Language Testing in Asia

RESEARCH Open Access

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

Thuy Ho Hoang Nguyen[1], Bao Trang Thi Nguyen[1*], Giang Thi Linh Hoang[1], Nhung Thi Hong Pham[1] and Tu Thi Cam Dang[1]

*Correspondence:
ntbtrang@hueuni.edu.vn

[1] Faculty of English, University of Foreign Languages and International Studies, Hue University, 57 Nguyen Khoa Chiem, Hue, Vietnam

Abstract

The present study explored the comparability in performance scores between the computer-delivered and face-to-face modes for the two speaking tests in the Vietnamese Standardized Test of English Proficiency (VSTEP) (the VSTEP.2 and VSTEP.3–5 Speaking tests) according to Vietnam's Six-Level Foreign Language Proficiency Framework (VNFLPF) and test takers' experiences. Data were collected from 75 and 82 VSTEP.2 and VSTEP.3–5 university English-majored test takers respectively in both computer-delivered and face-to-face conditions. A counterbalanced research design was adopted to minimise mode order effects. After test completion, 30 of the test takers, 15 from each proficiency test, were interviewed in the focus group format of 3–4 members per group. The results indicated mixed, selective effects of the testing mode. Overall, test scores were comparable in the VSTEP.2 Speaking test but significantly higher in favour of the face-to-face mode for the VSTEP.3–5 Speaking test. However, the statistically significant difference was observed in only one measure of the many analytical criteria (content development in the former test, and pronunciation in the latter test) with mixed mode advantages. The interview data has provided rich refreshing insights into how test takers viewed each testing mode against real-life communication. Their experiences further revealed a wide range of affective preferences involved in the inherent affordances or constraints of each testing mode and their communication and performance/outcome orientation. The findings offer important implications for extrapolation, test preparation and administration, and test taker/rater training in the particular context of the two English speaking proficiency tests in Vietnam and perhaps beyond.

Keywords: Vietnam, Speaking proficiency tests, VSTEP.2, VSTEP.3–5, Face-to-face, Computer-delivered, Performance scores, Test takers' perspectives

Introduction

In today's contemporary world, technological advances have transformed almost all aspects of language education and computer-delivered (semi-direct) language assessment has become a growing trend. For oral assessment, semi-direct testing has increasingly been adopted in many educational contexts alongside its traditional direct method. The former usually involves test takers talking to a computer and their task performance is recorded for subsequent grading while the latter means conducting a speaking test in the face-to-face (live) mode with the presence of a human interlocutor (Kyle et al., 2016; Qian, 2009). In the live testing environment, test takers have an opportunity to showcase their ability to interact (Mirzaei et al., 2016) and it has generally been acknowledged that direct testing has more face validity than semi-direct testing (Brahim, 2023; Kyle et al., 2016; van Lier, 1989), though test takers' performance might depend to a large extent on how the human raters administer the test (Chang et al., 2018; Kiddle & Kormos, 2011). Semi-direct testing, on the other hand, follows a standardised procedure as it provides similar forms of test input to all test takers (Leaper & Riazi, 2014). It is also considered fairer because the subjectivity to the skills of the present raters is no longer an issue (Stansfield, 1991) and more cost-effective since it could be administered on a large scale, thus saving human resources (Garcia-Laborda, 2007). The computer-based test mode might well address the growing demand for language proficiency assessment, particularly among English learners in the world (Marian & Jesus, 2017).

In Vietnam, the National Foreign Language Project was established in 2008 with an initial attempt to reassess and then improve Vietnamese learners' language proficiency. Central to the goals of the Project is the adoption of the Common European Framework of Reference for Languages (CEFR) for setting English teachers' professionalism standards, developing standard-based learning outcomes, and renewing English language curriculum. With the CEFR being adopted as a responsive move to the national need for a localised proficiency framework, the Vietnam's Six-Level Foreign Language Proficiency Framework (VNFLPF) was officially launched in 2014 by the Ministry of Education and Training of Vietnam (MOET) (MOET, 2014). The framework maps language competency onto three broad levels of Beginner, Intermediate, and Advanced, which are further sub-divided into six different levels from level 1 to 6, targeting the six corresponding levels of the CEFR from A1 to C2 (MOET, 2014).

Drawing on the VNFLPF, the Vietnamese Standardized Test of English Proficiency (VSTEP) was initially introduced in Vietnam in 2015, with five test levels, among which the level 3–5 English proficiency test and the level 2 English proficiency test (henceforth VSTEP.2 and VSTEP.3–5 respectively) have gained most popularity (MOET, 2015, 2016). These two tests aim to assess Vietnamese EFL learners' English proficiency in all the four skills (listening, reading, writing, and speaking) and offer a range of test options for Vietnamese EFL learners to select for their own needs. The VSTEP.3–5 test targets a wider range of test takers than the VSTEP.2 test. For example, the former is for undergraduate students majoring in English language teaching (ELT) or English linguistics who are required to obtain level 5 (C1-CEFR equivalent) upon graduation, as well as students in other disciplines who should achieve level 3 (B1-CEFR equivalent) to complete their BA studies. Postgraduate students aiming to obtain a level 4 (B2-CEFR equivalent) certificate, teachers of different disciplines, civil servants or those in search of jobs can also

opt for the VSTEP.3–5 test. This test construct ranges in difficulty from levels 3 to 5, and whatever level test takers' proficiency reaches will they be recognised at that level. Specifically, on a 10-point scale, if test takers score from 4 to 5.5 points, their English proficiency will be recognised at level 3; if they score from 6 to 8, they are at level 4; and those whose scores range from 8.5 to 10 are level 5 English users. On the other hand, other test takers might choose to sit the VSTEP.2 test for their job or studies that require a level 2 proficiency (A2-CEFR equivalent). The cut score of the VSTEP.2 test is 6.5 on a 10-point scale. These proficiency tests were originally designed to be administered in the paper-based format. However, in 2017, MOET promoted the administration of these tests in both direct and semi-direct modes, thus the delivery of the respective speaking tests following suit in the two said modes, with the same test structures and formats.

Against such a backdrop of test mode shifting, computer-delivered testing has increasingly garnered considerable attention in Vietnam. However, little empirical knowledge has been acquired about the compatibility of direct and semi-direct test delivery in particular regard to the two target speaking proficiency tests. As they are high-stakes tests in Vietnam, there is a clear need to investigate how they are compatible in terms of the performance scores and test takers' perceptions. The recent stipulation from MOET (2021) that computer-based testing be employed in EFL assessment in order to save time and human resources further necessitates the collection of validity evidence to inform their appropriate adoption. Furthermore, given the inconclusive findings on the equivalence of the direct and semi-direct testing modes in prior research to date (e.g. Bijani, 2019; Jeong & Hashizume, 2011; Nakatsuhara et al., 2021; Quaid & Barrett, 2021; Zhou, 2015), further research is warranted.

The present study thus aimed to explore the comparability in test scores between the face-to-face and computer-based modes of delivery for the VSTEP.2 and VSTEP.3–5 speaking proficiency tests according to the VNFLPF, and test takers' experiences. The significance of the study is threefolded. Firstly, empirical evidence in terms of scores and test takers' perceptions will inform test administration and extrapolation inference for appropriate conclusions to be drawn from the obtained results. Secondly, the study will further shed lights on how to orchestrate rater and test taker training as well as preparation for a supportive testing environment. Thirdly, it is hoped to extend existing scholarship on oral testing modes by responding to the pressing call for contextualised research (Frost & McNamara, 2018) that addresses the two speaking proficiency tests in Vietnam.

Literature review

Growing research has been devoted to comparing computer-based testing with the traditional face-to-face mode with the presence of a human interlocutor. Earlier research tended to employ correlation-based approaches to seek empirical evidence for the comparability of direct and semi-direct testing (see Stansfield & Kenyon, 1992 for a review). This line of research examined the extent to which two different raters agreed with each other when scoring the same test performances in both conditions or whether the same test takers received similar scores in both test formats. Stansfield and Kenyon (1992) reported high inter-rater reliability (high correlation coefficients) in both test formats. In other words, the reliability of both modes of testing was highly comparable. However, utilising one single measure of correlation coefficients is not empirically sufficient

to claim the equivalence between the two testing modes (Shohamy, 1994). Accumulating research has begun to investigate the comparability of computer-based and face-to-face testing via comparing performance scores and/or qualitatively analysing language use, and/or probing test takers' perceptions.

Comparability in terms of test scores and/or language use

The first point of comparison is test scores and/or language use and the results are quite mixed. To begin with, some studies have revealed no significant difference in test scores of the two testing modes (e.g. Nakatsuhara et al., 2021; Quaid & Barrett, 2021; Zhou, 2015). For example, Zhou (2015) examined test performances by 61 undergraduate students and 18 EFL high school students in Japan in two speaking tasks: (i) tell stories and (ii) express personal opinions, with the same content in the computer-based and face-to-face formats. No significant difference was found between the two testing conditions in terms of test scores in four analytical criteria: pronunciation, lexical resource, grammar and fluency. In a similar Japanese EFL context, Yonezaki (2016) asked 43 first-year students to perform a storytelling task and an opinion-giving task in the direct and semi-direct modes and test takers' performances were rated in four aspects, namely volume, content, fluency and accuracy. Again, no statistically significant differences were found between the two test modes in any criterion, though direct testing was perceived more favourably. Targeting four Chinese EFL learners, and their fluency in an Aptis General speaking test delivered in the direct mode and on a computer, Quaid and Barrett (2021) observed that the testing mode did not affect the overall fluency as measured by speaking speed, pauses and repair, though many unfilled pauses were recorded in the latter format. Khabbazbashi (2017) employed the multi-faceted Rasch program (FACETS) to analyse test scores by 83 test takers in a speaking test delivered in the live and semi-direct modes and found high comparability between the two test formats. All these findings here suggest that the two modes of testing could be reliable substitutes.

