test.

A number of studies highlight issues related to construct validity in CASTs, emphasizing whether the tests accurately measure speaking ability as intended. According to Hu et al. (2025), in their systematic review of computer-based English speaking tests, raters and test takers perceived CASTs as valid but point to technical and design factors that can threaten this validity, such as audio quality and test interface usability. Isbell (2024) examined how the Duolingo English is perceived by university stakeholders, and reported that concerns revolved around content and construct validity, with some stakeholders desiring more transparency on scoring criteria and task relevance. CASTs were also questioned regarding the limitations of task types as well as the absence of interlocutors which might lead to the lack of interactional competence (Elder & Iwashita, 2005; Chapelle & Douglas, 2006). Test takers do not have enough chances to demonstrate their speaking ability compared to being interviewed by a teacher. In various instances, Chapelle and Douglas (2006), Jeong et al. (2011), and Zhou (2015) noted that computer-assisted oral assessments might restrict the variety of task types and limit the assessment construct due to the absence of an interactive element. Furthermore, this mode of delivery fails to reflect the intricacies of natural language usage as effectively as a face-to-face interview format (Douglas & Hegelheimer, 2007).

2.2.2. Test Reliability

Test reliability refers to the scoring consistency, which should be maintained in various test conditions. In the shift from face-to-face interview to computer-based speaking, it is open to question whether the reliability can be maintained, or even improved. One important point in computerized tests is that raters can not see test takers in person, therefore they are not influenced by test takers' characteristics, especially facial expressions and non-verbal languages (Pizarro & Laborda, 2017). Koizumi et al. (2017) also pointed out raters reached high agreement and self-consistency in the Global Test of English Communication Computer Based Testing despite a small degree of severity. This can be partly explained in the aspect that typical errors, namely harshness or leniency, central tendency can be controlled in automated assessment.

On the other hand, while automated systems generally exhibit high inter-rater reliability, they may be a little bit more lenient compared to human examiners. (Hayashi & Kondo, 2024).

Also, the lower intra-rater reliability of computer-based rating compared to human raters is also raised in the previous study of Liu and Zhang (2020). It can be seen that there exist concerns regarding both the inter-rater and intra-rater reliability of computer-based speaking tests.

2.2.3. Test Fairness

Research suggests that virtual assessment offers significant advantages in both procedural efficiency and perceived fairness. For example, a comparability study of the Vietnamese standardized English proficiency tests (VSTEP) across B1 to C1 levels, conducted by Nguyen et al. (2024), found that the computer-delivered speaking test format (semi-direct) enhanced fairness when compared to the traditional face-to-face mode.

A primary reason for this preference was the reduction of potential rater bias: test takers felt that when raters were unable to see the examinee's face, they were less influenced by facial expressions and gestures, thus minimizing rater subjectivity and variation. Furthermore, equal time allocation for every test taker, which is standard in the computer-delivered environment, was also noted as a factor that boosted fairness.

Procedurally, examiners benefit significantly from technical support, which Yulianto and Mujatahin (2021) showed makes the distribution of testing materials easier. Moreover, several studies (Dreher et al., 2011; Baleni, 2015) indicated that educators highly value virtual assessment because it allows automatic grading and cuts down on administrative expenditures.

2.2.4. Test Impact

Some test takers perceive CBSA positively, citing several benefits. Kenyon and Malabonga (2001) found that participants generally preferred computerized oral tests because they were seen as less challenging, offered a more equitable set of questions, and reduced anxiety, ultimately allowing for a more accurate representation of proficiency. Beyond test-taker experience, virtual assessment offers strong administrative benefits. Studies showed that examiners highly value web-based systems for their automatic grading capabilities and significant reduction in administrative expenditures (Dreher et al., 2011; Baleni, 2015). Furthermore, Yulianto and Mujtahin (2021) highlighted that technical support can significantly ease the distribution of testing materials, contributing to greater procedural efficiency.

