LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 16, No: 2, July – December 2023

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Test-Takers' Performances on and Perceptions of Two Different Modes of Online Speaking Tests

Wiramon Sangsuwan[a,*], Anchana Rukthong[b]

[a] 6311120004@psu.ac.th, Faculty of Arts, Prince of Songkla University, Thailand
[b] anchana.r@psu.ac.th, Faculty of Arts, Prince of Songkla University, Thailand
[*] Corresponding author, 6311120004@psu.ac.th

APA Citation:
Sangsuwan, W., & Rukthong, A. (2023). Test-takers' performances on and perceptions of two different modes of online Speaking tests. *LEARN Journal: Language Education and Acquisition Research Network*, *16*(2), 168-183.

Received
22/11/2022

Received in revised form
12/03/2023

Accepted
05/04/2023

ABSTRACT

A direct test of English speaking is important to evaluate what learners can do in real-life situations. However, due to challenges in test administration, especially with a large number of test-takers, a direct speaking test may not be feasible in many contexts and thus indirect tests, such as conversational cloze tests, are mainly used. In response to this problem, this study utilized communication technology to create speaking tests with two different delivery modes: Real-Time Interview with a human interviewer (RTI) and Pre-Recorded Video (PRV). The tests were given to a group of 40 first-year university students to complete, followed by a perception questionnaire and a group interview to collect data about test-takers' perceptions of the tasks. Results showed that the participants performed significantly better on the PRV test tasks and they perceived both tasks positively. The strongest quality of both test tasks, as perceived by the participants, was authenticity. While the RTI tasks were perceived to significantly have more impact and interactiveness than the PRV tasks, the test-takers shared in the interview that they felt more comfortable and less anxious while completing the PRV tasks.

Keywords: English speaking test, performance-based assessment, online test, test-takers' perceptions

Introduction

Assessment of communicative language has received great research attention, and much scholarly discussion has surrounding the use of communicative language tests (Harding, 2014). Speaking is arguably the most difficult skill to evaluate because factors like rating and test administrating may influence the interpretation of test scores (Plough, 2018). One significant challenge for oral assessments, particularly for large-scale tests, is ensuring that the test assesses the construct of oral communication while actually being practical for administration (Bachman et al., 2010; Plough et al., 2018). While it is important to provide tests which are manageable in a setting, the test must be valid (Galaczi & Taylor, 2018).

Continuous validation can provide evidence that a test is assessing its intended construct and that an individual's test performance can be generalized to future non-test settings. This poses a great challenge to large-scale testing where the multiple-choice (MC) format is extensively employed for language assessment. Despite its benefits, one major challenge in assessing speaking with MC questions is that the actual ability to articulate the target language is not performed (Ginther, 2013; Hughes & Reed, 2016). This does not only affect the generalizability of test scores, but could negatively impact classroom teaching and learning (Bachman et al., 2010). Classroom teachers may end up focusing on training students to pass an MC test by proving test-taking strategies rather than encouraging them to participate in communicative language activities.

To assess actual abilities to communicate in the target language, it is important that performance-based assessment is organized. Not only does this approach require learners to use the target language to complete more authentic test tasks, it also helps promote learning, activate, and encourage the use of knowledge and skills required in real-world situations (Qutaishat et al., 2014). Despite their advantages, one limitation of performance-based assessments is they typically consume a significant amount of time for test administration and scoring, and therefore the test administration with a large number of test-takers may not be feasible in several contexts.

To alleviate the limitation, the use of communication technology seems to be an option. As can be observed in the two English standardized tests: the International English Language Testing System (IELTS) and the Test of English as a Foreign Language (TOEFL), computer technology has been used to deliver test items and score test responses. For IELTS, its computer-based test has offered the listening, reading, speaking, and writing test components. Computer-based tests have been increasingly used, especially during the spread of Coronavirus disease (COVID-19), which has caused a sudden shift from traditional on-site classroom teaching and testing to online. A review of related research has shown attempts to explore and assess the quality of online, computer-based tests (CBT) compared to paper-based tests (PBT), and those previous studies have showed conflicting results. While some studies (e.g., see Ebrahimi et al., 2019; Hüseyin & Özturan, 2018) discovered no significant differences between the performances on the CBT and PBT versions, others such as Panjan and Palanukulwong (2016) and Yao (2020), revealed that scores obtained from the CBT and the PBT versions for both general and academic English tests were significantly different.

In addition to the implementation of CBT in language assessment, online assessment has become important due to worldwide demand for high specifications, expanding the uses of assessment and their usefulness in high-stakes testing (Cerezo et al., 2014). However, an online assessment on speaking appears to have received less attention, compared to other skills. A review of the two language testing journals, Language Testing from 2010 to 2020 and Language Assessment Quarterly from 2009 to 2020 (see Han, 2019), suggested that only a few publications were focused on online oral assessment. The main aim of this study was to provide a better understanding of how test-takers perform in speaking tests with two different modes of delivery and how they perceive their usefulness. In this way, the study can provide its audience with guidelines of how computer technology can be used to enable speaking assessment especially in the context where human resources are limited.