However, other prior research has documented an impact of testing modes, though this effect was not always consistent with either mode of testing. Stansfield and Kenyon (1992) mentioned earlier reported the results of Shohamy et al.'s (1991) study that the number of grammatical errors test takers committed did not differ statistically in the two testing modes, though the semi-direct test elicited more formal language use than the traditional face-to-face test and examinees tended to repeat test questions in the latter mode. In a subsequent study, Shohamy (1994), via an analysis of 20 speaking test samples, observed that the language characteristics featured in the two testing modes were not statistically significant, though self-correction and paraphrasing were employed more frequently in the computer-based testing mode. This indicates that test takers were more aware of the need to use correct language when talking to a computer. Bijani (2019) found similar results in a study in an Irian EFL context which elicited test performances via five speaking tasks (description, narration, summarising, role-play and exposition). Despite the highly compatible test scores in the two test methods, test takers were also more oriented towards accuracy in the semi-direct format by displaying more self-correction and paraphrasing. These findings might suggest that the two test modes do not measure the same thing.

Mixed effects of test formats were also found in other studies. For example, Choi (2014) analysed the oral English performance samples by 15 Korean EFL test takers at an advanced level. They performed three monologic, non-interactive tasks, namely description, narration and opinion giving in both direct and semi-direct modes, with different task topics in each testing condition. This study uncovered that the overall performances under the direct and semi-direct conditions were not statistically different and direct and semi-direct performances were strongly correlated. Despite this overall high comparability, test takers committed more errors related to verb structures and displayed a higher level of lexical density in the computer-based performances, with some moderate effects of task types.

Several other studies have found an impact of testing modes on selective aspects of test performances or language use. Nakatsuhara et al. (2021) studied the comparability of IELTS speaking test scores and language functions used by test takers between the semi-direct and face-to-face speaking modes. In their study, 99 examinees sat for an IELTS speaking test in both conditions. The results revealed that test takers achieved similar test scores and used almost the same language functions in both modes, except for *asking for clarification*. In research by Kiddle and Kormos (2011), 42 students (year 1 to year 4), aged 18–37 at a university in Chile, performed three monologic tasks: (i) introduce themselves and talk about their future plans, (ii) describe a thief based on a video-recorded street robbery and (iii) record a voice message to present a problem and ask for a solution. The results of many-facet Rasch analysis showed that test takers scored significantly higher on pronunciation in the face-to-face test than in the computer-based test.

Overall, prior research has generally indicated high comparability in terms of total test scores between the two test formats, but the mode effects differed on different criterion scores and selective aspects of language use. This latter inconsistent impact of test modes could be attributed to the fact that past research on the compatibility of the two test formats has been conducted in a wide range of contexts with different pools of test takers. Furthermore, notably different task types/topics and rating scales have been adopted in different settings. All this diversity could make it challenging to compare results across studies while at the same time, interestingly, indicates that testing is typically context-specific and context-driven. In this regard, there is an obvious need to investigate how the two test formats are in use in particular educational settings and especially how comparable they are from multiple perspectives that go beyond performance scores to include test takers' perceptions and preferences. Research on test takers' perceptions is reviewed next.

Comparability in terms of test takers' perceptions

The perceptions of test takers towards the two modes of testing are an additional point of comparison, with inconsistent results across studies. For example, surveying 300 Iranian adult test takers about their perceptions of performing five given speaking tasks (description, narration, summarisation, role-play and exposition) in the computer-based and direct modes, Bijani and Khabiri (2017) found that the two testing modes were rated the same in terms of time pressure and difficulty. Preferences were reported for the direct format as it was perceived to better reflect test takers' speaking

ability and cause less anxiety for them. Unfamiliarity with recording techniques was also cited as one major disadvantage leading to unsatisfactory responses. In a similar vein, in Bijani's (2019) study, test takers found it more challenging and stressful to perform in the semi-direct mode. Preferences for the direct mode were also reported in other studies (Jeong & Hashizume, 2011; Khabbazbashi, 2017), though the two modes were comparable in terms of performance scores. Obviously, the testing mode could affect the test takers cognitively as well as emotionally.

This affective influence is, however, less clear in Qian's (2009) study, which found that a majority of the test takers (57.6%) had a neutral opinion towards the two testing modes, and a sizable portion of them (40.9%) were willing to accept both testing formats and many (32.8%) reported a preference for the direct test. Kiddle and Kormos (2011) mentioned earlier also found that test takers had positive perceptions of both testing modes, though a majority of them considered direct testing a fairer way of assessing oral performances.

By asking test takers to perform two description tasks in the two testing modes and complete a survey about their preferences, Baralt and Gurzynski-Weiss (2011) found that their Spanish intermediate test takers reported similar levels of anxiety, and greater familiarity with either testing mode reduced apprehension. On the contrary, Chang et al. (2018) reported higher levels of anxiety in the face-to-face mode. This finding also corroborates those in Quaid and Barrett's (2021) study that being unable to read non-verbal feedback from the examiner in the computer-based test was a reported cause of anxiety, which resulted in lower fluency while face-to-face testing was perceived to motivate talk and enhance fluency. Generally, the mixed preferences and experiences found in research on test takers' perceptions were related to how they perceived the constraints and affordances associated with each testing mode.

In brief, mixed results have been reported in prior research on the equivalence of the face-to-face and computer-based testing formats both in terms of test scores/language use and test takers' perceptions. This warrants further exploration to inform their application. Furthermore, a wide range of learner groups and popular standardised tests such as IELTS or TOEFL iBT or researchers' tasks have quite often been targeted in prior studies. Scant research has compared the direct and semi-direct testing modes of the language proficiency tests that are mandated in a particular educational context. Given the importance of the VSTEP.2 and VSTEP.3–5 tests for different groups of EFL learners in Vietnam, examining test score compatibility and test takers' experiences in the computer-delivered and face-to-face testing modes of the target tests will provide valuable feedback for test administration and extrapolation.

The present study addressed the following research questions (RQs):

> RQ1. Are there any differences in test scores of the VSTEP.2 Speaking Test for Vietnamese EFL learners in the computer-delivered and face-to-face testing modes?
>
> RQ2. Are there any differences in test scores of the VSTEP.3–5 Speaking Test for Vietnamese EFL learners in the computer-delivered and face-to-face testing modes?
>
> RQ3. How do Vietnamese EFL test takers perceive sitting for each of the two proficiency tests in the computer-delivered and face-to-face testing modes?

Methods and materials

The participants

The test takers and teacher raters were recruited on a voluntary basis. They were informed of the research and all had given their consent before data collection began.

The VSTEP.2 test takers

The VSTEP.2 test takers were 75 first year English major university students at a university of foreign languages in central Vietnam. A majority of them (59/75) were female and their ages ranged from 19 to 21. By the time they sat the VSTEP.2 Speaking test in the present study, they had completed a speaking course that targeted lower A2 in the first semester of their bachelor program. None of them had experienced the semi-direct speaking mode before. Although they were in the same year level, their English proficiency varied from A1 to A2 according to the Common European Framework of Reference for Languages (CEFR) as informed by their class teachers. They had learnt English for about 8 years since their secondary school.

The VSTEP.3–5 test takers

Eighty-two English-major students (60 females and 22 males) at the same research site volunteered to take the VSTEP.3–5 Speaking test in both direct and semi-direct test modes of delivery. They were aged from 21 to 23. Fifty-eight of them were third year students whose proficiency levels ranged from B1 to B2 according to CEFR. Twenty-four were in their second semester of their fourth year at the university and they were required to achieve VSTEP Level 5 (C1 equivalent) upon graduation. At the time of test admiration, the third year and fourth year students had completed a two-credit English speaking course aiming at lower B2 and higher B2 levels respectively in their BA program. None had sat any VSTEP.3–5 Speaking test in the semi-direct format, though ten of them had some prior experience with the computer-based mode in testing of other foreign languages such as Japanese, Korean and Chinese, and English examinations at the high school level. They had learnt English for about 9–10 years since grade 6.

The raters

Sixteen Vietnamese EFL lecturers who taught at the research site were recruited to rate direct and computer-based performances. They were officially certified VSTEP raters who received intensive training on VSTEP.2 and VSTEP.3–5 scoring. They had at least 2 years of experience in assessing semi-direct and direct test performances of these two particular proficiency tests. All of them were female and aged from 35 to 45. Eight raters (four pairs) scored the performances of each target proficiency test in both testing modes. They were trained to administer the tests before its actual test day and the training focused on how to interpret, grade the different analytical categories of the marking scheme for each speaking test format and calculate final scores. This short training aimed to remind raters of how to score the VSTEP.2 and VSTEP.3–5 Speaking tests again.

The speaking tests

Two test forms of each English proficiency test (VSTEP.2 and VSTEP.3–5) were employed for the face-to-face and computer-based modes. They were two separate test forms which had the same test structure, though with a different content. Table 1 presents the task structure for the VSTEP.2 and VSTEP.3–5 Speaking test. The speaking tasks were sourced from a prepared bank of items developed by National Testing Center (NTC), Quality Control Department, MOET. The items in each test had been officially ensured for appropriateness of item difficulty through a standardised procedure, and both forms of the same test were validated by NTC through a rigorous process that involves multiple steps as stipulated in Circular 23 (MOET, 2017): (i) piloting each target test in two modes of delivery with at least 50 test takers; (ii) preparing and compiling the piloted test scores before analysis; (iii) arranging VSTEP experts to utilise specialised programs to analyse and validate the test, evaluate test difficulty and

Table 1 The structures of the VSTEP.2 and VSTEP.3–5 Speaking tests

VSTEP.2		**VSTEP.3–5**	
Task	**Timing**	**Task**	**Timing**
1. Greetings Test takers responded to examiners' greetings and 3 set-up questions (e.g. Good morning/afternoon, sit down please, What's your name, How are you?)	0.5 min (no preparation time)	***1. Social interaction*** Test takers were asked 3–6 questions on 2 familiar topics	3 min (no preparation time)
2. Social interaction Test takers were asked about 3 questions and raised about 1–2 questions about daily activities	1.5 min (no preparation time)	***2. Solution discussion*** Test takers were given a situation in which they were to choose one most suitable solution to the given situation among the three provided. They also need to justify their choice	4 min (1 min preparation)
3. Description Test takers were given a card and required to describe a specific person, thing, or experience that is familiar to their daily life After their description, they were asked 1–2 follow-up questions	4 min (1 min preparation)	***3. Topic development*** Test takers were required to develop a given topic (e.g. the benefits of reading books) by using the suggested ideas and/or their own Following their talk, they were asked 1–3 follow-up questions	5 min (1 min preparation)
4. Discussion Test takers were given a controversial topic (e.g. whether children should use smart phones at an earlier age) and required to express their opinion on it After their talk, they were asked 1–2 follow-up questions	4 min (1 min preparation)	n/a	n/a

difficulty equivalence between two test forms (face-to-face and computer-based); and (iv) suggesting items to be revised or eliminated as necessary.

