Pertanika J. Soc. Sci. & Hum. 29 (S3): 223 - 242 (2021)

PERTANIKA JOURNALS

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Perspectives of Test Examiners of the Localized Speaking Assessment Framework: A Case Study in Vietnam

Thi Nhu Ngoc Truong[1]*, Arshad Abd Samad[2] and Thi Thanh Phan[3]

[1]University of Economics Ho Chi Minh City, 59C Nguyen Dinh Chieu Street, Ward 6, District 3, 700 000 Ho Chi Minh City, Vietnam
[2]Faculty of Educational Studies,Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia
[3]University of Labor and Social Affairs (Campus 2), 1018 To Ky Street, Tan Chanh Hiep Ward, District 12, 700 000 Ho Chi Minh City, Vietnam

ABSTRACT

The present study explores the test examiners' perspectives on the role and qualitative aspects of the current localized speaking assessment framework used in Vietnam. A case study with two experienced test examiner-cum-English lecturers was conducted. Inductive content analysis was used to analyze the qualitative data findings obtained from individual semi-structured interviews. Drawbacks, merits, and standardization issues of the current localized speaking assessment frameworks, i.e., the Vietnamese Standardized Test of English Proficiency (VSTEP), were also discussed, especially in comparison to internationally recognized examinations and frameworks such as the International English Language Testing System (IELTS) and the Certificate in Advanced English (CAE) as well as the Common European Framework of Reference for Languages (CEFR). The study informed both English educators and policymakers to improve localized speaking assessment to suit the local teaching needs while still meeting the requirements of widely accepted international proficiency tests.

Keywords: CEFR, speaking assessment, speaking skill, test examiners, VSTEP

ARTICLE INFO

Article history:
Received: 16 July 2021
Accepted: 04 October 2021
Published: 30 November 2021

DOI: https://doi.org/10.47836/pjssh.29.S3.12

E-mail addresses:
ngocttn@ueh.edu.vn (Thi Nhu Ngoc Truong)
arshad@upm.edu.my (Arshad Abd Samad)
thanhpth@ldxh.edu.vn (Thi Thanh Phan)
*Corresponding author

INTRODUCTION

Assessing oral production is often a challenging task as the nature of language comprises explicit knowledge, which

ISSN: 0128-7702
e-ISSN: 2231-8534

students learn through formal schooling, and implicit knowledge when they are exposed to multimedia sources and real-life communicative settings. Moreover, during oral tests, students have to process information to use grammar, vocabulary, and phonology appropriately and may also be called upon to demonstrate sociolinguistic competence (Canale & Swain, 1980; Liontas & Siegel, 2019). All these expectations in speaking assessment represent a challenge for the students to produce native-like speech (Seifoori & Vahidi, 2012). A major concern for language examiners, thus, revolves around the need for explicitly delineated objective criteria for marking oral skills that take into consideration all the aspects of effective speaking ability.

Efforts to develop and improve criterion-referenced assessment for the speaking skill have been highlighted by language scholars (Liu & Jia, 2017). International assessments and frameworks such as the Common European Framework of Reference (CEFR) for languages have been very influential in this respect in the past few years. Interestingly, there have been increased calls for more localized tests and assessment frameworks in language assessment practices to meet the demands of various groups of learners in EFL countries, for example, the General English Proficiency Test (GEPT) in Taiwan (Wu, 2012) and the Fudan English Test in China (Fan & Ji, 2014). In line with this trend, the Vietnamese Standardized Test of English Proficiency (VSTEP), approved in 2015, has been gradually used by many local educational institutions to replace international English proficiency exams in Vietnam (Nguyen et al., 2020) and is considered an alternative to international tests such as TOEIC, PET, KET, and IELTS (T. N. Q. Nguyen, 2018).

Since VSTEP was introduced, many training programs have been conducted for test examiners, writers, and validators. A few studies have reported the effectiveness of those training programs on test validity or ratings (T. N. Q. Nguyen, 2018; Nguyen et al., 2020; T. P. T. Nguyen, 2018). Nevertheless, none of these studies have focused on the qualitative aspects and practices of the VSTEP speaking test from the practitioners' perspective. This study, therefore, reports on the views of two experienced test examiners who are English university instructors in Vietnam regarding the current speaking assessment practice in using the existing localized speaking test, especially within the prevailing context of internationalization. In particular, this study seeks to answer the following questions:

1. What are the participants' perceptions of the VSTEP speaking test?
2. What are the participants' perceptions of the VSTEP speaking assessment practice?
3. What do the participants think about standardization in speaking assessment practices in Vietnam?

LITERATURE REVIEW

English Language Assessment Practices in the Vietnamese Context

Overview of the English Language Assessment Practices in Vietnam. English is a mandatory subject in the Vietnamese educational system for all academic levels and a compulsory national examination for high school students to enter university (Hoang, 2010). English language assessment in Vietnam has undergone three main phases (Vu, 2016). During the pre-scientific phase in the 1990s, pre-constructed test papers were mainly designed by lecturers in top universities; around 100 to 200 preconstructed mock test papers for each subject, including English, were released to the public for students to review before the official exam dates (Vu, 2016). Selected universities, designated by the Ministry of Education and Training (MOET), subsequently chose a set of test papers randomly, edited and censored them to produce the official test paper. These pre-constructed test papers were designed to narrow the curricular content as an early form of standardization for the whole country; however, they became counterproductive because they encouraged teachers' teaching to the tests and learners' rote learning.