Despite the benefits, significant concerns exist regarding CBSA, primarily related to motivation, anxiety, and technical issues. The lack of a human interlocutor and face-to-face interaction can reduce test-taker motivation (Elder and Iwashita, 2005). A major issue is the increased anxiety reported when test takers are unable to read non-verbal feedback from an examiner (Quaid and Barett, 2021), which can result in lower fluency. Communication-oriented test takers frequently prefer the traditional interview, as they value the two-way interaction and timely verbal and non-verbal feedback that the "emotionless" computer screen fails to provide (Nguyen et al., 2024). Technical constraints also pose a threat: concurrent speaking by other test takers can create noise and distraction, leading to low-quality recordings (Nguyen et al., 2024). For examiners, general web-based assessment systems raise concerns about validity, practicality, and reliability (Dermo, 2009; Yulianto & Mujtahin, 2021)

In summary, computer-assisted speaking tests have been reported with both advantages and disadvantages in different aspects of test qualities. Stakeholders generally view CBSA as valid, but concerns arise about construct validity due to limited interaction and task variety. Reliability shows mixed results; raters have high agreement though systems may be more lenient. CBSA also improves fairness by minimizing rater bias and ensuring equal time. As can be seen from the review of previous studies, more studies have been conducted with a focus on

test takers' perspectives in different contexts throughout the world. There exists a research gap in the context of Vietnam, with few studies on stakeholders' perceptions of CBSA, especially examiners' views as the ones are directly involved in the process.

3. Research Methodology

3.1. Research Setting

The study examined the speaking component of the Vietnamese Standardized Test of English Proficiency (VSTEP), the official national English language proficiency test of Vietnam, administered at a test site in Hanoi, Vietnam. VSTEP was first launched in 2015 by the University of Languages and International Studies, Vietnam National University as a paper-based test with the face-to-face speaking component. Since the implementation of Decision 24/2021 issued by the Ministry of Education and Training, the VSTEP speaking test has been delivered via computer, consistent with the other skill components. Test takers are required to attend the test site and complete the speaking section using the computer-based testing system, where their responses are digitally recorded. Test takers basically contact the computer, read the question on the screen, and record their responses via the testing system. The whole speaking test lasts for 12 minutes, including preparation time.

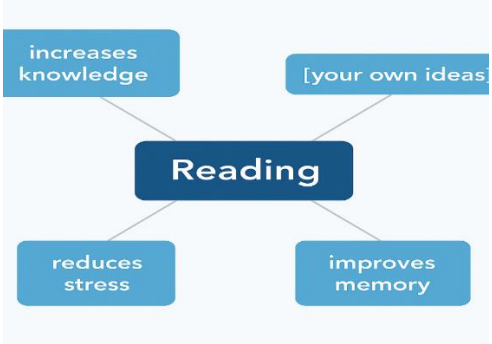
The speaking test consists of three parts - Social Interaction, Solution Discussion, and Topic Development - as outlined in Table 1. Each part includes audio instructions to guide candidates, while the corresponding questions are displayed on the screen. Test takers manage their response time according to the exam timer provided. Prior to recording, they are given the opportunity to check their microphone and ensure recording quality. Upon completion, candidates may review their recordings to identify any technical issues; if problems are detected, they are permitted to re-record their performance.

All recordings are subsequently evaluated in two rounds by certified raters, many of whom previously conducted the test in its face-to-face interview format.

Table 1

The Sample VSTEP Speaking Test

Part 1: Social Interaction (3')
Let's talk about your free time activities.
- What do you often do in your free time?
- Do you watch TV? If no, why not? If yes, which TV channel do you like best? Why?
- Do you read books? If no, why not? If yes, what kinds of books do you like best? Why?
Let's talk about your neighborhood.
- Can you tell me something about your neighborhood?
- What do you like most about it?
- Do you plan to live there for a long time? Why/why not?
Part 2: Solution Discussion (4')
Situation: A group of people is planning a trip from Danang to Hanoi. Three means of transport are suggested: by train, by plane, and by coach. Which means of transport do you think is the best choice?
Part 3: Topic Development (5')
Topic: Reading habit should be encouraged among teenagers.



Follow up questions

- What is the difference between the kinds of books read by your parents' generation and those read by your generation?
- Do you think that governments should support free books for all people?
- In what way can parents help children develop their interest in reading?