Test administration

The test takers performed their respective VSTEP.2 or VSTEP.3–5 Speaking test, in two modes: direct and semi-direct in a counterbalanced design to avoid the possible impact of mode order. As shown in Table 2, test takers in each proficiency test (VSTEP.2 and VSTEP.3–5) were randomly divided into two groups. Group 1 performed the target test in the face-to-face mode first (face-to-face → computer-based) while the reverse mode order was adopted for group 2 (computer-based → face-to-face). In this way, for the whole sample of each test, there were test takers who did each of the two testing modes in a counterbalanced manner.

In the face-to-face condition, the test was conducted in different rooms with about 18–21 test takers per test room who took turns to do the test. Only one candidate was present in the test room at a time and those who were waiting for their turns were grouped in a separate room on the same floor with the actual test location. They were not allowed to keep their belongings or any supporting materials in the waiting room. During the test, test takers were provided with blank papers to take notes as they wished, but they were not permitted to use any resources. Each test taker performed the target test with two live human raters: an interlocutor and an examiner. Altogether, test takers sat the VSTEP.2 and VSTEP.3–5 speaking test for 10 and 12 min respectively (also see Table 1). Upon completing the intended test, each test taker was directed not to return to the waiting room, but to leave the research site. An external supervisor was present to guide test takers to exit.

The computer-based mode of testing was conducted on the same day with the face-to-face testing. In the former format, the test prompts were inputted into a testing software by the NTC prior to test administration. Each candidate sat at an assigned computer, which displayed textual test prompts, followed the given task instructions and performed the test. An explicit digital countdown timer appeared on the computer screen for the candidate to manage response time. The semi-direct test was conducted in numerous rooms concurrently with about 18–21 participants per room. The setup involved arranging computers 1 m away from each other and installed with closed block walls in order to prevent noise from other test takers when the test was in progress. Test takers wore headphones while taking the test. In each testing room, one EFL teacher worked as test supervisor and one technician provided technical assistance as needed. Test takers did not receive support of any other form during the test, but they were allowed to take notes on a blank paper as they wished and they were to audio-record their task responses via a record button on the computer, and then submit their

Table 2 Test administration design

VSTEP.2 (*n* = 75)			**VSTEP.3–5 (*n* = 82)**		
	1st speaking	***2nd speaking***		***1st speaking***	***2nd speaking***
Group 1 (*n* = 37)	Face-to-face	Computer-based	Group 1 (*n* = 41)	Face-to-face	Computer-based
Group 2 (*n* = 38)	Computer-based	Face-to-face	Group 2 (*n* = 41)	Computer-based	Face-to-face

performances. Prior to the official test administration, they had been trained on how to perform the test via the computer-based mode.

Test scoring

Rating scales and score calculation

Both live and computer-based performances in each target speaking test were graded using the same scoring criteria. The first part of the VSTEP.2 Speaking test was graded in distinct ways from the remaining ones: If test takers sufficiently responded to all the four questions from the examiners (greetings and three set-up questions), they were awarded a full score of 1.0; if they could only respond to 1–3 questions, they received no point. In contrast, the VSTEP.2 performances of parts 2–4 were rated according to six analytical criteria (grammar, pronunciation, vocabulary, fluency, content development and communication strategies). Each criterion was scored on a five-level scale, from 0 (no attempt) to four (mastery). It was found that for the greetings part, every test taker was awarded a maximum score of 1.0 in both testing modes. For this reason, and the fact that the greetings part was rated differently, the focus of the current analysis was exclusively on parts 2–4 of the VSTEP.2 Speaking test to allow for more accurate comparison between the direct and semi-direct modes. The maximum total score for all the criteria was therefore 24 points.

For the VSTEP.3–5 Speaking test, the marking scheme targeted five analytical dimensions (grammar, vocabulary, pronunciation, fluency and content development) and each was scored on an 11-point scale, the lowest being 0 (no attempt) to the highest 10 (mastery). The total maximum score for all the criteria was 50 points. The computer-based and face-to-face test performances of each part per test taker were graded via the same marking scheme mentioned above. The overall score for the entire test was the average score of the three parts calculated for each analytical criterion separately and for all the criteria combined.

As described earlier, the raters were well trained and certified in scoring those tests. They received a printed version of the marking scheme with detailed descriptions of the target criteria in the pre-scoring training session and again on the test administration/scoring day and ready-made paper score sheets to record the scores in each part of the test and the total score for each individual test taker.

Scoring procedures

In the face-to-face mode, two different raters graded the same live performances according to the six analytical criteria for the VSTEP.2 Speaking Test and the five analytical elements for the VSTEP.3–5 Speaking Test as described above. The raters administered the test with two individual candidates first and awarded individual scores for each candidate. Test takers' performances were also audio-recorded for the purpose of subsequent remarking as needed. Any score discrepancies larger than the maximum allowed difference stated in Circular 23 (MOET, 2017) were discussed. After the first two candidates' performances, the raters discussed with each other if any discrepancies occurred. The raters then continued to deliver the live test with the remaining test takers. They had as much time as they needed to compare and discuss the scores after each candidate.

Table 3 Pearson correlation coefficients of rater scoring

VSTEP.2 (*N* = 75)			**VSTEP.3–5 (*N* = 82)**		
Raters	***Computer-based***	***Face-to-face***	***Raters***	***Computer-based***	***Face-to-face***
Pair 1 (*n* = 18)	0.95	0.91	Pair 1 (*n* = 20)	0.92	0.92
Pair 2 (*n* = 19)	0.93	0.90	Pair 2 (*n* = 21)	0.94	0.95
Pair 3 (*n* = 19)	0.90	0.94	Pair 3 (*n* = 20)	0.92	0.92
Pair 4 (*n* = 19)	0.91	0.96	Pair 4 (*n* = 21)	0.95	0.93

N, number of test takers; *n*, number of performances to score in each mode of testing

Table 4 The interviewees (*n* = 30)

	Test takers	**Age**	**Gender**	**English proficiency level**	**Semi-direct testing experience**
VSTEP.2	15 first-year English-majored students	19–21	13 females 2 males	A2-A2 (CEFR)	None
VSTEP.3–5	15 English majors (5 third year & 10 fourth year)	21–23	11 females 4 males	B1-B2 (CEFR)	One

In the computer-based mode, the audio-recorded performances of each speaking test (VSTEP.2 and VSTEP.3–5) were graded 1 day after the test by the same four pairs of trained raters who rated the face-to-face performances. Like the face-to-face mode, the same scoring criteria were applied and so was a similar scoring procedure. The two raters sat at two different computers, listening to the first two audio recorded performances and rated them independently. Any score differences beyond the maximum allowed difference specified in Circular 23 (MOET, 2017) were further discussed for score finalisation. After the first two test takers, scoring continued with the remainder of the assigned recordings in similar manners.

For the two target speaking tests, the scoring results of the rater pairs demonstrated high inter-reliability, with the agreement percentages ranging from 92 to 95% of the scores within one point difference for the three different parts of each test. In addition, Pearson correlation coefficients between the final scores awarded by each pair of raters were from 0.90 to 0.96 for each testing mode ($p < 0.001$), indicating high reliability (Table 3).

Interviews

Immediately after completion of the VSTEP.2 and VSTEP.3–5 Speaking tests, 30 of the test takers (15 per target test) participated in focus group interviews with 3–4 students per group. Most of the interviewees were female, 13 and 11 from the VSTEP.2 and the VSTEP.3–5 tests respectively. Only one of them (VSTEP.3–5 test taker) had previously experienced the computer-based format in which she completed a different English test in her high school years (further see Table 4). The interviewees were selected on a voluntary basis as they expressed willingness to attend a post-test interview together with their given consent to participate in the present research. The interviews aimed to understand test takers' experiences of the two testing modes, centering around their preferences and

the reasons underpinning their preferred mode of testing. In other words, they were guided by two overarching open-ended questions "How did you experience the two testing modes? Which do you prefer and why?" The interviews were conducted in Vietnamese to maximise comfort and understanding for easy expression of their experiences performing the target speaking tests in the different modes. Each group interview lasted appropriately 1 h and were audio recorded with their prior permission. The interviewees were not informed of their test scores in any mode at the interview time.

Data analysis

Test scores awarded by the raters in the two modes of delivery (face-to-face and computer-based) for each type of test (VSTEP.2 and VSTEP.3–5) were inputted into SPSS (version 26.0) and checked carefully for accuracy by the authors before analysis. The data were checked for normality of distribution via the Kolmogorov–Smirnov test as recommended by Field (2017) and the results showed non-normality ($p < 0.05$, two-tailed). In order to compare the speaking performances between the two delivery formats, a number of non-parametric Wilcoxon signed rank tests were run as the data were not normally distributed. The significance level of 0.05 was selected as a conventional value for all analyses. The percentage variance effect size (r) as recommended by Larson-Hall (2010) was also reported where statistically significant differences were found. The r values of 0.10, 0.30 and 0.50 were considered cut-off values for small, medium and large effect sizes respectively (Field, 2017).

The audio-recorded interviews were transcribed in their entireties by four authors, two of whom were in charge of each level. Then, each pair cross-checked the accuracy of the transcription before the fifth author conducted a comprehensive check of the whole set of transcripts. The data were then analysed in the original language of the interviews (Vietnamese). By conducting the analysis in the source language, the intended meanings were well retained (Casanave, 2010). This was an iterative open process of coding through generating and regenerating themes, confirming and disconfirming them in an iterative manner from provisional to confirmational as informed by the data (Newman, 2014). Yin (2015) recommended that both original and translated interview quotes should be known to the reader for their own interpretation. However, due to space constraint, only the translated extracts of interview were presented in the present paper. The precision of the translated texts was checked by an experienced EFL teacher. Pseudonyms instead of the real names of the interviewees were co-presented with interview quotes for de-identification and confidentiality.