The second period, 1996 to 2007, was directed towards standardization for reliability (Vu, 2016). Electronic marking with closed-ended questions was first piloted in 1996 for national exams to avoid raters' subjectivity and errors in scoring. In 2002, the multiple-choice question university entrance exam was promulgated under the three policies: same paper, exam date, and results, which all universities used for admission decisions. To enhance the English teaching and learning quality and meet the challenges of globalization, in 2008, the MOET approved the National Foreign Language Project 2020, aiming to produce a skilled workforce able to communicate competently, independently, and confidently in a multicultural and multilingual environment (The Prime Minister of the Socialist Republic of Vietnam, 2008). A standardized speaking test and assessment framework that meets international standards while being localized to suit the national needs was essential to fulfill the goal.

The National Foreign Language Project 2020 marked the third stage, standardization for reliability and validity (Vu, 2016). This project adopted the CEFR and proposed the six-level foreign language competency framework for Vietnam in 2012, called the Common European Framework of Reference - Vietnam (CEFR-V). The six levels of competency in the CEFR-V, parallel to those of the CEFR, were localized to orientate English curriculum design and assessment (Hoang, 2010; Le et al., 2017; T. Nguyen, 2017; T. N. Q. Nguyen, 2019; Pham & Bui, 2019). Accordingly, CEFR-V was introduced at the primary and secondary school levels. Meanwhile, at the tertiary level, the foreign language curriculum is decided by each institution following guidelines provided by the government. Following this policy, Vietnamese students

can choose either a domestic (i.e., VSTEP) or an international English language proficiency test (e.g., FCE, IELTS, and TOEFL) to take as long as they obtain at least level 3 of the CEFR-V or B1 on the CEFR to graduate (Le et al., 2017; Pham & Bui, 2019).

Vietnamese Standardized Test of English Proficiency (VSTEP). The first localized proficiency test, i.e., VSTEP 3-5, was designed by Vietnam National University in 2012, assessing four skills: listening, speaking, reading, and writing (T. N. Q. Nguyen, 2019). After three years of planning, designing, and piloting, in 2015, the MOET approved the official utilization of the national standardized VSTEP.3-5 and the CEFR_V as the benchmark for English language assessment nationwide (Nguyen et al., 2020). Following the VSTEP.3-5 test format (assessing English proficiency levels 3, 4, and 5 of CEFR-V, equivalent to levels B1, B2, C1 of CEFR), other variants of VSTEP were designed such as VSTEP.1 (i.e., level 1 or A1-CEFR), VSTEP.2 (level 2 or A2-CEFR), and even level 6 (or C2-CEFR) which is supposed to be beyond the English capacity of the majority of Vietnamese people (T. N. Q. Nguyen, 2019). VSTEP tests were designed with a globalized quality and localized to meet the national standards. In effect, these tests are considered a reliable instrument to measure the English ability of Vietnamese adult learners from different professions and levels of qualification (T. N. Q. Nguyen, 2019).

Because of its more comprehensive range of users compared to other VSTEP tests, in this paper, we focused on VSTEP.3-5, which is more common for most employees and students in Vietnam. The test aims to test interaction, discussion, problem-solving, and presentation skills and includes three parts: social interaction (comprising 3-6 questions about two different topics), solution discussion (requiring students to select, present, and defend their solution to a given situation from three suggested solutions), and topic development (requiring students to ask questions about a given topic using prompts to develop their ideas) (MOET, 2015) (see Appendix B for a VSTEP sample test). VSTEP.3-5 scores are measured on a scale from 0 to 10, based on five marking criteria: grammar (range and accuracy), vocabulary (range and control), pronunciation (individual sounds, stress, and intonation), fluency (hesitation and extended speech), and discourse management (coherence, cohesion, and thematic development) (MOET, 2015).

Factors Relevant to Oral Performance

Speaking assessment practices can be affected by diverse factors such as task and interlocutor characteristics, test validity and reliability, assessment criteria (Fan & Yan, 2020; Kang & Wang, 2014), rater effects (McNamara et al., 2019), and rater training (Kang et al., 2019). Studies have shown that interaction tasks involving interactions with examiners can be unnatural compared to paired or oral group tests, in which the test taker interacts

with another test candidate (Brooks, 2009; Winke, 2013). O'Sullivan (2002) found an acquaintanceship effect in an experimental study with 32 Japanese students for decision making, personal information exchange, and narrative tasks, subsequently confirmed by Norton (2005) in the document analysis of 15 transcribed recordings of pairs of candidates for the FCE test in the UK. These studies indicate that subjects achieved higher scores when collaborating with a friend rather than a stranger. Interaction effects between the gender of the interlocutor and acquaintanceship were also found for grammatical accuracy (O'Sullivan, 2002).

Ahmadi and Sadeghi (2016) found that accuracy, fluency, and complexity differed across three tasks (monologue, interview, and group discussion), and accuracy was significantly correlated with the analytical and holistic assessment scores. The score differences between these two assessment methods were also documented in prior studies (Namaziandost, 2019; Namaziandost et al., 2019). Moreover, because candidates' oral performance is assessed through speech features, dependent on the test purpose and construct, these features may have different score weightings (Plough, 2018). Also, various tasks require different responses and rating scales, affecting the test-takers performance (Chalhoub-Deville & Wigglesworth, 2005). For instance, responsive and interactive tasks require the test-taker to interact more with an interlocutor peer or test examiner than the imitative, intensive, and extensive tasks (Brown & Abeywickrama, 2010).