3.2. Research Method

Choosing a qualitative approach is essential for this study because it honors the individual voices of examiners in computer-assisted speaking tests. While quantitative data can tell us how many examiners feel a certain way, only a qualitative lens can capture the "human story" behind the screen - the specific frustrations, successes, and moments of professional adaptation that define their experience. By focusing on "common themes" and "divergent perceptions," this method allows for a rich, conversational exploration of how technology affects the examiner's role. A semi-structured interview or focus group discussion can be proper tools to collect insightful information. They provide the space for participants to describe, reflect and evaluate the rating process that they have got involved. The present study employed Focus Group Discussions (FGDs) to collect rich, in-depth data on participants' perspectives. Focus group discussions are widely recognized in social sciences as an effective data collection technique for exploring complex phenomena through participant interactions (Morgan, 1997; Krueger & Casey, 2000). In language assessment research, FGDs have been used to explore stakeholders' perceptions and experiences with language tests, facilitating insights into test fairness, usability, and validity (Zhu & Flaitz, 2005). They provide a valuable forum for participants to discuss computer-based speaking tests, shedding light on cognitive and affective factors involved (Nirma et al., 2024). Purposive sampling and skilled moderation enhance the reliability and depth of data obtained (Guest et al., 2017; Kitzinger, 1995). These methodological principles guided the design and implementation of the FGDs in this study, aligning with best practices in social science and language assessment research.

3.3. Research Participants

Six participants were selected for this study based on their extensive background as speaking examiners. Each participant is a certified VSTEP rater with over ten years of experience, having transitioned from traditional face-to-face testing to the current computer-based format. In recent years, due to the changes in the system, they have been working with the computer-based English speaking test system as raters. Their role involves a standard double-marking process, where they evaluate test-taker recordings across two independent rounds to ensure scoring consistency. For this pilot study, the participants were organized into two focus groups of three. This smaller group size was chosen to encourage an open,

comfortable dialogue while still capturing a variety of professional perspectives. As a preliminary inquiry, this sample size provides a focused starting point for exploring these examiner experiences in more depth.

3.4. Data Collection Procedure

Each focus group session lasted approximately one hour. Participants were invited to share their beliefs and experiences related to the targeted computer-based speaking test. A trained facilitator guided the discussions using a semi-structured interview guide, which included open-ended questions to encourage in-depth sharing without restricting participant expression. The interview questions were designed to elicit detailed narratives from the participants, starting with their long-term history as VSTEP examiners. We encouraged them to reflect on their interactions with the digital system, specifically focusing on the user interface, the clarity of the audio recordings, and the technical flow of the marking process. Furthermore, participants shared their professional insights on how test-takers perform in this digital environment compared to the traditional face-to-face format, highlighting any shifts they observed in candidate behavior or language production.

Sessions were audio-recorded with participant consent to ensure accurate data capture. The facilitator fostered an inclusive and respectful environment, which encouraged equal participation and managing group dynamics to prevent dominance by any individual. Field notes were also taken to supplement the audio recordings by capturing non-verbal cues and contextual information.

3.5. Data Analysis

The focus group recordings were transcribed verbatim and anonymized to ensure participant confidentiality, with each rater assigned a code from T1 to T6. The data were then analyzed using thematic analysis, a process involving the systematic identification and categorization of patterns within the participants' professional experiences. This iterative coding process allowed for a more nuanced exploration of how examiners perceive the shift to computer-based speaking assessments. By moving beyond surface-level observations, the analysis sought to capture the underlying beliefs and practical challenges that characterize the examiners' transition to the digital format.

4. Results and Discussion

4.1. Research Question 1: Common Themes in Examiners' Perceptions

4.1.1. Increased Objectivity and Reliability

Examiners commonly recognized that computer-assisted speaking tests enhance objectivity in scoring. Participants noted that automated or recorded assessment reduces subjective bias inherent in face-to-face scoring.

> "*In my experience listening to VSTEP computer-based speaking tests, I do not observe personal narratives or subjective impressions; candidates strictly respond to specific task questions."* (T1)

One teacher observed that unlike traditional exams where examiners' attitudes or fatigue may affect scores, computer-based tests rely on audio recordings that can be replayed and re-evaluated, promoting fairness and consistency.