Results

Computer-delivered vs. face-to-face speaking performances of the VSTEP.2 Speaking Test

The first research question sought to compare test takers' scores in the two testing modes of the VSTEP.2 Speaking test. Descriptive statistics for the VTEP.2 Speaking Test performances are presented in Table 5.

In the VSTEP.2 speaking test, the mean scores for all the analytical criteria tended to be higher in the semi-direct mode except pronunciation and fluency. The median values in vocabulary and content development were also higher in this test mode, but a higher median was recorded for communication strategies in the direct format. Variability

Table 5 The descriptive statistics for the VSTEP.2 speaking performances in the computer-delivered and face-to-face modes

Grading criteria	Computer-delivered ($n=75$)						Face-to-face ($n=75$)					
	Min	Max	Mean	SD	Mdn	Range	Min	Max	Mean	SD	Mdn	Range
Grammar	2.00	4.00	3.40	0.57	3.00	2.00	2.00	4.00	3.33	0.64	3.00	2.00
Pronunciation	2.00	4.00	3.28	0.56	3.00	2.00	2.00	4.00	3.35	0.60	3.00	2.00
Vocabulary	2.00	4.00	3.45	0.62	4.00	2.00	2.00	4.00	3.33	0.64	3.00	2.00
Fluency	2.00	4.00	3.24	0.54	3.00	2.00	1.00	4.00	3.24	0.77	3.00	3.00
Content development	2.00	4.00	3.57	0.57	4.00	2.00	1.00	4.00	3.21	0.83	3.00	3.00
Communication strategies	2.00	4.00	3.44	0.58	3.00	2.00	1.00	4.00	3.36	0.75	4.00	3.00
Total score	*12.00*	*24.00*	*20.39*	*2.67*	*22.00*	*13.00*	*9.00*	*24.00*	*19.83*	*3.55*	*22.00*	15.00

Table 6 Ranks of test scores in the VSTEP.2 Speaking test in the computer-delivered and face-to-face modes

		n	Mean rank	Sum of ranks
Grammar f2f-C	Negative ranks	17	15.00	255.00
	Positive ranks	12	15.00	180.00
	Ties	46		
	Total	75		
Pronunciation f2f-C	Negative ranks	9	11.22	101.00
	Positive ranks	13	11.69	152.00
	Ties	53		
	Total	75		
Vocabulary f2f-C	Negative ranks	18	16.50	297.00
	Positive ranks	12	14.00	168.00
	Ties	45		
	Total	75		
Fluency f2f-C	Negative ranks	15	17.63	264.50
	Positive ranks	17	15.50	263.50
	Ties	43		
	Total	75		
Content development f2f-C	Negative ranks	32	21.56	690.00
	Positive ranks	9	19.00	171.00
	Ties	34		
	Total	75		
Communication strategies f2f-C	Negative ranks	16	14.34	229.50
	Positive ranks	11	13.50	148.50
	Ties	48		
	Total	75		
Total scores f2f-C	Negative ranks	33	33.26	1097.50
	Positive ranks	27	27.13	732.50
	Ties	15		
	Total	75		

f2f, face-to-face; *C*, computer-delivered

Table 7 The Wilcoxon signed ranks test results for the VSTEP.2 Speaking test in the computer-delivered and face-to-face modes

	Grammar f2f-C	Pronunciation f2f-C	Vocabulary f2f-C	Fluency f2f-C	Content Development f2f-C	Communication Strategies f2f-C	Total score f2f-C
Z	−0.928[a]	−0.898[b]	−1.459[a]	−0.010[a]	−3.707[a]	−1.095[a]	−1.354[a]
Asymp. Sig. (2-tailed)	0.353	0.369	0.145	0.992	0.000	0.273	0.176

f2f, face-to-face; *C*, computer-delivered; *a*, based on positive ranks; *b*, based on negative ranks

tended to be greater in the direct mode (higher range values in the last three categories and in the total score). Since the data were not normally distributed, Wilcoxon signed-rank tests were run and the results (Tables 6 and 7) revealed no significant difference in the overall total scores between the two testing forms, $Z = -1.354$, $p = 0.176$. Regarding the six analytical criteria, the results further show that there were no significant

differences between the two formats of testing in performance scores in all the components except content development. In particular, in the computer-based mode, test takers performed better in terms of content than speaking with the presence of a human rater, $Z = -3.707$, $p < 0.001$, with a large effect size ($r = 0.58$). As further seen from Table 6, 32 test takers (32 negative ranks) had a higher content development score in the computer-based mode while only nine did so in the face-to-face mode, and 34 had similar scores in the two modes (34 ties). This suggests that the computer-delivered mode of testing was more facilitative to students, though only in organising content of their speaking performance.

Computer-delivered vs. face-to-face speaking performances of the VSTEP.3–5 speaking test

The second research question sought to compare test takers' scores of the VSTEP.3–5 Speaking test in the face-to-face and the semi-direct formats. Descriptive statistics for the VSTEP.3–5. test performances are presented in Table 8.

As shown in Table 8, for the VSTEP.3–5 Speaking test, all the mean values appeared to be higher in the face-to-face mode, though the medians were the same. Like the VSTEP.2 test, greater variation was observed in the direct mode than the computer-delivered format (greater range values in all the analytical criteria and the total scores). The results of the Wilcoxon signed rank test (for the non-normally distributed data) are summarised in Tables 9 and 10. They show that when taking the VSTEP.3–5 test with a live examiner, the total test score was significantly higher than when responding to task prompts from a computer, $Z = -2.020$, $p = 0.043$, though the effect size was small ($r = 0.23$). However, regarding the individual scores in relation to the different dimensions of the VSTEP.3–5 Speaking performances, the results of the Wilcoxon signed rank test indicated that the difference was not statistically significant in the dimensions of grammar ($Z = -1.212$, $p = 0.225$), vocabulary ($Z = -1.109$, $p = 0.267$), fluency ($Z = -1.396$; $p = 0.163$) and content development ($Z = -1.868$, $p = 0.062$). Yet, a significantly higher score was given for pronunciation when test takers sat the VSTEP.3–5 test with the presence of a human, $Z = -2.547$, $p = 0.011$, $r = 0.33$. In particular, as seen from Table 9, 41 test takers (41 positive ranks) scored higher in pronunciation in the face-to-face mode while only 18 did so in the computer-delivered format. The findings speak in favour of the direct oral testing format in terms of English pronunciation for this particular VSTEP.3–5 test. Notably, the difference in the content score approached significance ($p = 0.062$) in favour

Table 8 The descriptive statistics for the VSTEP.3–5 speaking performances in the computer-delivered and face-to-face modes

Grading criteria	Computer-delivered ($n = 82$)						Face-to-face ($n = 82$)					
	Min	Max	Mean	SD	Mdn	Range	Min	Max	Mean	SD	Mdn	Range
Grammar	5.00	9.00	7.01	1.01	7.00	4.00	4.00	9.00	7.15	1.17	7.00	5.00
Vocabulary	5.00	10.00	7.18	1.00	7.00	5.00	4.00	10.00	7.33	1.25	7.00	6.00
Pronunciation	5.00	10.00	7.24	1.07	7.00	5.00	5.00	10.00	7.57	1.16	7.00	5.00
Fluency	5.00	9.00	7.01	0.99	7.00	4.00	3.00	10.00	7.21	1.23	7.00	7.00
Content development	5.00	9.00	7.07	0.98	7.00	4.00	4.00	10.00	7.34	1.38	7.00	6.00
Total score	*25.00*	*47.00*	*35.52*	*4.60*	*35.00*	*22.00*	*21.00*	*49.00*	*36.60*	*5.72*	*37.00*	*28.00*

Table 9 Ranks of test scores in the VSTEP.3–5 Speaking test in the computer-delivered and face-to-face modes

		n	Mean rank	Sum of ranks
Grammar f2f-C	Negative ranks	21	24.67	518.00
	Positive ranks	29	26.10	757.00
	Ties	32		
	Total	82		
Vocabulary f2f-C	Negative ranks	23	29.07	668.50
	Positive ranks	33	28.11	927.50
	Ties	26		
	Total	82		
Pronunciation f2f-C	Negative ranks	18	31.56	568.00
	Positive ranks	41	29.32	1202.00
	Ties	23		
	Total	82		
Fluency f2f-C	Negative ranks	22	33.41	735.00
	Positive ranks	38	28.82	1095.00
	Ties	22		
	Total	82		
Content development f2f-C	Negative ranks	23	30.26	696.00
	Positive ranks	38	31.45	1195.00
	Ties	21		
	Total	82		
Total f2f-C	Negative ranks	28	39.45	1104.50
	Positive ranks	49	38.74	1898.50
	Ties	5		
	Total	82		

f2f, face-to-face; *C*, computer-delivered

Table 10 The Wilcoxon signed ranks test results for the VSTEP.3–5 Speaking test in the computer-delivered and face-to-face modes

	Grammar f2f-C	Vocabularyf2f-C	Pronunciation f2f-C	Fluency f2f-C	Content development f2f-C	Total score f2f-C
Z	− 1.212[a]	− 1.109[a]	− 2.547[a]	− 1.396[a]	− 1.868[a]	− 2.020[a]
Asymp. Sig (2-tailed)	0.225	0.267	0.011	0.163	0.062	0.043

f2f, face-to-face; *C*, computer-delivered; a. Based on negative ranks

of the face-to-face testing mode, with more test takers (38 vs. 23) scoring higher in content development for the VSTEP.3–5 Speaking test in this study.