Despite these multi factors, the basics of speaking assessment involving scales, raters, and methods should be considered to ensure the reliability and validity of speaking assessment (Ginther, 2020). While methods and scales are objective and predetermined, rater characteristics and rater bias were reported to be inconsistent over time (Lumley & McNamara, 1995). Moreover, raters' elicitations of demonstrating speaking competence, structuring talk sequences, and questioning techniques lead to variations in the impressions of candidates' ability (Brown, 2000). However, regular rater training can improve rating accuracy and minimize rating bias (Bijani, 2018; Kang et al., 2019). Surprisingly, there were no significant differences between non-native and native speakers as assessors in the outcome scores in some studies (Rossiter, 2009; Zhang & Elder, 2014), although EFL teachers viewed native speakers and their pronunciation as ideal models (Walkinshaw & Duong, 2012).

To sum up, although numerous factors such as rating scales, task characteristics, and rater effects have been reported in previous studies as influential variables to speaking assessment quality (Fan & Yan, 2020; McNamara et al., 2019; O'Sullivan, 2002), very few studies provide insights into test examiners' perspectives about these factors. Besides, qualitative aspects of VSTEP have not been extensively reported in the literature (Nguyen, 2015). Thus, exploring expert raters' perceptions of VSTEP and its speaking assessment practice can provide initial insights into influential

variables and issues that have not been reported in previous studies on speaking assessment in Vietnam (Nguyen et al., 2020; T. P. T. Nguyen, 2018).

METHODOLOGY

Participants

Vietnamese university lecturers were invited through a TESOL network in Vietnam to participate in the study. Two female English lecturers aged 35 and 38 were subsequently recruited based on their experience in teaching and testing as VSTEP speaking examiners. Using the pseudonyms Anna and Jane, Table 1 provides demographic data of the participants.

Table 1

Participants' demographic information

Name	Institution	Years of working as an English lecturer	Years of working as a general English-speaking examiner.	Years of working as VSTEP speaking examiner	Familiarity with speaking assessment frameworks
Anna	Private University	9	7	1	IELTS, CEFR & VSTEP
June	Public University	8	7	1	IELTS, CEFR & VSTEP

Data Collection

Data were collected using in-depth semi-structured interviews. The first and third authors contacted participants via phone calls to introduce themselves, explain the purpose of the study, and schedule meetings. The participants who agreed to participate in the study signed an informed consent form. The authors agreed on an interview protocol, and the third author interviewed the participants using Skype video calls after informing them the interviews would be recorded. The interview questions involved teaching and testing experiences, perception of the localized VSTEP speaking test and assessment practice, and views regarding standardization of speaking assessment practices in Vietnam (see Appendix A). All their personal information, participation in the study, and recorded interviews were kept confidential.

Data Analysis

Data were gathered, collected, transcribed, and analyzed using inductive content analysis guidelines suggested by Creswell's (2002) guidelines. The researcher organized the qualitative data through open coding and created categories for abstraction. Accordingly, the researcher clarified the

content by writing notes and headings during reading and rereading. The final coding scheme comprises inductive codes. When common patterns were found within and across cases, the researcher identified disconfirming cases and patterns before checking and rechecking codes with data and clustering them into categories. The researcher continued revising and refining the category system, and within each category, the researcher searched for sub-topics, including contradictory viewpoints and new insights. Suitable direct quotes from the interviews were used to illustrate, support, validate the findings (Thomas, 2006).

Reliability and Credibility

Our findings were based on raw data. We employed reliability procedures, including conducting multiple transcripts reviews to reduce mistakes in the participants' narratives of their experiences (Creswell, 2007). Multiple authors were involved in the coding process. Our positionality was employed as a form of reliability (Merriam & Tisdell, 2015). As the researchers, we were aware that reflexivity affected how we made meaning of the participants' worldviews. The position of the first and third authors as full-time university English lecturers and speaking examiners in Vietnam also provided access to and acceptance by our participants.

RESULTS AND DISCUSSION

This section presents and discusses the findings in three categories: perceptions of the VSTEP speaking test, rater training and styles, and beliefs about standardization in speaking assessment practices in Vietnam.

Perceptions of the VSTEP Speaking Test

Both examiners agree that the format of the VSTEP speaking test is well-organized and localized, with three main parts varying from basic to a higher level of task difficulty (social interaction, solution discussion, and topic development). Also, participants shared similar opinions in that the assessment criteria for VSTEP are detailed, although the test lacks natural interaction and a high level of reasoning skills.

Localization and Authenticity. Anna believed that the VSTEP test was localized and reliable. In addition, the language of instruction in the VSTEP test is easy to understand as test writers considered different language backgrounds, especially students from low to high levels of English.

> *I think the test is somehow reliable, and it is localized. However, the sentences and questions in VSTEP are very short though the wording of the task requirements is clear. It may be because the test writers think that candidates' general English proficiency is not high, so when they write the test, they aim at students of the average English proficiency level. (Anna)*

Besides, both Jane and Anna found the three parts in a VSTEP speaking test similar to those in the international standardized

test, e.g., IELTS. However, they both opined that VSTEP topics were sensitive to the Vietnamese contexts for international exchange inside the country. For Anna, this might be a unique feature of VSTEP. Anna also elaborated that VSTEP topics had higher authentic features involving real-life situations.