Literature Review

Test Usefulness

Regarding test design and development, Bachman and Palmer (1996) as well as Bachman et al. (2010) suggest tests should be measured by six qualities of test usefulness – reliability, construct validity, authenticity, interactiveness, impact, and practicability. Test reliability refers to the consistency of test scores obtained on different occasions, through different means of measurement or when different tests are used (see also Luoma, 2004). Construct validity refers to capability of the tests to measure the construct or abilities that they are intended to measure. According to Bachman and Palmer (1996), the interpretation of test-taker ability is meaningful and appropriate when the test has construct validity. Authenticity is the potential of the test tasks to simulate the characteristics of real-world communication. When the test tasks closely relate to real-world tasks, the language abilities assessed by test tasks are likely to be in congruence with what is required for real-world communication. This allows for the justification of the predictions and generalizations about each test-taker's ability in non-test situations made on the basis of scores obtained in test situations. Interactiveness is described as the use of the individual's communicative competence and strategic and metacognitive competence to complete the test. An interactive test can raise test-takers' curiosity and interest in test and connect their language ability to real-world language use (Weigle, 2002). Impact, sometimes referred to as test washback, is the influence of tests on classroom teaching and learning and educational management (Weir, 2005). The last aspect of test quality is practicality. The practicality of tests refers to the proportion of resources required for test development and ability to make test organization feasible in practice (Weigle, 2002). In this study, these six qualities were used to guide the test design and investigate whether the designed tests possessed these qualities through the investigation of test-takers' perceptions towards the tests.

Speaking Test

A variety of test formats have been employed to assess spoken interaction, ranging from indirect item types, such as a conversation cloze test, to a direct assessment of speaking abilities, e.g., a face-to-face interview (Underhill, 1987). Indirect speaking tests have been extensively used in language assessment both at international and local levels because of the ease of administration as well as speed and reliability of marking (Bailey & Nunan, 2005). However, the negative impact of such tests on classroom teaching has been pointed out. Instead of focusing on practicing communicative language skills, classroom teachers have been found to train learners to the test (Sundayana et al., 2018). A direct speaking test, on the other hand, has been found to have positive washback on classroom teaching and learning. Allen (2016) found that IELTS, which requires a face-to-face conversation with a human examiner, provided a positive impact on test-takers, as a group of Japanese university students perceived that they had to practice speaking and writing to be better at English. A direct assessment of speaking abilities is, in fact, emphasized and delivered in several international English tests. In addition to IELTS, TOEFL iBT assesses speaking abilities by requiring its test-takers to listen, read, and orally discuss the issues related to listening and reading texts (https://www.ets.org/toefl.html). TOEIC instructs test-takers to orally describe a picture, respond to questions, and expressing opinions in the speaking component (https://www.ets.org/toeic/test-takers/about/speaking-writing.html).

Online assessment of language abilities, involving the application of computer technology, have seen a significant increase in use since 2020, partly because of the effects of the COVID-19 pandemic. Several studies have been conducted in order to explore the reliability and validity of computer-based tests. Dai (2011), for example, compared the effects of face-to-face oral proficiency interviews (OPI), where an examiner interacted orally with a test-taker in a test room, and computer-based oral proficiency interviews (COPI), or a version of OPI conducted through

an online platform. This study reported a high degree of comparability between test platforms as well as the consistency of scores obtained from the raters evaluating OPI and COPI test performances. Although the participants in this study thought that the COPI tasks were less interactive than the OPI tasks, a correlation between the face-to-face speaking test and the computer-based speaking test was found to be at a high level (0.91). In the present study, the COPI version of Dai's (2011) study was replicated.

Test-Takers' Perceptions

One source of research data used for investigating test quality is test-takers' perception (Zhou & Yoshitomi, 2019). Previous studies found such data can provide useful information regarding test validity (e.g., Zhou, 2012), test interactiveness (Sato & Ikeda, 2015), and test difficulty (e.g., Elder et al., 2002). Since test-takers directly experience the test and are affected by the test results, their perceptions help reflect the quality of the test, especially regarding the skills needed for completing the test tasks, providing essential information about test validity (Brooks & Swain, 2015). Previous research, such as Brooks and Swain (2015) and Poonpon (2021), has found test-takers' perceptions reveal construct irrelevances that had gone unnoticed by test developers and may affect the measurement of target constructs and test interface used in online tests. Brooks and Swain (2015) found that issues related to the lack of interaction and immediate feedback and time constraints, reflected by the Test of English as a Foreign Language Internet-based Test (TOEFL iBT) test-takers, impacted test performances.