> "*One benefit of the computer-based speaking test is that I can replay the recording anytime I want. This provides me with more details to give the scores, while it's impossible to do it with face-to-face interview."* (T4)

When it comes to increased objectivity, this perception generally reflects a widespread trust in technology's capacity to improve reliability in language assessment. This finding is similar to what Pizarro & Laborda (2017) and Koizumi et al. (2017) presented in their study.

4.1.2. Flexibility and Efficiency in the Assessment Process

Flexibility was a recurrent theme. Examiners appreciated the computer-assisted model as it enables asynchronous scoring, allowing raters to review their scoring themselves throughout their evaluation, or when scoring gaps between two rounds happen and require two raters to discuss for an agreement.

> "*Previously scoring required two separate rounds with limited breaks, but now two rounds can be completed in one session, allowing immediate resolution of scoring disagreements..."* (T2)

They commented that this flexibility makes the assessment process more efficient and manageable, especially when compared to intensive real-time face-to-face grading. In the interview mode of VSTEP, the rating used to follow two separate rounds, one is direct interview, and the another is recording marking. The interval between two rounds was normally about one week, so when there existed disagreements, it took time for both raters to listen again and negotiate. In the computer-based speaking test, at the same time, one recording is marked by two raters in two different rooms, which not only saves time but also reflects the test reliability. Therefore, the elimination of multiple scoring rounds into a streamlined single session was highly valued for saving time and resources. This finding is confirmed by all research participants in their discussion.

4.1.3. Test-Taker Accessibility and Reduced Anxiety

Another widely acknowledged benefit was related to learners' experience. Examiners believed that students might feel less intimidated speaking to a computer than a live examiner.

> "*Students tend to feel less anxious speaking to a computer than facing a live examiner, which allows more confident responses."* (T5)

> "*When test takers talk to the computer screen, it seems that they feel free to talk. The anxiety is reduced when they do not have to perform in front of the teacher as usual, and have chances to read the questions on the screen instead of listening to the examiner."* (T6)

This reduced social pressure, combined with the ability to see questions clearly via the test interface, was perceived as creating a more supportive testing environment, especially for lower-level or less confident English speakers. The system's user-friendly interface and built-in audio instructions were viewed as student-friendly features that enhance transparency and reduce anxiety. When raters can not see the test taker's face, it is also the matter of test fairness which will be maintained. This reflects the findings in previous studies such as Kenyon and Malabonga (2001), Dreher et al. (2011), and Baleni (2015).

4.2. Research Question 2: Divergent Aspects in Examiners' Perceptions

4.2.1. Concern Over Validity and Authenticity of Assessment

Despite recognizing objectivity benefits, some examiners expressed reservations about test validity. They questioned whether computer-assisted formats could adequately capture the interactive and spontaneous nature of oral communication.

Some noted that the lack of live interaction limits evaluation of pragmatic competence, turn-taking, and non-verbal cues. This divergence points to an ongoing tension between the controlled, standardized format of computerized tests and the complexities of real-world speaking.

> *"When the test takers can see the questions on the screen and provide their responses, it does not reflect the real communication. We normally have two-way interaction in a speaking test with a human examiner. With the computer-based speaking test like this, it happens one way, it is the test takers who have to work on their own to show their understanding of the questions and the ability to give responses. I wonder if this kind of test can assess the interactional competence of test takers."* (T3)

> *"Even though the scoring rubric does not have criteria assessing interactive competence, it should be taken into consideration with the current test mode. We cannot observe test takers' body language and facial expressions. These features are not assessed in the scoring rubric but they are parts of real-life communication. I also wonder if the test can assess the spontaneous nature of communication when students have time to read the questions instead of listening to the examiner in Part 1 and the follow-up of Part 3."* (T4)

On the other hand, some participants considered that this was not a serious issue as in real life many speaking activities occur without two-way interaction.