Test takers' experiences of the computer-delivered and face-to-face testing

The third research question aimed to understand how the test takers experienced the two testing formats. Their recounts in the interviews revealed refreshing insightful information on the different affective preferences involved in the process of being orally assessed with a human assessor and without in a computer-based condition for the two popular English speaking proficiency tests in Vietnam. In general, 11 and 12 out of 15

interviewees in the VSTEP.2 and VSTEP.3–5 tests respectively preferred the face-to-face mode while the remaining favoured the computer-based format. However, those who reported preferring a given test mode did not necessarily achieve a higher overall score in that format (Table 11). For example, in the case of VSTEP.2 Speaking test, of the four test takers who expressed a liking for the computer-based format, two achieved a higher score (interviewees 4, 8) and two equal scores (interviewees 11 and 15). Similarly, those who favoured the face-to-face format were not always rated higher in this test mode. For example, while interviewees 2, 9, 10 and 14 achieved higher scores, interviewees 1, 3, 6, 7 and 11 to name a few and others (e.g. interviewees 5, 12, 13) did not. A similar pattern was noted for the VSTEP.3–5 test takers. For instance, interviewees 2 and 14 had a lower score in the semi-direct mode, though it was their preference. In addition, interviewees 1, 3, 13 and 15 rated the face-to-face test more favourably and also had a higher score (41 vs. 29, 46 vs. 39, 47 vs. 42 and 42 vs. 37 respectively) whereas others (e.g. interviewees 8, 11) were rated lower in their preferred test mode. In other words, there was great variation among individual interviewees. However, this quantitative finding needs to be interpreted with care since the number of interviewees was quite small (15 per test). Regarding the test modes, it is essential to understand the qualitative insights from test takers' perspectives that is how they viewed the different aspects of the test modes.

Table 11 Interviewees' preferences for the test modes and their total scores

	VSTEP.2 (*n* = 15)			**VSTEP.3–5 (*n* = 15)**		
	Computer-based	**Face-to-face**	**Preferred test mode**	**Computer-based**	**Face-to-face**	**Preferred test mode**
	Total score	***Total score***		***Total score***	***Total score***	
Interviewee 1	22	22	Face-to-face	29	41	Face-to-face
Interviewee 2	22	23	Face-to-face	34	37	Computer-based
Interviewee 3	24	24	Face-to-face	39	46	Face-to-face
Interviewee 4	16	15	Computer-based	37	35	Face-to-face
Interviewee 5	21	19	Face-to-face	37	37	Computer-based
Interviewee 6	24	24	Face-to-face	43	46	Face-to-face
Interviewee 7	24	24	Face-to-face	40	40	Face-to-face
Interviewee 8	18	15	Computer-based	40	35	Face-to-face
Interviewee 9	21	24	Face-to-face	38	39	Face-to-face
Interviewee 10	23	24	Face-to-face	34	37	Face-to-face
Interviewee 11	24	24	Computer-based	32	31	Face-to-face
Interviewee 12	18	13	Face-to-face	36	31	Face-to-face
Interviewee 13	23	18	Face-to-face	42	47	Face-to-face
Interviewee 14	17	22	Face-to-face	43	47	Computer-based
Interviewee 15	21	21	Computer-based	37	42	Face-to-face

Their mixed preferences were related to the extent to which they viewed how each test mode replicates real-life communication and provides a supportive testing environment or otherwise. This mainly centered around the presence or absence of a human rater and further revealed the concurrent tensions between being *communication-oriented* and *performance-driven* in this and other inherent affordances or constraints of the two modes of delivery.

Genuine communication as supportive

The physical presence of a human rater in the direct testing condition was repeatedly mentioned as one key attribute that prompted content generation and increased the amount of talk. Many candidates conceptualised the human rater in their face-to-face test as *"a person to talk to"* or *"a person who is there to listen to me"*, which enabled more idea generation:

> *In the face-to-face speaking test, having a person to talk to encourages me to have more ideas than speaking to the machine in the computer-based test. (Linh, VSTEP.3-5)*

Obviously, the rater was affectively viewed as a human interlocutor being there to talk and co-talk rather than an assessor per se. In other words, in front of a human rater, these test takers tended to reveal themselves to be *communication-oriented*.

In this respect, the two-way interaction and back channeling in the direct mode was another affordance which reportedly motivated test takers to "try to talk" to sustain communication. For them, timely feedback was obtained via various verbal and non-verbal means in the presence of a live assessor. Verbal means were deployed by the test takers themselves through asking for clarification and through examiners' responses:

> *I prefer the direct speaking test because I can ask clarification questions if needed. Two-way communication with the interlocutor engages me more, and motivates me to speak English more than talking to the computer. (Minh, VSTEP.2)*

> *I would go for the face-to-face test as I feel that I have more ideas to talk when I interact with the teacher. In addition, the teacher can repeat a question so that I can speak. (Quang, VSTEP.3-5)*

The comments above have shown that the direct mode allows test takers to respond and clarify, which was considered a form of support. Interestingly, several test takers narrated the self-inflicted "guilt" that could ensue if they were silent in front of a live human rater. That "I can't keep silent" drove them to speak:

> *I feel that I cannot keep silent if someone is there with me. Saying nothing when being with someone makes me feel guilty, so I am motivated to speak more. (Vy, VSTEP.3-5)*

Test takers' self-perceived responsibility to talk in the physical presence of an interlocutor points to the *human* aspect of interpersonal communication where emotions and affective factors come into play in the speaking act, especially in a testing situation. This

resonates with a range of negative emotions test takers experienced in the non-human delivery mode:

> *I could neither develop ideas well nor use good vocabulary when I was with the computer. I was sitting there, staring at the emotionless cold, plain computer screen, and I couldn't use any strategies at all. Only when I am with someone else, I can start brainstorming ideas, planning what to say, reminding myself to avoid repetition, and using body language. I can also use "you know" to fill the pauses in my talk. However, if someone is there in front of me, I cannot just keep silent. (Chau, VSTEP.3-5)*

For this test taker, the "cold" "emotionless" "plain" non-human computer screen was experienced as a constraint which hindered vocabulary and strategy use and reduced talk. Such a comparison is a subtle indication that this particular candidate and some others with similar views needed the human rater to function better. An absence of a true need for communication, a missing human interlocutor, was perceived to cause distraction and disfluency:

> *When I took the computer-based test, the feeling that no-one was listening to me or looking at me easily made me distracted. Sometimes I didn't notice that the timer on the screen had started counting down, and I lost a few seconds before I started to speak. (Thuy, VSTEP.3-5)*
> *I was advised that I should stay focused on the topic … but it was a real challenge to really focus when you were there talking to the computer. It is really hard to imagine that you are interacting with a real person if what is in front of you is just the computer screen. (Huy, VSTEP.3-5)*

The irresponsive computer screen was clearly a block to communication in test takers' view. Accordingly, a lack of authentic communication was reportedly a major disadvantage of the semi-direct mode due to its unnatural delivery:

> *When I took the computer-based speaking test, my feeling was that I was delivering a speech, not interacting with others in real-life communication. (Nhan, VSTEP.2)*

Rater non-verbal feedback as both enabling and constraining

Several candidates verbalised the appeal of the live testing mode was to support them through non-verbal responses from the human raters (e.g. smiles, nods) which even predict test outcome or performance quality:

> *What I like best about the face-to-face test is that I could guess my score by seeing the teacher's smile or looking at her eyes. These things also gave me some hints to adjust my talk. (Mai, VSTEP.3-5)*

> *During the direct speaking test, I could tell if my performance was good or not by reading the teacher's facial expressions. (Minh, VSTEP.3-5)*

While verbal and non-verbal feedback from a human rater was a source of support and motivation for many test takers to interact and sustain communication, the presence of a human examiner was negatively viewed by many others who preferred the reduced

negative emotions in the computer-mediated context. For them, rater impact was absent as they did not have to confront a live rater and as such not experiencing their negative verbal or nonverbal feedback that could be fear/anxiety-inducing:

> *During the computer-based test, I did not have to perform in front of the examiner. It scared me whenever the examiner frowned. Whenever I saw that, I was at a loss for words. (Tri, VSTEP.3-5)*

> *Taking the computer-based test meant that you were with the computer only, so there was no extra pressure from being there with the examiner as in the face-to-face condition. (Ngoc, VSTEP.2)*

The candidate responses were mixed. While the physical presence of a human rater was generally perceived as advantageous for many test takers, five interviewees in the VSTEP.2 Speaking test and six in the VSTEP.3–5 Speaking test considered sitting for the test in the live condition were more constraining than the semi-direct mode. In particular, raters' attitudes and non-linguistic expressions could trigger negative emotions for some candidates, such as worry and anxiety which prevented further talk. One VSTEP.2 test candidate narrated:

> *I was worried that the examiner was too strict. Her facial expressions such as frowning or discontentment really made me worried, which discouraged me from completing my speaking test. (Duyen, VSTEP.2)*

Notably, the negative non-verbal feedback from the rater such as frowning or dissatisfaction was perceived to be related to *rater harshness or severity*, which could be devastating for students, inhibiting communication or cause embarrassment:

> *My biggest fear was being with a strict examiner. An unhappy look or a frown from her can make me worried, which easily messed up my speaking performance. (Nhi, VSTEP.2)*

It is intriguing and at the same time poignant to note that these test takers interpreted rater severity/difficulty through their negative non-linguistic reactions. This reference to rater severity was felt at a greater level of intensity by candidates of the two tests. This was particularly true for those who considered being likely to be affected by raters' emotional reactions:

> *I am a type of emotional person. I am easily affected by others' comments about me, so just a frown from the examiner can scare me. (Giang, VSTEP.2)*

Ten test takers additionally cited worry and anxiety stemming from the possible failure on the part of the examiner to understand their talk, as well as their own fear of not successfully communicating intended ideas:

> *I am extremely worried that I have to face the examiner. My listening skill is not good enough, so I cannot understand what the teacher says. (Yen, VSTEP.2)*

Confronting' teachers (raters), who were authorities and more competent interlocutors in the testing room posed great pressure in the live speaking session. One candidate commented:

I am not used to speaking English to my teachers, so I easily get nervous although I know they won't do anything to scare me at all. (Huong, VSTEP.2)

Rater vs. textual prompts as conductor

Earlier, test takers cited the subjective benefits and constraints related to the presence of a human interlocutor and its inherent communication-oriented characteristics. Seven of them explicitly reported the *objective* advantages of the semi-direct testing format in terms of *being led* in the speaking session:

About other 'objective' benefits, I think as the teacher was there, she could guide me through the different speaking tasks, which saves me time in planning what to do next. (Tien, VSTEP.3-5)

In their views, such guidance saved time and alleviated the cognitive burden of subsequent planning. One common message from those who saw the benefit in this way was that "*Just follow the teacher rater's instructions and everything will be alright.*" Overall, those test takers who viewed the direct mode positively reported a sense of reassurance as they had the examiner steer their live speaking session.