Regarding speaking tests, previous studies investigated test-takers' perceptions in relation to test validity and the results showed test-takers perceived the construct, content, and predictive validity of technology-based speaking tests positively (Fan, 2014; Zhou, 2012; Zhou & Yoshitomi, 2019). However, a study of the computer delivered TOEIC speaking test indicated some Japanese university students had reservations about the test (Zhou & Yoshitomi, 2019). In comparison to computer-delivered tests for other skills (see e.g., Stricker & Attali, 2010) and face-to-face tests (see e.g., Brooks & Swain, 2015), Japanese students appeared to be more conservative about technology-mediated speaking tests, citing a lack of interaction in the CBT as the primary reason for their negative perception. Noticeably, they had mixed feelings about technology-based speaking tests. The results of Zhou and Yoshitomi (2019), to some extent, indicate that test-takers' perceptions of the test used depend on different factors, including testing context, student demographics and experience. The present study was extended to investigate test-takers' perceptions of the test tasks and for this purpose, a perception questionnaire and a group interview were employed to collect data.

In conclusion, while a direct test of English speaking is important to justify what learners can do in communicative settings while providing valid results, a test of direct speaking is not commonly observed in practice. While there have been previous attempts to investigate the effectiveness of online assessments, the benefits and drawbacks of CBT, and test-taker attitudes towards the test, test-taker performance on and perceptions about the tests are likely task specific. Previous research provided inconsistent results regarding test-takers' performances and their perceptions of the tasks used, especially when compared between the CBT and PBT versions of the tests. With the aim of providing a different means for testing oral communication abilities, this study attempted to create a speaking test which employed different modes of online delivery, Real-Time Interview with a human interviewer (RTI) and Pre-Recorded Videos (PRV), and investigated how test-takers responded to the tests and how they perceived test usefulness. The research questions addressed by this study were:

1. Are there any significant differences in the test-takers' performances on real-time and pre-recorded online speaking test tasks?

2. What are the test-takers' perceptions of real-time and pre-recorded online speaking test tasks?

Methodology

Both quantitative and qualitative data were collected to investigate how test-takers performed on two different types of online speaking tests and how they perceived the usefulness of the tests. The quantitative data were the test-takers' performing scores on the two tests and their responses to a perception questionnaire. Qualitative data included test-taker's responses to open-ended questions on the questionnaire and during a group interview. Detailed information regarding the participants, experimental procedure, and data analyses is described in the following subsections.

Participants

A total of 40 first-year university students (M = 18; F = 22) from the Faculty of Business Administration at a university in Thailand took part in this study. Convenience sampling was used to recruit the participants. At the time of data collection, the participants had previously completed one fundamental English course, which aimed to improve their basic listening, reading, speaking, and writing skills. Their ages ranged from 18 to 22 years old, with an average age of 19.2 years. They had studied English for an average of 12.08 years since primary school. However, only a small number of the participants (13%) had previous experience in taking an English speaking test, such as IELTS, participating in an exchange program interview, or other types of spoken English proficiency assessment.

Instruments

Research Tasks: *Two Speaking Tests with Different Modes of Delivery*

Two parallel English-speaking tests, Real-Time Interview with a human interviewer (RTI) and Pre-Recorded Videos (PRV), were developed to assess test-takers' abilities to use English for communication based on the same construct. The construct was designed in accordance with the Common European Framework of Reference for Languages (CEFR) levels A1-B1. This framework has been used in Thailand to guide language teaching and assessment since 2015, and is used widely across the world. Each test consisted of three speaking tasks, which were: Task 1: Making a conversation about yourself, Task 2: Listening to a text and sharing your ideas about it, and Task 3: Describing and discussing pictures. The first task of the test was aimed at assessing the ability to use familiar everyday expressions and provide personal information. The second task was integrated with the listening prompt to elicit different language functions, including expressing an opinion and agreeing, or disagreeing. The final task was designed to assess test-takers' ability to describe, compare, and contrast two concepts and justify their opinions (see Table 1).

Table 1

Components of the Speaking Tests

Tasks	Language Construct	Topic	Level	Time Allocated (minutes)
1	• Giving personal information	Name/Hometown/ Family/Hobby/Interest	A1	3
2	• Listening and responding to a listening source text • Expressing an opinion, showing agreement or disagreement	Study/Health	A2	4
3	• Describing pictures • Comparing and contrasting two ideas • Justifying opinions	Online vs on-site class/ Online vs in-store shopping	B1	4