> *"I don't think it is a big issue since presentations or talks in the real world also happen as monologues first. The test only reflects part of the real-life communication, it does not necessarily need to reflect exactly the same process."* (T6)

However, based on the discussions, more participants showed their concern about the missing of interactive communication in the test. The absence of two-way interaction with an examiner reduces the authenticity of communicative exchange and raises questions about the extent to which interactional competence can be evaluated. Furthermore, key elements of real-life communication, including body language, facial expressions, and spontaneous responses, are not captured within the current scoring rubric. The practice of reading questions from a screen, rather than responding to a live interlocutor, may also diminish the natural flow of conversation. These issues suggest that the test may inadequately represent communicative competence. This is no longer a new issue, which has been reported in a number of studies (Chapelle & Douglas, 2006; Jeong et al., 2011; Zhou, 2015; Nguyen et al., 2024). The concern that existing papers and the current study raise is all about the real nature of communication. When computer-based speaking tests become more and more popular, this issue should be taken into thorough consideration, as this may have further impact on students' learning behaviors. For further discussion, it is also the matter of test impact, when test takers may choose a way to learn speaking skills without real communication with humans as some participants worried.

4.2.2. Impact on Student Responses and Speaking Performance

Playing the role of raters, some examiners also reported issues related to the quality of test takers' responses. What they are concerned reflects typical cases of the system. Firstly, it is the blurred understanding of requirements.

> "*Many test takers did not listen to the audio instruction in each part. Also, they haven't prepared for the test well. That explains why they did not talk about the topic in Part 3 by following the mind map. They only answered the follow-up questions. If only it was a direct interview, I could have noticed them.*" (T1)

> "*Some of them did not know that they just need to choose one topic to answer in Part 1, and they were in a rush to cover all the questions in Part 1. This actually could have been dealt with if they had listened to the audio instructions*." (T6)

These represent two of several cases reported by raters during the marking process. In fact, this is the responsibility of test takers to follow instructions carefully, it is not the matter of the testing system according to the majority of the participants. However, some examiners proposed measures to prevent such problems if system modifications are possible. For instance,

rather than displaying both the mind map and follow-up questions simultaneously in Part 3, the system could first present the mind map for a set period before revealing the follow-up questions. Clearer guidance should also emphasize the necessity of listening to the audio instructions, ensuring that candidates attend to the system's directions.

Failure to meet these requirements has led to reduced scores for many test takers - an outcome which participants believed could have been avoided with clearer instructions. Consequently, questions have been raised about whether the computer-based testing system can accurately assess speaking performance. Some participants argued that scores did not reflect test takers' actual ability, as misunderstandings of the requirements - or insufficient clarity in the instructions - prevented them from demonstrating their true competence.

4.2.3. Adaptations in Scoring and Assessment Criteria

There was debate about whether traditional scoring rubrics need adjustment for computer-assisted formats. Even though there were only six participants, they had different ideas.

> "*Actually, I do not think it is necessary to adjust the rating scales. The current rubric still works well because it targets key components of speaking. Even though the interactional skills may be affected in the computer-based mode, it is also not assessed in the current scales*." (T5)

> "*I think that when we apply a new testing mode, we should reconsider the scoring rubric. As we discussed earlier, we were worried about the missing of interactional competence in computer-based tests, there should be a criterion to assess this aspect. It may refer to the test takers' ability to interact adequately with the questions, the system or maybe a virtual examiner in the future*." (T2)

It can be observed that while the core scoring criteria for speaking remain applicable, certain aspects of interactive responsiveness might require reassessment or supplementary measures, given the modified testing environment. This divergence highlights the evolving nature of language testing practices in response to technological integration. It would take more time to find the answer whether the scoring criteria need adjusting as this is related to the test format and the testing system as well.

In summary, the findings highlight a complex interplay between the benefits of technological efficiency and concerns regarding communicative authenticity. Regarding research question 1, examiners identified several "common themes" of the advantages of the digital format. Specifically, the participants emphasized the increased reliability and fairness of test, noting that the ability to replay recordings reduces the subjective bias and fatigue inherent in face-to-face interviews. The system was also praised for its flexibility and efficiency, as the streamlined, asynchronous marking process saves significant institutional resources. Furthermore, examiners perceived the computer-based interface as a tool for reducing test-taker anxiety, providing a more supportive environment for less confident speakers by removing direct social pressure. The findings show similarities with previous research such as Malone (2014), Pizarro and Laborda (2017), Nguyen et al. (2024). In contrast, research question 2 revealed divergent perceptions concerning the test's fundamental nature. While many participants valued the standardization, others raised significant concerns regarding validity and authenticity, arguing that the lack of two-way interaction fails to capture interactional competence. This divergence extended to the impact on student performance, where some raters attributed poor scores to system-led misunderstandings of instructions, while others viewed this as a matter of candidate preparation. This is what Elder and Iwashita (2005), Quaid and Barett (2021), Nguyen et al. (2024) pointed out. Finally, the study found conflicting views on scoring criteria, with debate over whether current rubrics should be adapted to reflect the unique demands of computer-mediated communication. Such findings need more investigation in the future research.