> *The task in part 2 is designed to suit the Vietnamese context, more practical, authentic, and applicable, compared to part 2 in IELTS, which is related to personal topics. Part 2 in VSTEP requires test-takers to explain a problem to a person and the choice they go for to solve the problem…This test is more useful in real life than IELTS because students need to communicate and discuss and persuade others regarding practical problems. (Jane)*

> *[In] IELTS, students just talk about a given topic like a book they prefer… However, sometimes, many students do not like reading books [and] may not have any ideas to talk about...For this reason, I think the authenticity of the IELTS test is not high. In VSTEP, students are presented with three options, and at least students get interested in one of the three given options. So, the authenticity is high. (Anna)*

However, Anna felt that test questions written by non-native speakers were still less reliable and natural than those written by native speakers. She justified her view by stating that she sometimes found grammar and spelling mistakes in the VSTEP speaking tests written by Vietnamese test writers compared to the questions written for international standardized exams.

> *Regarding the wording of the test questions and description of tasks, I think as VSTEP tests are written by the Vietnamese, they may not sound as good as those in IELTS written by native English speakers…Sometimes, there are some typing or spelling mistakes, lacking the verb "to be" or auxiliary verbs in VSTEP tests. (Anna)*

Detailed Assessment Criteria. Both participants have positive attitudes towards the VSTEP rating scale at the macro level, stating that the marking rubrics are clear and detailed. The new criterion, which they did not find in other familiar frameworks such as CEFR and IELTS, concerns discourse management, including coherence, cohesion, and theme development. When comparing VSTEP with IELTS, Anna highlighted the importance of the discourse management criterion. She explained that this criterion enabled her to give an accurate assessment of other criteria.

> *In IELTS, there are no discourse management criteria, so I think this is a drawback in IELTS as discourse management is very important. If we give an accurate assessment of students' discourse management which comprises thematic development, coherence, and cohesion, in VSTEP we can mark the remaining criteria accurately. However,*

if we give the wrong assessment of students' discourse management, we may not accurately assess other criteria such as grammar, vocabulary, pronunciation, and fluency. (Anna)

Likewise, Jane considered discourse management the "new assessment criterion," although she did not highlight its importance. She added that VSTEP assessment criteria were designed for analytical assessment, enabling her to assess students' performance accurately because "it is possible to give scores on a scale from 1 to 10 more precisely." However, when comparing tasks in VSTEP with those in the CAE exam, Jane mentioned that VSTEP speaking tasks were less interactive. For example, in the VSTEP part 1, although the candidate interacted with the assessor, the interaction was not natural.

[T]he CAE test is more interactive… because in some tasks, candidates interact with each other, and they have to discuss and share opinions. Although part 1 of the VSTEP speaking test is titled Social Interaction, the examiners ask only one candidate for a speaking test session, and the candidate shares their personal experience. Thus, they are not actually interacting socially with the examiner. (Jane)

Sharing the same opinion, Anna emphasized that interaction criterion be included for VSTEP because this could ensure authenticity and identify candidates memorizing prepared notes.

[In] CEFR, they have the interaction criterion to avoid students' preparation of tasks in advance…In CEFR, the interaction task is designed to test students' natural interaction with others...I think interaction should be a criterion in VSTEP. (Anna)

However, Anna cautioned that the interlocutor's characteristics could affect the test taker's performance in paired oral tasks. For example, she explained, "sometimes one student says something the other student cannot grasp the main idea due to bad pronunciation, and this creates some difficulty for the test-taker."

In terms of reasoning skills, Anna and Jane posited that the purpose of part 3 in both VSTEP and IELTS is similar, i.e., assessing reasoning skills and a higher level of linguistic ability. However, they commented that VSTEP part 3 was more manageable than IELTS part 3 because candidates were given a mind map with three provided ideas as prompts to enable test-takers to elicit their ideas. Meanwhile, in IELTS part 3, candidates must discuss follow-up topics at a more macro level without hints, clues, or prompts, requiring a higher level of cognitive thinking and knowing a wide range of social and educational topics.

In part 3 of the IELTS test, candidates will be asked deeper argumentative questions to share their views on

more macro issues... Part 3 of VSTEP speaking is similar when candidates are asked about macro follow-up questions but only after being given a topic with a concept map. The suggestion on this mind map facilitates the candidate to answer and give them hints… In my opinion, part 3 of the two tests is quite similar because both assess the linguistic ability and reasoning ability of candidates. (Jane)

About task 3 in VSTEP speaking test, there is a mind map of ideas, so for candidates who need time to think about ideas, they can still base on 3 suggested ideas to come up with their ideas. (Anna)

In general, both participants had positive views about the VSTEP speaking test and its assessment criteria, which contribute to content validity information of the VSTEP test that was previously validated for the reading and writing skills (T. N. Q. Nguyen, 2018 & T. P. T. Nguyen, 2018). Furthermore, because candidates could apply the solution situation in VSTEP topics to the real-world context, sociocultural expectations were considered in constructing the VSTEP tests. However, as Anna revealed that VSTEP oral tests written by non-native speakers contain errors and do not sound natural in terms of wording, it seems that Anna may view native speakers as ideal models of standard English, consistent with EFL teachers' beliefs about native speakers reported by Walkinshaw and Duong (2012). Besides, because paired tasks and the assessment of interaction are not included, participants indicated that the authentic interaction level in the one-to-one interview in VSTEP speaking tasks was less natural, which coincides with Brooks' (2009) and Winke's (2013) empirical findings on the superiority of the interactive nature in paired tasks.