Delivery Modes of the Speaking Tests

The speaking tests used in this study were delivered online to the participants using two different modes: pre-recorded video and a real-time mode via ZOOM, a reliable cloud platform for video and audio-conferencing. To investigate the extent to which the test tasks measured what they aimed to measure, test validation was completed prior to data collection process. A panel of expert was set up, consisting of three experienced language teachers and researchers, and the experts were invited to rate different tasks of the tests by using the Item-Objective Congruence (IOC) form with three different rating scales: congruent = +1, questionable = 0, and incongruent = -1. The results showed that the underlying constructs of the test in the RTI were all scored at 1.00, showing that the test items captured their intended construct. For the PRV test tasks, however, the results showed that the underlying construct of Task 3 was questionable, so this task was revised by rewording its instructions and turned in the planned conversation, so that the task was able to elicit the use of language functions aimed at by the test. Also, issues concerning sound quality and the test interface of the PRV task were improved according to the comments. After validation and modification, the tests were piloted with five students whose backgrounds were comparable to those of the participants in the main study to ensure the tasks would test the desired constructs.

Perception Questionnaire

A questionnaire was developed to investigate the test-takers' perceptions of the test tasks. The questionnaire was first developed in English and then translated into the participants' first language to ensure the participants understood the items accurately. For its validity, the validation process carried out for the test items was followed. The questionnaire had three main sections. The first was for the participants to provide their demographic information, i.e., age, education, gender, English learning experience, and previous online testing experience. The second section contained 15 items with a five-point Likert scale ranging from strongly agree (5) to strongly disagree (1). These items were designed to capture the participants' perceptions of test usefulness in relation to six qualities of the useful test described in the Literature Review (Test Usefulness). Since two negative statements were included, reverse score calculation was used with these items (see Appendix 1). The last section of the questionnaire contained three open-ended questions that invited participants to reflect on the strengths and weaknesses of each delivery mode. Cronbach's alpha was used to test the reliability of the questionnaire. The reliability of the PRV test

questionnaire was 0.84 and that of the RTI test was 0.78, suggesting an acceptable level of reliability.

Group Interview

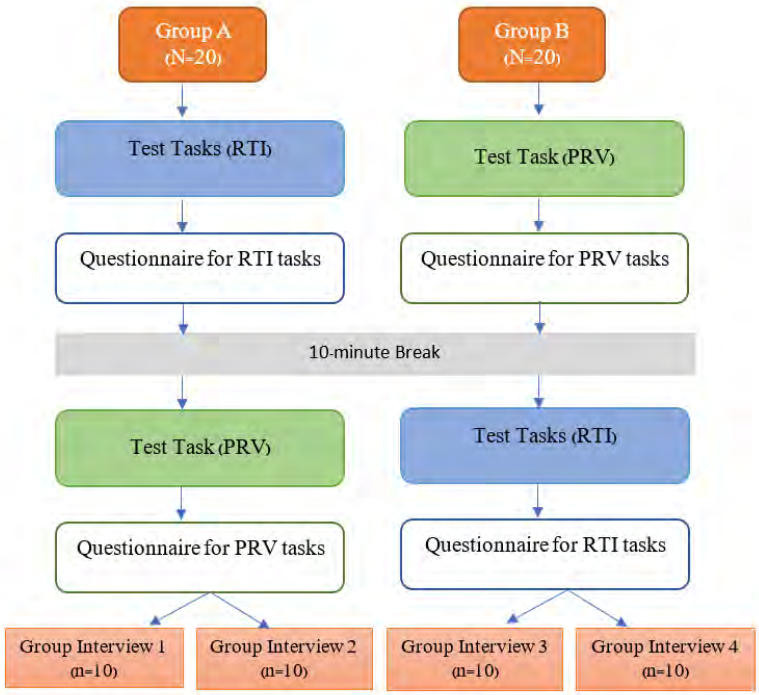
The focus of this interview was to 1) gain detailed insight into test-takers' perceptions of the two testing delivery modes and 2) to gather each test mode's strengths and weaknesses, as reflected by the test-takers. The interview questions built upon open-ended questions given in the questionnaire. Ten test-takers were randomly selected for each group interview session, which was carried out after their test completion.

Data Collection Procedures

After the validation process, the test materials and the perception questionnaire were used to collect data from the group of 40 participants. An online orientation session explained to the participants what participation would involve. A counterbalanced design was used to collect data. The participants were divided at random into two groups, A and B, with 20 participants in each group. As illustrated in Figure 1, the participants were asked to complete both the RTI and the PRV tests, though in a different order. The participants in group A started with the RTI test whereas those in Group B started with the PRV test. The perception questionnaire was delivered after they finished each test. After both tests and questionnaires were completed, the participants participated in a group interview which was moderated by the first researcher. A video was recorded during the data collection process.