5. Conclusion

This study investigated examiners' perceptions of a specific computer-based speaking test assessing English language level of Vietnamese adult learners. Findings from the study have shown the contribution of this study in the field. On the one hand, the study's results reaffirm what existing research pointed out. By using high-quality recordings and a more flexible, asynchronous marking process, the system helps eliminate common human errors like examiner fatigue or subjective bias that often occur in face-to-face interviews. On the other hand, concerns are raised in terms of the test validity and authenticity with a lack of interactional competence. This tension highlights the challenge of balancing reliability with tasks that reflect real-world communication. Technology undoubtedly enhances efficiency, but test design must continue to safeguard the human essence of language use.

Implications extend to high-stakes testing in Vietnam and beyond. Computer-based speaking tests such as VSTEP are increasingly adopted due to their convenience and scalability. Yet their success depends on addressing authenticity gaps to ensure construct validity and positive washback. Practical solutions include clearer instructions, sequential prompts, and AI-supported virtual interlocutors to restore interactional features and authentic communication.

The study however has some limitations with a small sample of six experienced raters in one context. Also, self-reported perceptions may overlook behavioral discrepancies. Therefore, future research could employ larger, diverse cohorts, triangulate with quantitative reliability metrics (e.g., inter-rater agreement analyses), or explore longitudinal washback on learners. The researcher would consider conducting experimental comparisons with face-to-face modes, alongside AI-enhanced interactivity to further validate these perceptions. Ultimately, examiners' voices are expected to guide us toward tests that not only measure but authentically cultivate communicative competence in the era of computerized assessment.

References

Alimorad, Z., & Saleki, A. (2022). Challenges and opportunities of web-based assessment in EFL courses as perceived by different stakeholders. *Applied Research on English Language*, *11*(2), 125-154. https://doi.org/10.22108/are.2022.132413.1843

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

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511733116

Cushing, S. T., Ren, H., & Tan, Y. (2024). *The use of TOEFL iBT® in admissions decisions: Stakeholder perceptions of policies and practices* (ETS Research Report Series). Educational Testing Service. https://files.eric.ed.gov/fulltext/EJ1459649.pdf

Dermo, J. (2009). e-Assessment and the student learning experience: A survey of student perceptions of e-assessment. *British Journal of Educational Technology*, *40*(2), 203-214. https://doi.org/10.1111/j.1467-8535.2008.00915.x

Douglas, D., & Hegelheimer, V. (2007). Assessing language using computer technology. *Annual Review of Applied Linguistics*, *27*, 115-132. https://doi.org/10.1017/S0267190508070062

Dreher, C., Reiners, T., & Dreher, H. (2011). Investigating factors affecting the uptake of automated assessment technology. *Journal of Information Technology Education: Research*, *10*, 161-181. https://doi.org/10.28945/1492

Elder, C., & Iwashita, N. (2005). Planning in language testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 217-238). John Benjamins. https://doi.org/10.1075/lllt.11?locatt=mode:legacy

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

European Union. https://www.researchgate.net/publication/281090096_FACE-TO-FACE_AND_COMPUTER-BASED_ASSESSMENT_OF_SPEAKING_CHALLENGES_AND_OPPORTUNITIES

Guest, G., Namey, E., & McKenna, K. (2017). How many focus groups are enough? Building an evidence base for nonprobability sample sizes. *Field Methods*, *29*(1), 3-22. https://doi.org/10.1177/1525822X16639015

Hayashi, Y., & Kondo, Y. (2024). Automated speech scoring of dialogue response by Japanese learners of English as a foreign language. *Innovation in Language Learning and Teaching*, *18*(1), 32-46. https://doi.org/10.1080/17501229.2023.2217181