Also, cognizant that the interlocutor's characteristics could influence the examinee's performance, as reported in previous studies (Norton, 2005; O'Sullivan, 2002), Anna may mean that various tasks should be included to cancel out weaknesses of each task. However, a need for different task types may mean different task-specific assessment scales, as Chalhoub-Deville and Wigglesworth (2005) posited. Thus, VSTEP designers should consider which tasks can bring more positive washback effects to improve the current oral test. Finally, the provision of prompts in VSTEP task 3 indicates that the VSTEP test considers EFL learners' characteristics and difficulties in language processing, which may also explain why the requirement of the reasoning skill is not so high in Vstep oral tasks.

Rater Training and Styles

Intensive Rater Preparation. Participants affirmed that training is an essential element for effective assessment. For example, Jane shared that to become an official VSTEP speaking examiner, she had to complete "a two-week training program," including "120 offline periods for writing and speaking assessment" and "240 periods for online studies" and mark at least ten students' oral performances together with

an experienced examiner. She added that the scores difference between her and another rater's marking "should not exceed two scores." Attending the same program, Anna reflected that the VSTEP training equipped her with useful assessment knowledge. She also observed a change in her assessment style because the detailed descriptions of the VSTEP rating scale that she was trained with rendered the assessment procedure more logical and transparent to her.

> *I think I learned many new things when participating in training programs. Before, I only gave a subjective assessment based on my experience, but when I attended the training, I was given the rating scale with detailed descriptions for every criterion. I think the speaking assessment becomes clearer and logical. (Anna)*

Positive Assessment Style and Rater Drifts. Both Jane and Anna disclosed that they based their assessment style on the 'can-do mindset,' i.e., the candidate's actual oral production guided in the rater training. Although the participants did not explain why they adopted the 'can-do mindset,' it can be inferred that the 'can-do statements' describing the proficiency levels in CEFR-V may be transformed into the 'can-do mindset' assessment, i.e., the positive assessment style for the VSTEP oral test.

> *[We] do not deduct students' marks but assess them based on the "can-do mindset." (Anna)*

> *For example, when a test-taker does not perform very well in part 1, but in part 2 and 3, they can perform well, I mark [their] performance based on what they have performed and what they can answer. I don't deduct scores. (Jane)*

However, although the VSTEP assessment rubrics were designed to assess students' oral performance more analytically, Jane shared that she often used a holistic instead of an analytic approach to assessing overall performance.

> *[M]ost often, I have a holistic assessment of students' performance after performing all three tasks. But this is not mentioned in the rating scale. (Jane)*

Likewise, Anna reflected that although the marking rubric was extensively descriptive, she usually did not have sufficient time to refer to the rubric during oral exam marking because she had to listen to the test taker's responses. Hence, she relied on her memorizing the general description for each band score and her subjective experiences to mark her students' responses.

> *[T]here is not enough time to simultaneously listen to students' performance and read the band descriptions to give them scores…I remember the general description of each band score, and based on my personal experience, to give students marks. (Anna)*

Besides, another reason for the change from an analytic to a holistic scoring approach after time elapsed from the training is that both participants found that several sub-criteria in the rubric were not clearly described.

> *[T]he descriptions in the band scores in some criteria are sometimes overlapped, such as band 5 and 6. Sometimes, I don't know whether to give the student 5 or 6 scores for their performance. (Anna)*
>
> *The scores in the middle range like 4,5,6,7 and band descriptors for these scores easily confuse test examiners… I am sometimes confused because I do not know which score in the rating scale I should go for. (Jane)*

Thus, it appears that from both participants' perspectives, training English lecturers to become test examiners was a necessary step towards standardization in speaking assessment practices, which echoes findings from previous studies that rater training mitigates rater bias and improves rating consistency (Bijani, 2018; Kang et al., 2019). However, the participants' assessment style changes are quite surprising because both Anna and Jane stated that they highly valued the detailed description of the VSTEP rating scale, which enabled them to give objective assessments to test takers. Just as Lumley and McNamara (1995) observed, this "rater drift," however, is somewhat reconciled through moderation at the end of the grading process. As task types are related to analytic and holistic approaches (Ahmadi & Sadeghi, 2016), and rating rubrics can affect effective assessment (Fan & Yan, 2020), more training on the differences among criteria descriptors for each band score and assessment approaches for different task types can lead to improved standardization in speaking assessment practices.

Beliefs about Standardization in Speaking Assessment Practices in Vietnam

Regarding the necessity for standardization in speaking assessment practices at the national level, Anna and Jane shared a similar viewpoint that standardization was indispensable to ensure fairness, equity, and consistency among universities.

> *Standardizing speaking assessment practices is necessary to ensure equity and fairness. Right now, each educational institution has its way of assessing its students. Thus, this lacks synchronization and accuracy in assessing students. (Anna)*
>
> *If talking about standardization in speaking assessment practices, the promulgation of general regulations for one common framework for all educational institutions to adapt to their context can be a good choice. (Jane)*

However, both were cautious about the inherent difficulties of unifying speaking assessment practices for all local educational institutions due to differences in university contexts, learner proficiency level, and training programs.