Figure 1

Data Collection Procedures



Data Analysis

The speaking tests were scored by two raters: the first researcher and a native English teacher. They scored each task on a scale of 1-3, to measure four aspects of speaking – fluency, accuracy, pronunciation, and lexical resource. The participants were videotaped while taking both versions of the test, so the raters were able to independently evaluate each student's ability based on the video footage. Scores from the two raters were analyzed for level of agreement using Pearson correlation. The average score that each rater assigned for each of the three tasks across both tests were: 4.48/6 (for the PRV) and 4.33/6 (for the RTI) for Rater 1 and 4.40/6 (for the PRV) and 4.53/6 (for the RTI) for Rater 2. The *r* value for the PRV and the RTI scores were 0.92 and 0.88 respectively, suggesting a high degree of consistency between the raters. To arrive at the final score for each participant on each task, an average score from the two raters was used. To see if there were significant differences regarding test performance on the PRV and the RTI tests, a paired sample T-test was employed.

An online questionnaire with the Likert-scale and open-ended questions was given to investigate the participants' perception towards the tests. The descriptive statistics for the responses to the rating scale questions were calculated (means and S.D.). Data obtained from the open-ended questions and group interviews were analyzed using thematic analysis. After the responses were transcribed, themes in the data were identified and relevant information was provided to explain each theme.

Results

This section presents the research results for each research question. First, participant performance on the test tasks is presented. Then, participant perceptions toward the test tasks is examined.

RQ1. Are there any significant differences in the test-takers' performance on real-time and pre-recorded online speaking test tasks?

Overall, the participants performed better on the PRV than the RTI tasks (Table 2). The mean test scores were 18.06/20 for the PRV test and 17.46/20 for the RTI test. With respect to each aspect of test performance – fluency, lexical resource, accuracy, and pronunciation – the results showed participants did well on all aspects, with average scores ranging between 4.19 and 4.69 out of 5.

Table 2

Mean Differences of Test-Takers' Performance on RTI and PRV Test Tasks

Task Types	Aspect	Min.	Max.	Mean	Std. Deviation
RTI	Fluency (5)	3.00	5.00	4.19	0.69
PRV		3.00	5.00	4.31	0.62
RTI	Lexical (5)	3.00	5.00	4.20	0.55
PRV	Resource (5)	3.00	5.00	4.49	0.6
RTI	Accuracy (5)	2.50	5.00	4.36	0.64
PRV		3.00	5.00	4.54	0.63
RTI	Pronunciation (5)	4.00	5.00	4.69	0.43

PRV		4.00	5.00	4.69	0.45
RTI	Total score (20)	14.50	20.00	17.46	1.53
PRV		14.50	20.00	18.06	1.52

Note. RTI = Real Time Interview, PRV = Pre-recorded Video

RQ2. What are the test-takers' perceptions of real-time and pre-recorded online speaking test tasks?

To further investigate whether there were significant differences in participant performance on the RTI and PRV test tasks, a paired sample t-test was performed (see Table 3). The results showed that the overall performances on the two tests were significantly different (t=4.18, $p \leq 0.05$) and the participants performed better on the PRV tasks. Besides, participants were found to perform better on the lexical resource and accuracy aspects, with t = 3.51, $p \leq 0.05$ and t = 2.11, $p \leq 0.05$ respectively. However, the performance regarding fluency and pronunciation were not significantly different.

Table 3

Pairwise Comparison Between Score Obtained from PRV and RTI Test Tasks

Aspects	Tasks	Mean difference	Std. Deviation	t	df	Sig. (2-tailed)
Fluency	PRV- RTI	0.13	0.57	1.38	39	0.18
Lexical resource	PRV- RTI	0.29	0.52	3.51	39	0.00*
Accuracy	PRV- RTI	0.18	0.53	2.11	39	0.04*
Pronunciation	PRV- RTI	0	0.28	0	39	1
Total score	PRV- RTI	0.60	0.95	4.18	39	0.00*

Note. RTI = Real Time Interview, PRV = Pre-recorded Video

To answer the second research question, descriptive statistics were employed to analyze the questionnaire data (see Table 4). The response averages were classified into the following rubric: 4.50-5.00 = strongly agree, 3.50-4.49 = agree, 2.50-3.49 = neutral, 1.50-2.49 = disagree, and 1.00-l.49 = strongly disagree. Hence, participants agreed that both test tasks contain the qualities of test usefulness. Regarding individual aspects, they expressed stronger agreement on the validity and authenticity of the RTI tasks than the PRV tasks. For the other four aspects of test usefulness: reliability, impact, interactiveness, and practicality, the participants agreed that both test tasks have these qualities.