Hu, H., Gong, Q., & Mohd-Said, N. (2025). Exploring a decade of research: A systematic review of computer-based English speaking tests. *Forum of Linguistic Studies*, *7*(4), 788-803. https://doi.org/10.30564/fls.v7i4.8978

Isbell, D. R., Crowther, D., & Nishizawa, H. (2024). Speaking performances, stakeholder perceptions, and test scores: Extrapolating from the Duolingo English Test to the university. *Language Testing*, *41*(2), 233-257. https://doi.org/10.1177/02655322231165984

Jeong, H., Hashizume, H., Sugiura, M., Sassa, Y., Yokoyama, S., & Shiozaki, S. (2011). Testing second language oral proficiency in direct and semidirect settings: A social-cognitive neuroscience perspective. *Language Learning*, *61*(3), 675-699. https://doi.org/10.1111/j.1467-9922.2011.00635.x

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology*, *5*(2), 60-83. https://www.learntechlib.org/p/90922/

Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *British Medical Journal*, *311*(7000), 299–302. https://doi.org/10.1136/bmj.311.7000.299

Krueger, R. A., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Sage Publications.

Kunnan, A. J. (2004). Test fairness. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Proceedings of the ALTE Barcelona Conference* (pp. 27-48). Cambridge University Press. https://assets.cambridge.org/97805215/35878/frontmatter/9780521535878_frontmatter.pdf

Liu, J., & Zhang, B. (2020). Multi-level Rasch model analysis of computer-assisted automated scoring of English listening and speaking tests. In *Proceedings of 2020 International Conference on Computer Engineering and Application* (pp. 632–636). https://doi.org/10.1109/ICCEA50009.2020.00138

Morgan, D. L. (1997). *Focus groups as qualitative research*. Sage Publications. https://doi.org/10.4135/9781412984287

Nguyen, T. H. H., Nguyen, B. T. T., Hoang, G. T. L., Pham, N. T. H., & Dang, T. T. T. (2024). Computer-delivered vs. face-to-face score comparability and test takers' perceptions: The case of the two English speaking proficiency tests for Vietnamese EFL learners. *Language Testing in Asia*, *14,* 6. https://doi.org/10.1186/s40468-024-00277-1

Nirma, Sujariati, & Ariana. (2024). Focus group discussion as communicative activity in enhancing students' speaking skill at SMAN 5 Barru. *English Language Teaching Methodology*, *4*(3), 335-345. https://doi.org/10.56983/eltm.v4i3.473

Pizarro, M. A., & Laborda, J. G. (2017). Analyzing test-takers' views on a computer-based speaking test. *Profile: Issues in Teachers' Professional Development*, *19*(Suppl. 1), 23-38. http://dx.doi.org/10.15446/profile.v19n_sup1.68447

Quaid, E., & Barrett, A. (2021). Interpretations of spoken utterance fluency in simulated and face-to-face oral proficiency interviews. *Language Education & Assessment*, *4*(1), 1-18. https://doi.org/10.29140/lea.v4n1.385

Sachdeva, S., Tamrakar, K. A., Perwez, E., Kapoor, P., & Gupta, D. (2024). Focus group discussion: An emerging qualitative tool for educational research. *International Journal of Research and Review*, *11*(9), 302-308. https://doi.org/10.52403/ijrr.20240932

Yulianto, D., & Mujtahin, N. M. (2021). Online assessment during Covid-19 pandemic: EFL examiners' perspectives and their practices. *Journal of English Teaching*, *7*(2), 229-242. https://doi.org/10.33541/jet.v7i2.2770

Zhou, Y. (2015). Computer-delivered or face-to-face: Effects of delivery mode on the testing of second language speaking. *Language Testing in Asia*, *5*(2), 1-16. https://doi.org/10.1186/s40468-014-0012-y

Zhu, W., & Flaitz, J. (2005). Using focus group methodology to understand international students' academic language needs: a comparison of perspectives. *The Electronic Journal of English as a Second Language, 8*(4). https://files.eric.ed.gov/fulltext/EJ1068108.pdf