Although it is necessary to standardize speaking assessment practice, I think it is also difficult…because students at public and private universities have different proficiency levels. (Anna)

[S]tandardizing the speaking skills assessment framework by applying the VSTEP framework…may not be necessary for some non-public organizations because they can follow the international frameworks which are more suitable for their teaching and learning context or the needs of overseas cooperation and study. (Jane)

In responding to which assessment framework should be used to standardize speaking assessment practices at the national level, Anna believed that VSTEP should be used because it was suitable "for most working people, secondary and tertiary students." However, she recommended that "English majors should study IELTS, and non-English majors should take VSTEP" because topics of VSTEP were "localized." Likewise, Jane suggested that students who planned to study overseas "should study IELTS [which] can benefit them in the long term" because VSTEP was not globally recognized. However, if students did not intend to study abroad, VSTEP could be a better choice because of "its low cost."

In general, participants hold a balanced view towards standardization in speaking assessment practices because proficiency evidence can be proved by either local or international standardized assessment dependent on the test-taking purposes and the training institutions. Since the locally produced VSTEP has not yet gained international recognition, both favored IELTS for overseas studies and academic advancement. Taking stock of the current speaking assessment practices in Vietnam, if VSTEP is to be gradually globally recognized just as other locally standardized proficiency tests, e.g., GEPT in Taiwan (Wu, 2012), validation of VSTEP speaking test and addressing challenges related to its speaking assessment rubrics is necessary.

IMPLICATIONS

To sum up, participants expressed the need to include local content in the test design, the interaction criterion in the rating scale, the importance of receiving training, and the necessity to balance standardization in speaking assessment. The participants emphasized a positive perspective towards completing the test task (as seen in their can-do mindset) and believed that assessing actual speaking ability should not be clouded by students first understanding foreign and unfamiliar contexts. VSTEP seems an appropriate assessment tool that considers localized contexts besides meeting localized objectives, especially for local employees and non-English majors. However, teachers continue to refer to the IELTS as it seems that the VSTEP has not yet received global acceptance. Despite this, the VSTEP is still a successful test. It measures language ability based on internationally accepted criteria (as indicated by its close and careful association to the CEFR), and many characteristics make VTEP a practical speaking test.

Notably, test task characteristics are important concerns raised by the two participants. Different test tasks will elicit various kinds of language as numerous factors can make speaking a complicated activity involving a high cognitive level of information processing and knowledge and consequently a difficult skill to assess. For example, more prompts were provided in the VSTEP than other tests as observed by one participant, which is understandable in an EFL context as it can encourage speech production. Although prompts make the task easier, they allow for greater speech to be produced and assessed. On the other hand, the same participant felt that interaction was not emphasized in the VSTEP. This concern needs to be addressed by exposing learners to a wide range of communicative situations and engaging them in various test tasks. Thus, English lecturers can familiarize learners with information processing in retrieving necessary core linguistic knowledge to solve the tasks. Examples of test tasks include discussing a situation, role plays, talking about a past event, solving a problem, and other real-life communicative tasks. Also, including various speech functions such as comparing, describing, expressing opinions, and persuading can increase the task difficulty, differentiate proficiency levels. Interestingly, however, there is mention of discourse management as a new criterion in VSTEP speaking assessment, which can clearly and eventually lead to a greater focus on interaction. It is also worth noting that including different speaking test tasks can help balance out the advantages and disadvantages of each task, and test designers might have considered this by including paired tasks when composing VSTEP speaking tests.

Finally, specialized training for the test examiners must be continually provided as, without it, test assessors may find it hard to assess students objectively. Speaking assessment is very demanding on the test examiners, especially when the test examiner holds dual roles as a grader and an interlocutor because this adds to the cognitive load they face. Thus, examiners may skip or ignore the details when the scoring criteria are extensive and switch to global and sometimes subjective assessments. Despite being trained, rater drift and rater variability cause concern as the effect of training may not last long. In this respect, paired and group tasks should be considered as one-to-one oral interviews have been criticized for failing to evaluate all aspects of oral proficiency (Ockey, 2018). Besides, it should be emphasized that there is no best practice for speaking assessment practices because different tasks with various difficulty levels are designed to suit diverse purposes ranging from personal, easy, concrete to non-personal, difficult, and abstract topics. Suppose the VSTEP speaking test is comparable to other international tests. In that case, test designers should ensure that test tasks follow a justifiable order of difficulty comparable to international frameworks (e.g., CEFR) and include a wide range of tasks and updated topics. Also, the predictive validity of

VSTEP in terms of language achievement should be examined.

CONCLUSION

Adopting an existing assessment wholesale may be easy, especially if it has been internationally recognized and accepted. However, as is the mantra today, assessment is not necessarily just for the sake of assessment and should also encourage learning. Hence, national and localized examinations such as the VSTEP are not surprisingly slowly becoming a more common occurrence. From the two participants' perspectives, VSTEP has positive features (e.g., localized topics, availability of prompts, and detailed assessment criteria) and drawbacks (e.g., the lack of high interactivity, extensive and overlapping criteria descriptions, and not being completely free from grammatical errors). Nevertheless, the interviewed participants believe that standardization in speaking assessment practices was essential to ensure assessment fairness, equity, and consistency, especially among local educational institutions. To achieve this, though, a balanced view of standardization in assessment practices at the national level should be adopted as various institutions had different training and educational purposes, and learners also had various study intentions. Furthermore, although VSTEP was designed based on an internationally accepted framework, it has yet to receive complete local acceptance, let alone global recognition. Thus, caution should be taken when imposing standardization practices using VSTEP for all local educational institutions.