Table 4

Descriptive Statistics of Aspects of Test Usefulness in RTI and PRV Test Tasks

Test Task	Aspect of Test Usefulness	n	Mean	Std. Deviation	Interpretation
PRV	Reliability	40	3.61	0.65	Agree
RTI			3.69	0.61	Agree
PRV	Validity	40	4.44	0.69	Agree
RTI			4.56	0.59	Strongly Agree
PRV	Authenticity	40	4.5	0.74	Agree
RTI			4.65	0.55	Strongly Agree
PRV	Impact	40	4.22	0.65	Agree
RTI			3.59	0.32	Agree

PRV	Interactiveness	40	4.31	0.84	Agree
RTI			4	0.55	Agree
PRV	Practicality	40	3.62	0.58	Agree
RTI			3.59	0.56	Agree
PRV	Total	40	3.89	0.52	Agree
RTI			4.07	0.4	Agree

Note. RTI = Real Time Interview, PRV = Pre-recorded Video

Table 5

Paired Samples Statistics of Aspect of Test Usefulness RTI and PRV Test Tasks

Aspects of Test Usefulness	**Paired Differences**			
	Mean Difference	**Std. Deviation**	**t**	**Sig. (2-tailed)**
Reliability_PRV Reliability_RTI	-0.08	0.81	-0.58	0.56
Validity_PRV Validity_RTI	-0.13	0.78	-1.01	0.32
Authenthicity_PRV Authenthicity_RTI	-0.15	0.86	-1.10	0.28
Impact_PRV Impact_RTI	0.63	0.65	6.12	0.00*
Interactiveness_PRV Interactiveness_RTI	0.31	0.93	2.09	0.04*
Practicality_PRV Practicality_RTI	0.03	0.73	0.22	0.83
Overall_PRV Overall _RTI	-0.18	0.57	-1.94	0.06

Note. RTI = Real Time Interview, PRV = Pre-recorded Video

Regarding whether there were significant differences between the perceptions of the RTI and PRV test tasks, no significant differences in the overall perception of the two test versions were found (see Table 5). However, when focusing on individual aspects, significant differences were found for the perceptions of impact and interactiveness. The participants perceived the RTI tasks as having more impact and interactiveness than the PRV tasks.

The qualitative data obtained from the group interview also support that test-takers perceived both test versions in a positive way. The analysis of the interview data showed that 80% indicated they were more comfortable and less anxious when performing the PRV tasks. For example, some students responded that:

> *"sitting with the computer, I felt like I had more time to think about the answer."*
> *"I felt quite comfortable as no one was watching me."*

Some participants found the PRV tasks more challenging than the RTI tasks because they could not request repetition or clarification when they did not understand either the questions or instructions. They pointed out the limitations of the PRV tasks, as some responded that:

"I do not like the PRV task because I think this is not a natural way of communication. I felt like I was talking to a robot.";
"I do not quite like the PRV task. I think it lacks a key component of humanization of genuine interaction. It was like one-way communication."
"*I think I had to pay rapt attention to the task and my listening as I could not probe for a repetition on the questions.*"

This seems to confirm Weir's view (2005) that attributes of normal conversation, such as ability in accommodating a real-time interaction, were lacking in the PRV test.

Regarding the RTI test, generally over 90% of the test-takers favored this test version because 1) the participants preferred an authentic conversation in real time with an interlocutor and 2) the flexibility in communication which such an interaction facilitated. For example, some indicated that:

"Doing the RTI task is like participating in a real-world conversation. I prefer to have two-way communication like this."
"RTI task is a lifelike situation. Immediate responses and facial expressions made it so real that I feel like I am talking to a native speaker in person."

Regarding the RTI test, the test-takers indicated they could see their own limitations and felt encouraged to improve their language skills. Nearly a half of the participants reflected they were influenced by performance-based assessment. They felt a certain level of accomplishment and were motivated to improve their English language skills. Some participants responded that:

"This speaking test provided me a room to speak English one-on-one with a native speaker."
"I believe I have gained confidence in speaking with native speakers. I had not had much of a chance before."
"I will practice more so that I can speak English more fluently and confidently."

Only a small number of participants were found to negatively perceive the RTI tasks because they did not like the nerve-racking experience of completing the test. They revealed that:

"I don't like it that much because I am too afraid to speak out. I was nervous."
"I think I was not confident with this test and that badly affected my performance."

With these negative comments, test-takers admitted that the examiner's character contributed to their performance. Participants thought the examiner in this study looked friendly and kind –welcoming them to talk. Hence, they felt encouraged to talk and participate more in the conversation. For example, some participants expressed that:

"I feel at ease because the examiner is kind and friendly."
"He is nice and kind."
"It is good that he is helpful when I asked for some repetitions."

In addition to the comments specific to each version of the test, some participants made general comments on the provision of a direct test of speaking. Overall, the participants agreed with the use of a direct speaking test. They agreed the test will motivate them to engage more in speaking activities outside of a test setting.

"The speaking test allows me to know my language ability. It is useful and I will keep practicing."
"I believe I have gained confidence in speaking with native speakers. I hadn't had much of a chance."

"Although I was very excited and my hands were shaking, I found it a very good way of improving my speaking ability."