Contrary to a more *communication orientation* in the direct testing mode, test takers who expressed preferences for the semi-direct mode displayed a more pronounced *performance/outcome-orientation*. About four test takers in both proficiency tests considered the textual display of the speaking task prompts on the computer screen enabled them to read as many times as they wished and thus gained time to think about what to say. While this was viewed negatively due to lack of authenticity by many candidates, being able to *read* the task questions in the computer-delivered testing context was a clear advantage for many others:

I prefer to take the computer-based speaking test because I can read the task prompts many times, which gives me enough time to plan my speech. (Binh, VSTEP.2)

Fairness and procedural issues

Again, tensions exist as individual preferences varied. Preferences for the semi-direct test delivery were related to its having a higher level of fairness as the issue of rater variation and subjectivity could be avoided, this time interestingly in the *absence of the examinee* who is speaking:

The computer-based format ensures more test fairness as the raters do not see the candidates' faces, and they do not know who is speaking. (Huong, VSTEP.2)

Implicit in examinees' comments were the likelihood of raters judging each visible candidate differently, leading to unfairness. Fairness was also mentioned because of the equal time length allocated for every test taker in the computer-delivered testing context:

The computer-based speaking test is fairer, as the time allotment for each and every candidate's performance is exactly the same. (Binh, VSTEP.3-5)

In the direct face-to-face testing environment, the *waiting time* involved further added pressure for some candidates, as they had to wait for their turn in a waiting room to be assessed directly by a teacher rater, which was perceived to lead to fatigue and minimal language production:

> *Sometimes the waiting game is tiring. When it is your turn to take the test, you cannot perform your best due to fatigue after a long wait. (Yen, VSTEP.3-5)*

On the other hand, the live condition was perceived to be more relaxing in terms of time. Test takers reported not to suffer time pressure when being interviewed by a human assessor while the countdown timer in the semi-direct mode was more pressing. Importantly, time limit was "*softer*" in the direct mode as raters would be flexible for overtalk:

> *I think during the direct speaking test, the examiner won't interrupt me if she thinks I still have ideas to share. On the contrary, the timer counting down on the computer screen really scares me. (Nhan, VSTEP.2)*

For the computer-delivered testing environment, technical issues were also reported. For example, many test takers (18/30) recounted negative emotions such as anxiety and fear of mistaken operation and failure to submit their test performance:

> *I was afraid that the recording quality was not up to standard, or that my mistake with using the computer would cause the loss of my speaking test. (Suong, VSTEP.2)*

The impending threat that things could go wrong was perceived to induce a high level of anxiety and pressure among test takers. Furthermore, even though the setup for the computer-delivered testing involved enclosed cabins in much separation from each other as logistically allowed, noise originating from other test takers speaking concurrently in the test room was a major source of distraction and low-quality recordings. This is well captured in the following comments:

> *I was unable to concentrate on the test because of the fear that the computer could record my friend's voice instead of mine. (Hoai, VSTEP.3-5)*

> *My biggest challenge was that I was easily distracted by the noises around me. Sometimes, I was talking fluently, and paused for a little while. During that pause, as other friends' voices landed on my ears, my mind was directed towards their talks, which distracted me from my own thoughts. (Tien, VSTEP.3-5)*

Several students additionally reported that the headphones were very rough and it hurt, especially when they had to wear glasses at the same time. Yet, the semi-direct mode was perceived to be more appropriate for active, independent self-reliant test takers:

> *We had to be self-reliant during the computer-based speaking test. Self-reliant students would perform their best under this testing mode. (Lien, VSTEP.2)*

The computer-delivered mode was also perceived to help students manage response time and adjust volume as long as there is sufficient practice:

During the computer-based test, I found it easier to time myself and adjust the volume. I knew how loud my voice was, so I could adjust the volume to the optimal level. It's also easy to time our speaking, so it just takes a bit of practice to perform well in the test. (Thu, VSTEP.3-5)

These comments suggest that it might take test takers' autonomy and proactivity to operate well in the non-live testing environment.

Discussion

This section discusses the findings of the study, in relation to the comparability in scores between the computer-delivered and face-to-face modes for the VSTEP.2 and VSTEP 0.3–5 Speaking tests according to VNFLPF, and test takers' experiences of the different formats of test delivery.

Score comparability in the computer-delivered and face-to-face testing modes

For the VSTEP.2 speaking test, the overall results indicated that test takers achieved similar outcomes regardless of testing modes. There were no significant differences between the two testing modes in all the analytical categories (grammar, pronunciation, vocabulary, fluency, content development and communication strategies) except content development; that is, test takers benefited only in the content development dimension in the semi-direct mode of testing. This finding partially corroborates those found in Zhou's (2015) study which shows that the two modes of testing were comparable in that Japanese high school EFL students achieved similar scores in all dimensions of pronunciation, lexical resource, grammar and fluency. The minimal effect of testing mode was also echoed in Quaid and Barrett's (2021) study, though on overall fluency. Generally high comparability also finds support in other prior research (e.g. Choi, 2014; Khabbazbashi, 2017; Yonezaki, 2016).

Yet, it is interesting that for the VSTEP.3–5 speaking test, an impact of testing mode was found with a higher overall average test score in favour of the direct mode than the semi-direct condition. It is even more interesting to observe the marked difference in the only category of pronunciation, but not in any other analytical criteria. The value approaching significance in the content development score ($p = 0.062$) might further suggest the greater advantage of the live oral testing for the Vietnamese EFL candidates in the VSTEP.3–5 speaking test in the present study. All these findings illustrate that testing mode had selective effects on performance scores for these groups of test takers and this impact might be subject to the type of proficiency test taken.

An important question to ask is why the impact of testing mode was inconsistent. In the present study, candidates in the VSTEP.2 speaking test were first year English majors who were in their second semester at the university. They had less prior experience with the direct test than their senior third year counterparts who took the VSTEP.3–5 test. It could be that with more extensive experience of the face-to-face test as a routine format of oral formative and summative assessment for these students at the research site, the VSTEP.3–5 test takers might have been more aware of employing strategies to be successful when talking with a live human rater, thus obtaining a higher overall test score than in the computer-delivered testing environment. Equally, they might have become more successful communicators with their higher English proficiency as they were

senior students. This again suggests that proficiency might have a moderating effect on the equivalence of the two testing conditions.

It is worthy to note that the grading scheme was not the same for the two proficiency tests with an absence of the measure of communication strategies in the VSTEP.3–5 Speaking test. It was unexpected that mode of test delivery had no discernable effect on the use of communication strategies in the VSTEP.2 Speaking test. Arguably, while the live environment with a human assessor could have been more enabling, it remains unclear how raters scored the full range of communication strategies without seeing the test takers' face while scoring the audio-recorded performances in the computer-based mode. It could be that raters might have attended to the audio input in both conditions or else test takers did not display use of strategies sufficiently in the direct mode. The greater advantage of the face-to-face test observed in the VSTEP.3–5 on pronunciation substantiates the finding by Kiddle and Kormos (2011) that test takers scored significantly higher on this measure in the face-to-face test than in the computer-based test. This finding could be attributable to the fact that pronunciation is often rated intuitively (Derwing & Munro, 2005) and EFL teachers/raters have different orientations towards nativelikeness (Brown et al., 2005; Deterding, 2010). Intuitive scoring and subjective orientation are more likely in view of the inherent affordances such as non-linguistic resources and proximity between the raters and candidate in the direct mode.

Test takers' experiences of the computer-delivered and face-to-face testing modes

Overall, more test takers in the present study valued the inherent affordances in the live environment such as verbal and non-verbal communication, real-life interaction and two-way communication than those who did not. This generally confirms that direct testing is more ecologically valid as it replicates real-life communication (Brahim, 2023; van Lier, 1989) and thus it has more face validity (Bijani, 2019; Qian, 2009; Yonezaki, 2016). Test takers have shown mixed preferences for each testing mode, and this finding is broadly congruent with prior research (Baralt & Gurzynski-Weiss, 2011; Bijani & Khabiri, 2017; Chang et al., 2018; Qian, 2009). Yet, examinees' experiences have shown nuanced insights into the wide range of positive and negative emotions involved in the particular affordances or constraints of each testing mode.

The greater correspondence between the direct testing mode and real-life communication did not always receive unanimous positive perceptions and many candidates still preferred the computer-delivered context without a human rater for its non-human interaction features. This could be attributed to how individual test takers viewed the opportunities and limitations in each testing environment. To be more legitimate, how they utilised the affordances and coped with the wide range of emotions involved to strive for their performance surfaced as important. The mixed affective responses could well be related to individual learning styles, for example, how field-independent and -dependent learners function in a noisy and distracting environment with multiple test takers speaking in the semi-direct context, or else how visual and auditory learners might need different support channels to best perform the intended test. That more VSTEP.3–5 than VSTEP.2 interviewees in the present study referred to how independent, self-reliant and goal-directed learners could fit better with the semi-direct mode might well indicate proficiency could be an individual factor. A link between affective

and learner-related factors and oral performance is not a novel finding (e.g. Jalilzadeh & Yeganehpour, 2021; Liu, 2018), but this connection needs further exploration in oral assessment with different delivery modes.

Above all, the confidence, psychological support, motivation to talk stemming from the self-inflicted "guilt" of being silent or co-constructed talk characteristic of the nature of speaking (Brooks & Swain, 2014; Brown, 2003; Swain, 2001), or the fear, the self-oriented or examiner-oriented anxiety in response to rater positive and negative non-verbal feedback as well as the plain and cold computer screen viewed in the computer-based speaking all denote idiosyncratic positions that might have reduced the impact of test mode on performance scores. They are perhaps not the mere manifestations of personal affective preferences but could well suggest the kind of testing environment that each individual test taker might need to best perform. Above all, the test takers' experiences have shown that the act of being assessed orally could be cognitively and emotionally taxing, just as cognition and emotion are inseparable in speaking (Brooks & Swain, 2014; Swain, 2013). In a high-stakes test, it makes sense that test takers could be strategic and proactive to curb negative emotions to strive to achieve a high score in a non-live environment. That test takers could be performance-oriented at the sacrifice of authentic communication does negate the value of communication, but rather point to the importance of how to best support test takers to perform in an oral test, which is addressed next.