The research findings provide useful information about the drawbacks and merits of VSTEP and localized speaking assessment practice for English test examiners and administrators. Vietnamese test examiner-cum-English lecturers' positive attitude and critical evaluation on VSTEP play an initiative role in inspiring other EFL countries to create their own localized English proficiency tests which are equivalent to other international standardized English proficiency tests in terms of quality and validity suitable for national or domestic use. However, due to the limited number of participants, not all Vietnamese VSTEP test examiners' views were represented. Therefore, future research should include more VSTEP test examiners from public and private institutions to confirm the findings. We also note the special need for further studies on the VSTEP in terms of methods. For example, recent studies to validate VSTEP speaking tests only used the inter-rater reliability method to determine reliability. Hence, to prove its reliability and validity and consequently gain larger global acceptance, future studies could consider other methods like discourse analysis to confirm test validity and test score reliability. Besides, future studies can also use generalizability theory (G-theory) to validate the test as G-theory allows the researcher to determine relevant facets that are related to the assessment context (Lynch & McNamara, 1998) and their relative effects on test scores (Bachman et al., 1995; Brennan, 1992).

ACKNOWLEDGEMENT

This research is funded by the University of Economics Ho Chi Minh City, Vietnam.

REFERENCES

Ahmadi, A., & Sadeghi, E. (2016). Assessing English language learners' oral performance: A comparison of monologue, interview, and group oral test. *Language Assessment Quarterly*, *13*(4), 341-358. https://doi.org/10.1080/15434303.2016.1236797

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing*, *12*(2), 238-257. https://doi.org/10.1177/026553229501200206

Bijani, H. (2018). Investigating the validity of oral assessment rater training program: A mixed-methods study of raters' perceptions and attitudes before and after training. *Cogent Education, 5*(1), Article 1460901. https://doi.org/10.1080/2331186X.2018.1460901

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement: Issues and Practice*, *11*(4), 27-34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing*, *26*(3), 341-366. https://doi.org/10.1177/0265532209104666

Brown, H. D. (2000). *Principles of language learning and teaching*. Addison Wesley Longman Inc.

Brown, H. D., & Abeywickrama, P. (2010). *Language assessment: Principles and classroom practices* (Vol. 10). Pearson Education.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1-47. https://doi.org/10.1093/applin/1.1.1

Chalhoub-Deville, M., & Wigglesworth, G. (2005). Rater judgment and English language speaking proficiency. *World Englishes*, *24*(3), 383-391. https://doi.org/10.1111/j.0083-2919.2005.00419.x

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative*. Prentice-Hall.

Creswell, J. W. (2007). *Qualitative inquiry and research design: Choosing among five approaches*. Sage Publications.

Fan, J., & Ji, P. (2014). Test candidates' attitudes and their test performance: The case of the Fudan English Test. *University of Sydney Papers in TESOL*, *9*, 1-35.

Fan, J., & Yan, X. (2020). Assessing speaking proficiency: A narrative review of speaking assessment research within the argument-based validation framework. *Frontiers in Psychology*, *11*, Article 330. https://doi.org/10.3389/fpsyg.2020.00330

Ginther, A. (2020). Assessment of speaking. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1-8). Wiley. https://doi.org/10.1002/9781405198431.wbeal0052.pub2

Hoang, V. V. (2010). The current situation and issues of the teaching of English in Vietnam. 立命館言語文化研究 [*Ritsumeikan Language and Culture Studies*], *22*(1), 7–18.

Kang, O., & Wang, L. (2014). Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Research Notes, 57*, 40-49.

Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing, 36*(4), 481-504. https://doi.org/10.1177/0265532219849522

Le, V. C. (2017). English language education in Vietnamese universities: National benchmarking in practice. In E. S. Park & B. Spolsky (Eds.),

English education at the tertiary level in Asia: From theory to practice (pp. 283–292). Routledge. https://doi.org/10.4324/9781315391588-11

Liontas, J. I., & Siegel, M. (2019). Cultural perspectives in teaching speaking. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-8). American Cancer Society. https://doi.org/10.1002/9781118784235.eelt0696

Liu, L., & Jia, G. (2017). Looking beyond Scores: Validating a CEFR-based university speaking assessment in Mainland China. *Language Testing in Asia*, *7*, Article 2. https://doi.org/10.1186/s40468-017-0034-3

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing*, *12*(1), 54-71. https://doi.org/10.1177/026553229501200104

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, *15*(2), 158-180. https://doi.org/10.1177/026553229801500202

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice & language assessment*. Oxford University Press.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

MOET. (2015). *Decision No. 729/QD-BGDDT: Approving VSTEP.3-5 format based on the Vietnamese version of the Common European Framework of Reference for Languages*. Thukyluat.vn. https://thukyluat.vn/vb/quyet-dinh-729-qd-bgddt-de-thi-danh-gia-nang-luc-su-dung-tieng-anh-tu-bac-3-den-bac-5-416b4.html

Namaziandost, E. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: A comparative study. *Applied Linguistics Research Journal*, *3*(2), 70-82. https://doi.org/10.14744/alrj.2019.83792

Namaziandost, E., Banari, R., & Momtaz, S. (2019). Evaluating oral proficiency skill through analytics and holistic ways of scoring. *Humanities & Social Sciences Reviews*, *7*(5), 424-433. https://doi.org/10.18510/hssr.2019.7547

Nguyen, A. T. (2015). Towards an examiners training model for standardized oral assessment qualities in Vietnam. *Malaysian Journal of ELT Research*, *11*(1), 41-51.