In short, test-takers had a positive perception towards both tests. The test-takers agreed the speaking tests were beneficial, accurately measured what was intended, and the test tasks were engaging. While they agreed the tests possess all six characteristics of test usefulness, they thought the RTI was more authentic and valid compared to the PRV test. Only a small number of participants expressed negative opinions about test anxiety regarding the RTI test and the inability to accommodate communicative interaction during the PRV test tasks.

Discussion

This study sought to see if different modes of test task delivery would affect participant speaking performance. Generally, test-takers were highly successful on both tests, although they performed significantly better on the PRV test. Considering the individual elements of task performance, the participants scored significantly higher on lexical resource and accuracy for the PRV test, while fluency and pronunciation did not show significant performance differences. This suggests that the modality used to deliver the tests (i.e., a human interviewer vs. a pre-recorded video of an interviewer) affects task performance to some extent. This finding is in line with Yao (2020), a previous comparative study on test delivery modes which suggested test-takers performed differently when different modes of test delivery were used. However, not many studies have discussed mode effects on direct speaking assessment in real-time and pre-recorded responses because most of them focused primarily on assessments of English vocabulary, listening, and reading comprehension.

Regarding test-taker perceptions of the tests, the results show test-takers agreed on the usefulness of the test tasks, although they expressed stronger agreement on the impact and interactiveness of the RTI tasks than the PRV tasks. During the RTI tasks, as shared by the participants in the group interviews, they felt they had an authentic interaction because they were asked to talk to a human being. In contrast, the PRV tasks made them feel like they were talking to a machine. It seems possible to say that a lack of interaction and naturalness contributed to test-takers' unfavorable perceptions of the PRV tasks. This result is supported by recent empirical evidence found by Zhou and Yoshitomi (2019) and other qualitative studies (Brooks & Swain, 2015; Fan, 2014). Additionally, it seems the characteristics of the real-time interviewer added to the positive perceptions of the participants. Regarding the present study, the participants perceived the interviewer as nice and friendly, and so they were encouraged to participate in the conversation. However, one commonly shared reason the participants cited for disfavoring the RTI tasks is test anxiety, since having a real-time interview with a native interlocutor has been reported as anxiety-provoking. This assumption is in line with previous studies exploring student English-speaking anxiety levels (Behforouz et al., 2022). Those studies discovered students were less anxious in an online speaking environment than in a traditional classroom setting. Based on this idea, it is possible to suggest that the main reason for test-takers to feel that the RTI test tasks are more useful that the PRV test tasks is the impact and interactiveness of the task, whereas the other qualities, i.e., reliability, validity, authenticity, and practicality, are not at issue.

Conclusion

This study aimed to investigate the effectiveness of the RTI and the PRV test tasks. To achieve this aim, 40 first year university students participated in tests of English speaking and their perceptions of the tests were gathered through questionnaire responses and group interviews. Participant responses on both test tasks were scored for fluency, accuracy, lexical resource, and pronunciation. The scores were then analyzed and compared to see whether the participants

performed differently across the tests. The results showed that participants performed significantly better on lexical resource and accuracy during the PRV test and their overall performance on the PRV was significantly better than that on the RTI. Besides, the analysis of the perception questionnaire and group-interview data showed that participants thought the two test types were very useful. However, when looking into the perceptions of each individual quality of test usefulness, the results showed the RTI test was superior in terms of impact and interactiveness, while the other qualities (reliability, construct validity, authenticity, and practicality) were not differently perceived.

Implications

A key issue for assessing English speaking abilities identified by several scholars (Brown & Abeywickrama, 2010; Ginther, 2013) is that its construct, including interactional competence, is complicated and not easily tapped into by commonly used item types, like multiple choice (MC). So, it is important to look for alternatives for speaking tests. The results of this study suggest both the RTI and the PRV test tasks provided advantages for language assessment and should be implemented in actual testing practice. The RTI and the PRV tasks allowed test-takers to perform actual spoken responses. Both types of tasks tapped into test-takers' speaking abilities with some significant differences in lexical resource, accuracy and overall scores. While the RTI test appeared to be favored by participants due to its greater authenticity and naturalness compared to the PRV test, participants did not score as highly on the RTI test as they did on the PRV test. One possible reason for this evidenced in the qualitative data is that test-takers felt more engaged during the PRV than the RTI test. The majority of participants were not familiar with a direct speaking test and were too nervous to express their ideas in front of a native interviewer.

The study has shed light on the limitations of both versions of the speaking test. While the RTI test required more time to complete because the interviewer had to conduct interviews on a one-on-one basis, the PRV test seemed less authentic to the test-takers than the RTI test, although it was more inviting. However, the PRV test appeared to take less time to administer because it was delivered online to all test-takers simultaneously. Hence, this study suggests including the PRV test task as an item type in assessments of English for communicative purposes. This would enable a test development team to observe test-takers' actual abilities and limitations in speaking.