Implications and conclusions

In the present study, 75 and 82 Vietnamese EFL test takers sat the VSTEP.2 and VSTEP.3–5 speaking tests respectively in both computer-delivered and face-to-face modes in a counterbalanced manner. The results revealed mixed effects of the testing mode, with the overall test scores being compatible in the VSTEP.2 test, but significantly higher in the face-to-face mode for the VSTEP.3–5 test. Furthermore, for the VSTEP.2 test, test takers performed significantly better on one single measure of content development in the computer-based condition while only pronunciation benefited in the face-to-face mode for the VSTEP.3–5 test. Test takers also showed mixed affective preferences in view of the inherent affordances or constraints of each testing mode.

Implications

In light of the findings of the present research, important implications for test administration in terms of rater and test taker training, and extrapolation are addressed.

Firstly, that the overall test score did not differ statistically in both direct and semi-direct testing modes of the VSTEP.2 Speaking test points to the validity of the computer-delivered as a viable alternative to the traditional face-to-face format for this particular test. Decision-making needs to consider which form of assessment would be appropriate for assessment purposes in different contexts.

Secondly, considering the fact that test takers were much more familiar with the direct testing mode than the semi-direct one where they had to talk to a computer, the significant difference though in only one measure of the many categories (content development in the VSTEP.2 test and pronunciation in the VSTEP.3–5 Speaking test, though mode effects were mixed) could be taken as a positive sign of the latter mode of delivery.

Training and practice need to be in place to reduce novelty, as Baralt and Gurzynski-Weiss (2011) found a lower level of apprehension when test takers became more familiar with the testing mode, be it direct or semi-direct. As the semi-direct mode of testing was less familiar with the test takers than the direct format in both the VSTEP.2 and VSTEP.3–5 Speaking tests in the present study, more empirical knowledge needs to be acquired from other contexts of Vietnam to better inform test administration.

Thirdly, speaking in the computer-delivered context as reported by many test takers in the present study does not characterise real-life interaction, which is otherwise more possible in the presence of a live rater. This might hinder accurate interpretations from computer-based test performances to performances in real life situations. Kiddle and Kormos (2011) well noted "the threat of construct underrepresentation through the lack of interaction in computer-based tests" (p. 342). It is thus crucial to define the construct of interest (Bachman & Palmer, 1996) to measure it accurately in any delivery practice for the purpose of extrapolation. Likewise, if the inherent nature of speaking is co-constructed as seen in authentic two-way interaction (Brooks & Swain, 2014; Brown, 2003; Swain, 2001), how to replicate this co-construction at least in the direct test if it is the sole testing mode choice is important. Equally important, rater and test taker variability could well lead to unfairness and unreliability. While this could be reduced by scripting interlocutors in live testing to standardise its procedure, it is important to organise rater training that should go beyond a linguistic focus to include norms or rules on how to behave non-verbally in direct oral testing.

In view of authentic communication, given the constraints of a computer-delivered speaking test, it is common to see monologic tasks being used in many oral proficiency testing contexts. This kind of task clearly limits the level of authenticity of test and its correspondence to real-life contexts where multiple interlocutors might be involved in oral discourse. Task design features or pair exams could be further considered, of course in light of the practicality that each testing site could afford.

To achieve the equivalence of the two testing conditions, how to replicate communicative features in the semi-direct delivery practice is an additional concern. The question should not always be which mode of delivery to use in place of the other, but rather which factors to consider when decision is made upon administering a certain test, especially high-stakes ones in either mode. As the semi-direct mode of delivery could be economical and fairer given its standardised procedures and its identical task input to all test takers (Leaper & Riazi, 2014), the authenticity of response which is lacking in the computer-delivered mode could be enhanced by incorporating varied forms of task input. Textual task prompts could be replaced by audio or video versions to increase interactiveness. More advanced technology with a higher level of interactiveness and friendliness (instead of a countdown timer) could be orchestrated with perhaps cross-disciplinary collaboration to create a more supportive testing environment. The findings of the present study revealed that test takers' preferences varied, depending on how they viewed they could be best supported by either mode of testing to perform at their full potential. In this regard, test providers should create a supportive anxiety-free environment for examinees, for example, reducing noise from adjacent speakers in the same test room, preparing quality headphones to enhance their confidence. Appropriate support needs to be orchestrated in different forms as it has been shown to correlate with

self-confidence (Rees & Freeman, 2007) and performance scores (Fu et al., 2021). With a high-stakes test like the level 3–5 test that could affect thousands of undergraduates and post-graduates in Vietnam, careful consideration of the constraints and affordances in each testing mode is crucial in providing adequate support for test takers.

Limitations and suggestions for further studies

The current research has several shortcomings that need attention. First of all, quantitative data such as test scores as employed in the present study might not supply sufficient empirical evidence for the interchangeability of the two testing modes. Future research might need to consider analysing live and computer-mediated speaking performances qualitatively through a variety of linguistic features such as language functions, turn-taking, quality of use of vocabulary among others (Quaid, 2018). Next, the views of raters when grading live and recorded performances, their rating behaviours, what they expect test takers to perform, or their personal inclinations would provide richer information on how the testing mode could influence test results. This is a worthy avenue for further studies on the comparability of the testing modes in particular regard to the two proficiency tests under study in Vietnam to inform rater and test taker training.

Though the counterbalanced research design was adopted, the "recency effect" (Quaid, 2018), the effect of the fact that test takers have performed a similar task regardless of test modes could be an additional confounding factor. Future research could employ more complex analyses such as many-facet Rasch model or general linear mixed-effects model test (GLMT) to better understand individual factors in accounting for any variance in test scores. Furthermore, the present study only utilised the raw test scores, thus preventing more nuanced insights to be gleaned from the use of Item Response Theory (IRT) in linking and equating test forms and test scores. Besides, it did not explore test takers' perceived test difficulty which could have been a confounding factor contributing to inconsistent mode effects. For example, the VSTEP.2 speaking test that targets an A2 level might be more "suitable" for the current VSTEP.2 test takers with proficiency levels ranging from A1 to A2. Meanwhile, the VSTEP.3–5 test targeting B1-C1 levels could be more challenging for the VSTEP.3–5 test takers (B1-B2 levels) in the present study. However, this account is only speculative and awaits research that additionally examines how test takers rate the difficulty of the target tests to better understand the comparability of the direct and semi-direct modes in each test and across different tests.

In addition, a majority of the test takers and all raters in the present study were female, suggesting that the impact of gender could be further explored in future studies, as some research (e.g. O'Sullivan, 2000) has shown female participants felt greater comfort in conversing with an female interviewer. Equally, the test takers in our study were English-majored university students, who might have experienced the employed testing modes differently from other groups of EFL learners of different proficiency levels, learning styles and prior English learning experience. Therefore, more extensive research in other contexts of Vietnam and on a larger scale with a larger sample size to further testify the equivalence of the two testing modes for the administration of the target English speaking tests is needed. Particularly, a larger number of interviewees could further allow a more systematic analysis of the relationship between test takers' test mode preferences and their test performances.

Abbreviations

CEFR	Common European Framework of Reference for Languages
EFL	English as a foreign language
MOET	Ministry of Education and Training of Vietnam
NTC	National Testing Center
TEFL	Teaching English as a foreign language
VNFLPF	Vietnam's Six-Level Foreign Language Proficiency Framework
VSTEP	The Vietnamese Standardized Test of English Proficiency

Acknowledgements
The authors are grateful to the Vietnamese students and raters for their participation in this research.

Authors' contributions
All the authors were responsible for the design of the study, its execution and respective parts of data collection and analysis. Further details related to the formation of the paper are described as follows. Thuy Ho Hoang Nguyen: conceptualisation, study design, data collection, data analysis, writing, revising and editing. Bao Trang Thi Nguyen: conceptualisation, study design, data collection, data analysis, writing, revising, editing, submitting/correspondence. Giang Thi Linh Hoang: conceptualisation, study design, data collection, data analysis, revising. Nhung Thi Hong Pham: conceptualisation, study design, data collection, data analysis, revising. Tu Thi Cam Dang: conceptualisation, study design, data collection, data analysis, literature review, revising.

Notes on authors
Thuy Ho Hoang Nguyen obtained an MA and a PhD both in Applied Linguistics from the University of Queensland, Australia. She is currently a lecturer at the Faculty of English, University of Foreign Languages and International Studies, Hue University, Vietnam. Her research interest and publications are mainly in Applied Linguistics and English language education.
Bao Trang Thi Nguyen works as a lecturer at the Faculty of English, University of Foreign Languages and International Studies, Hue University, Vietnam. Her field of research is task-based language teaching and learning, learner proficiency and SLA. She has a number of book chapters published by John Benjamins, Bloombury, Springer and Routledge. Her research articles also appear in different journals such as Language Teaching Research, TESOL Journal, Asia Pacific Journal of Education, International Journal of Comparative Education and Development, Language Related Research and Canadian Journal of Applied Linguistics.
Giang Thi Linh Hoang is a lecturer at the Faculty of English, University of Foreign Languages and International Studies, Hue University, Vietnam. She obtained an MA in TESOL and a PhD in Applied Linguistics and has disseminated research results at international conferences in Applied Linguistics. She has published articles in journals like Language Assessment Quarterly, Language Related Research and the ***JALT CALL Journal***. Her research interests are language assessment, especially automated writing assessment, and second language acquisition research.
Nhung Thi Hong Pham obtained an MA and a PhD in Applied Linguistics with the University of Queensland, Australia. She has taught various postgraduate cosurses for the Faculty of English, University of Foreign Languages and International Studies, Hue University, Vietnam, and published in the field of Applied Linguistics and Language Education.
Tu Thi Cam Dang works as a lecturer at the Faculty of English, University of Foreign Languages and International Studies, Hue University, Vietnam. She gained her BA in English Language Teaching at Hue University in 2011 and Master Degree in Applied Linguistics at Victoria University of Wellington, New Zealand, in 2015. Her research interests include Language Teaching Methodology and Applied Linguistics.