Nguyen, T. (2017, March 26-29). *Vietnam's National Foreign Language 2020 Project after 9 years: A difficult stage* [Paper presentation]. The Asian Conference on Education & International Development. Japan.

Nguyen, T. N. Q. (2018). A study on the validity of Vstep writing tests for the sake of regional and international integration. *VNU Journal of Foreign Studies*, *34*(4), 115-128. https://doi.org/10.25073/2525-2445/vnufs.4285

Nguyen, T. N. Q. (2019). Vietnamese standardized test of English proficiency: A panorama. In L. I. Su, C. J. Weir, & J. R. W. Wu (Eds.), *English language proficiency testing in Asia: A new paradigm bridging global and local contexts* (pp. 71–100). Routledge. https://doi.org/10.4324/9781351254021-4

Nguyen, T. N. Q., Nguyen, T. Q. Y., Tran, T. T. H., Nguyen, T. P. T., Bui, T. S., Nguyen, T. C., & Nguyen, Q. H. (2020). The effectiveness of Vstep.3-5 speaking rater training. *VNU Journal of Foreign Studies*, *36*(4), 99-112. https://doi.org/10.25073/2525-2445/vnufs.4577

Nguyen, T. P. T. (2018). An investigation into the content validity of a Vietnamese standardized test of English proficiency (Vstep.3-5) reading test. *VNU Journal of Foreign Studies*, *34*(4), 129-142. https://doi.org/10.25073/2525-2445/vnufs.4286

Norton, J. (2005). The paired format in the Cambridge speaking tests. *ELT Journal*, *59*(4), 287-297. https://doi.org/10.1093/elt/cci057

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295. https://doi.org/10.1191/0265532202lt205oa

Ockey, G. J. (2018). Oral language proficiency tests. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-5). American Cancer Society. https://doi.org/10.1002/9781118784235.eelt0234

Pham, T. N., & Bui, L. T. P. (2019). An exploration of students' voices on the English graduation benchmark policy across Northern, Central and Southern Vietnam. *Language Testing in Asia*, 9, 15. https://doi.org/10.1186/s40468-019-0091-x

Plough, I. C. (2018). Speaking assessment for high-stakes testing. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1-6). American Cancer Society. https://doi.org/10.1002/9781118784235.eelt0235

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, *65*(3), 395-412. https://doi.org/10.3138/cmlr.65.3.395

Seifoori, Z., & Vahidi, Z. (2012). The impact of fluency strategy training on Iranian EFL learners' speech under online planning conditions. *Language Awareness*, *21*(1-2), 101-112. https://doi.org/10.1080/09658416.2011.639894

The Prime Minister of the Socialist Republic of Vietnam (2008). *Decision No 1400/QD-TTg*. https://vanbanphapluat.co/1400-qd-ttg

Thomas, D. R. (2006). A general inductive approach for analyzing qualitative evaluation data. *American Journal of Evaluation*, *27*(2), 237-246. https://doi.org/10.1177/1098214005283748

Vu, T. P. A. (2016, October 13-14). *25 years of language assessment in Vietnam: Looking back and looking forward* [Paper presentation]. International Conference on English Language Assessment. Hanoi, Vietnam.

Walkinshaw, I., & Duong, O. T. H. (2012). Native- and non-native speaking English teachers in Vietnam: Weighing the Benefits. *TESL-EJ*, *16*(3).

Winke, P. (2013). The effectiveness of interactive group orals for placement testing. In K. McDonough & A. Mackey (Eds.), *Second language interaction in diverse educational contexts* (pp. 247-268). John Benjamins.

Wu, J. R. W. (2012). GEPT and English language teaching and testing in Taiwan. *Language Assessment Quarterly*, *9*(1), 11-25. https://doi.org/10.1080/15434303.2011.553251

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). *Assessment in Education: Principles, Policy & Practice*, *21*(3), 306-325. https://doi.org/10.1080/0969594X.2013.845547

APPENDICES

Appendix A

Semi-structured interview questions

1. How many years have you been working as an English instructor/ speaking examiner?
2. Have you ever taken VSTEP before?
3. Have you ever participated in any speaking assessment training programs?
   If yes, can you share with me your experiences of participating in those programs?
4. What speaking assessment frameworks are you familiar with?
5. What do you think about the VSTEP test and assessment criteria? How would you compare VSTEP with other tests and assessment frameworks such as CEFR or IELTS (e.g., marking criteria and test components)? What difficulties and benefits do you think test examiners may have when using localized VSTEP rating rubrics?
6. What aspects of VSTEP speaking assessment do you think need changing?
7. What do you think about standardization in speaking assessment practices for all the Vietnamese educational institutions?
8. Do you have any other suggestions to improve the current speaking assessment practices in our country?

Appendix B

A sample VSTEP speaking test

Part 1: Social interaction (3 minutes)

Let's talk about your free time activities.

- What do you often do in your free time?
- Do you watch TV? If no, why not? If yes, which TV channel do you like best? Why?
- Do you read books? If no, why not? If yes, what kinds of books do you like best? Why?

Let's talk about your neighborhood.

- Can you tell me something about your neighborhood?
- What do you like most about it?
- Do you plan to live there for a long time? Why/why not?

Part 2: Solution discussion (4 minutes)

Situation: A group of people is planning a trip from Da Nang to Hanoi. Three means of transport are suggested by train, by plane, and by coach. Which means of transport do you think is the best choice?

Part 3: Topic development (5 minutes)

Topic: Reading habits should be encouraged among teenagers