Funding
The authors would like to thank the National Foreign Language Project, Ministry of Education and Training, Vietnam, for funding this research project.

Availability of data and materials
Due to the privacy of the data related to the current study, they are not made publicly available, but will be provided upon reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Received: 9 November 2023 Accepted: 12 February 2024
Published online: 26 February 2024

References

Bachman, L. F., & Palmer, L. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.
Baralt, M., & Gurzynski-Weiss, L. (2011). Comparing learners' state anxiety during task-based interaction in computer-mediated and face-to-face communication. *Language Teaching Research,15*(2), 201–229. https://doi.org/10.1177/0265532210388717

Bijani, H. (2019). Evaluating the effectiveness of the training program on direct and semi-direct oral proficiency assessment: A case of multifaceted Rasch analysis. *Cogent Education,6*(1), 1670592. https://doi.org/10.1080/2331186X.2019.1670592
Bijani, H., & Khabiri, M. (2017). Direct and semi-direct validation: Test takers' perceptions, evaluations and anxiety towards speaking module of an English proficiency test. *Journal of Language and Translation.,1*(13), 25–41.
Brahim, Y. (2023). Computer-based vs face-to-face speaking assessment: Fitness for purpose from a communicative language testing view. *International Journal of Social Science and Human Research,6*(1), 22–30. https://doi.org/10.1177/0265532210388717
Brooks, L., & Swain, M. (2014). Contextualizing performances: Comparing performances during TOEFL iBTTM and real-life academic speaking activities. *Language Assessment Quarterly,11*(4), 353–373. https://doi.org/10.1080/15434303.2014.947532
Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language Testing,20*(1), 1–25. https://doi.org/10.1191/0265532203lt242o
Brown, A., Iwashita, N., & McNamara, T. (2005). An examination of rater orientations and test-taker performance on English-for-academic-purposes speaking tasks. *ETS Research Report Series, 2005*(1), i-157.
Casanave, C. P. (2010). *Distancing: From real life experiences to final research report in qualitative inquiry with multilingual participants*. In *the 2010 Symposium on Second Language Writing.* University of Murcia, Spain.
Chang, S. L., Lee, S. D., & Lee, S. P. (2018). Divergent effects of direct and semi-direct oral assessment on psychological anxiety and physiological response in EFL college students. *Asian EFL Journal,20*(12), 252–269.
Choi, I. (2014). The comparability of direct and semi-direct oral proficiency interviews in a foreign language context: A case study with advanced Korean learners of English. *Language Research,50*(2), 545–567.
Derwing, T., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly,39*(3), 379–397. https://doi.org/10.2307/3588486
Deterding, D. (2010). ELF-based pronunciation teaching in China. *Chinese Journal of Applied Linguistics,33*(6), 3–15.
Field, A. (2017). *Discovering statistics using SPSS* (5th ed.). SAGE.
Frost, K., & McNamara, T. (2018). Language tests, language policy, and citizenship. In J. W. Tollefson and M., Pérez-Milans (eds.), *The Oxford handbook of language policy and planning* (pp. 280–298). Oxford University Press.
Fu, D., Hase, A., Goolamallee, M., Godwin, G., & Freeman, P. (2021). The effects of support (in)adequacy on self-confidence and performance: Two experimental studies. *Sport, Exercise, and Performance Psychology,10*(1), 15–26. https://doi.org/10.1037/spy0000206
Garcia-Laborda, J. (2007). On the net: Introducing standardized EFL/ESL exams. *Language Learning and Technology,11*(2), 3–9.
Jalilzadeh, K., & Yeganehpour, P. (2021). The relationship between intermediate EFL students' oral performance, communicative willingness, as well as emotional intelligence. *The Reading Matrix: An International Online Journal,21*(2), 165–179.
Jeong, H., & Hashizume, H. (2011). Testing second language oral proficiency in direct and semi-direct settings: A social-cognitive neuroscience perspective. *Language Learning,61*(3), 675–699. https://doi.org/10.1111/j.1467-9922.2011.00635.x
Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing,34*(1), 23–48. https://doi.org/10.1177/0265532215595666
Kiddle, T., & Kormos, J. (2011). The effect of mode of response on a semi-direct test of oral proficiency. *Language Assessment Quarterly,8*(4), 342–360. https://doi.org/10.1080/15434303.2011.613503
Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing,33*(3), 319–340. https://doi.org/10.1177/0265532215587
Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. Routledge.
Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing,31*(2), 177–204. https://doi.org/10.1177/0265532213498237
Liu, M. (2018). Interactive effects of English-speaking anxiety and strategy use on oral English test performance of high- and low-proficient Chinese university EFL learners. *Cogent Education,5*(1), 1562410. https://doi.org/10.1080/2331186X.2018.1562410
Marian, A. P., & Jesus, G. L. (2017). Analysing test takers' views on a computer-based speaking test. *Profile Issues in Teachers Professional Development. 19*(1):23–38. https://doi.org/10.15446/profile.v19n_sup1.68447
Mirzaei, A., Hashemian, M., & Farsani, M. (2016). Lexis-based instruction and IELTS candidates' development of L2 speaking ability: Use of formulaicity in monologic versus dialogic task. *Journal of Teaching Language Skills*, *35*(2), 69–98. https://doi.org/10.22099/jtls.2016.3816
MOET. (2014). *Circular No. 01/2014/TT-BGDDT dated January 24, 2014 of the MOET on promulgating The Six-level Foreign Language Proficiency Framework for Vietnam.* Retrieved from https://thuvienphapluat.vn/van-ban/Giao-duc/Thong-tu-01-2014-TT-BGDDT-Khung-nang-luc-ngoai-ngu-6-bac-Viet-Nam-220349.aspx
MOET. (2015). *Decision No. 729/QD-BGDDT dated March 11, 2015 of the MOET on promulgating Specifications of English Proficiency Test Level 3–5 in accordance with The Six-level Foreign Language Proficiency Framework for Vietnam.* Retrieved from https://thuvienphapluat.vn/van-ban/Giao-duc/Quyet-dinh-729-QD-BGDDT-2015-de-thi-danh-gia-nang-luc-su-dung-tieng-Anh-tu-bac-3-den-bac-5-267956.aspx
MOET. (2016). *Decision No. 1481/QD-BGDDT dated May 10, 2016 of the MOET on promulgating Specifications of English Proficiency Test Level 2 in accordance with The Six-level Foreign Language Proficiency Framework for Vietnam.* Retrieved from https://thuvienphapluat.vn/van-ban/Giao-duc/Quyet-dinh-1481-QD-BGDDT-de-thi-danh-gia-nang-luc-su-dung-tieng-Anh-bac-2-nguoi-lon-2016-311836.aspx
MOET. (2017). *Circular No. 23/2017/TT-BGDDT dated September 29, 2017 of the MOET on promulgating Regulations of Assessing Foreign Language Proficiency in accordance with The Six-level Foreign Language Proficiency Framework for Vietnam.* Retrieved from https://thuvienphapluat.vn/van-ban/Giao-duc/Thong-tu-23-2017-TT-BGDDT-thi-danh-gia-nang-luc-ngoai-ngu-theo-Khung-nang-luc-ngoai-ngu-6-bac-363395.aspx

MOET. (2021). *Circular No. 24/2021/TT-BGDDT dated September 8, 2021 of the MOET on amending and supplementing some articles of the Circular No. 23/2017/TT-BGDDT issued on September 29, 2017 of the MOET on promulgating Regulations of Assessing Foreign Language Proficiency in accordance with The Six-level Foreign Language Proficiency Framework for Vietnam.* Retrieved from https://thuvienphapluat.vn/van-ban/Giao-duc/Thong-tu-24-2021-TT-BGDDT-sua-doi-Thong-tu-23-2017-TT-BGDDT-nang-luc-ngoai-ngu-6-bac-Viet-Nam-461565.aspx

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2021). Video-conferencing speaking tests: Do they measure the same construct as face-to-face tests? *Assessment in Education: Principles, Policy & Practice,28*(4), 369–388. https://doi.org/10.1080/0969594X.2021.1951163

Newman, L.W. (2014). *Social research methods: Qualitative and quantitative approaches (7th ed)*. Pearson.

O'Sullivan, B. (2000). Exploring gender and oral proficiency interview performance. *System,28*(3), 373–386. https://doi.org/10.1016/S0346-251X(00)00018-X

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. (2018). Output register parallelism in an identical direct and semi-direct speaking test: A case study. *International Journal of Computer-Assisted Language Learning and Teaching,8*(2), 75–91. https://doi.org/10.4018/IJCALLT.2018040105

Quaid, E., & Barett, 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.

Rees, T., & Freeman, P. (2007). The effects of perceived and received support on self-confidence. *Journal of Sciences,25*(9), 1057–1065. https://doi.org/10.1080/02640410600982279

Shohamy, E., Shmueli, D., & Gordon, C. (1991). The validity of concurrent validity of a direct vs. semi-direct test of oral proficiency. In 13th Language Testing Research Colloquium, Princeton, NJ.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing,11*(2), 99–23. https://doi.org/10.1177/026553229401100202

Stansfield, C. W. (1991). *Simulated oral proficiency interviews: An update. ERIC Digest*. ERIC Clearinghouse on Languages and Linguistics.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System,20*(3), 347–364. https://doi.org/10.1016/0346-251X(92)90045-5

Swain, M. (2001). Examining dialogue: Another approach to content specifications and to validating inferences drawn from test scores. *Language Testing,18*(3), 275–302. https://doi.org/10.1177/0265532201018003

Swain, M. (2013). The inseparability of cognition and emotion in second language learning. *Language Teaching,46*(2), 195–207. https://doi.org/10.1017/S0261444811000486

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly,23*, 489–508. https://doi.org/10.2307/3586922

Yin, R. K. (2015). *Qualitative research from start to finish (2nd edition)*. The Guilford Press.

Yonezaki, M. (2016). A comparative analysis of semi-direct speaking testing and direct speaking testing for Japanese EFL learners. *International Journal of Curriculum Development and Practice*, *18*(1), 27–38. https://doi.org/10.18993/jcrdaen.18.1_27

Zhou, Y. J. (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

Publisher's Note