The ability to articulate the target language in a given situation is important for language learning and development and the absence of this has shown detrimental effects on the development of speaking skills (Hughes, 2010). Although the RTI and the PRV test tasks in this study were designed for assessing speaking, the tasks can be used as classroom activities as the findings of this study show that students positively perceived both versions of performance-based speaking tests. They also agreed that RTI and PRV tasks are essential in exposing them to actual English speaking. Interesting, most of the students who scored above the mean reflected that they prefer the RTI test due to it being an authentic setting in which they may engage an interlocutor in meaningful communication. On the other hand, lower-scoring students felt significant anxiety about the real-time speaking test, especially regarding interacting with a native speaker. So, it is recommended classroom teachers employ PRV and RTI tasks, according to the resources available and student language proficiency, to increase the chances for students to engage in communicative language activities – especially where English is mainly used in the classroom.

Limitations and Recommendations for Further Studies

Although this study was carefully designed, some limitations related to the method of data collection and the sample size need to be pointed out. From the outset, this study set out to examine how students respond to traditional in-person speaking tests to compare to a virtual version. Having online and onsite test delivery modes may reveal a wider range of effects on

student performance and perceptions. However, due to COVID-19, several restrictions on in-person meetings, direct contact with people, and site visits were imposed. Hence, the method of data collection was adapted to fit social distancing requirements. Another possible set of limitations arise from the small sample size employed and homogeneity of the participants. It is important to note that these findings do not lend themselves to broad generalization due to the modest sample size and specific group of students in one field of study. As the study showed positive participant perception of both performance-based speaking test tasks, a comparative study of test-taker performance in direct in-person interview tests and virtual real-time interviews, which employs a diverse sample of participants would be helpful to determine the optimal method for assessing each language context.

Acknowledgements

This study is funded by the National Research Council of Thailand (NRCT)

About the Authors

Wiramon Sangsuwan: A graduate student of Teaching English as an International Language at Prince of Songkla University, Hat Yai, Thailand. Her research interest encompasses second language assessment.

Anchana Rukthong: An assistant professor at Faculty of Liberal Arts, Prince of Songkla University, Hat Yai, Thailand. Her main research interests are language education, psycholinguistics, second language and foreign language assessment, focusing specifically on assessment of integrated language skills.

References

Allen, D. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language testing in Asia*, *6*(1), 1-20.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford University Press.

Bachman, L. F., Palmer, A. S., & Palmer, A. S. (2010*). Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford University Press.

Bailey, K. M., & Nunan, D. (2005). *Practical English language teaching: speaking.* McGraw-Hill.

Behforouz, B., Gallema, M. C., Waga, R. M. A., & Al Weshahi, S. (2022). The *Journal of Asia TEFL. 19*(2), 469-488.

Brooks, L., & Swain, M. (2015). Students' voices: The challenge of measuring speaking for academic contexts. In B. Spolsky, O. Inbar, & M. Tannenbaum (Eds.), *Challenges for language education and policy*: Making space for people (pp. 65–80). Routledge.

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

Cerezo, L., Baralt, M., Suh, B. R., & Leow, R. P. (2014). Does the medium really matter in L2 development? The validity of CALL research designs. *Computer Assisted Language Learning*, *27*(4), 294-310.

Dai, Z. (2011). A study of the reliability of computerized oral proficiency interview. *Computer-Assisted Foreign Language Education*, *33*(2), 45-50.

Ebrahimi, M. R., Toroujeni, S. M. H., & Shahbazi, V. (2019). Score equivalence, gender difference, and testing mode preference in a comparative study between computer-based testing and

paper-based testing. *International Journal of Emerging Technologies in Learning (Online)*, *14*(7), 128.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?. *Language Testing*, *19*(4), 347-368.

Fan, J. (2014). Chinese test-takers' attitudes towards the Versant English Test: A mixed-methods approach. *Language Testing in Asia*, *4*(1), 1-17.

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly*, *15*(3), 219-236.

Ginther, A. (2013). Assessment of speaking. *The encyclopedia of applied linguistics*, *1*.

Han, C. (2019). A generalizability theory study of optimal measurement design for a summative assessment of English/Chinese consecutive interpreting. *Language Testing*, *36*(3), 419-438.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language assessment quarterly*, *11*(2), 186-197.

Hughes, R. (2010). Materials to develop the speaking skill. *English language teaching materials: Theory and practice*, 207-224.

Hughes, R., & Reed, B. S. (2016). *Teaching and researching speaking*. Taylor & Francis.

Hüseyin, Ö. Z., & Özturan, T. (2018). Computer-based and paper-based testing: Does the test administration mode influence the reliability and validity of achievement tests? *Journal of Language and Linguistic Studies*, *14*(1), 67-85.

Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.

Panjan, S., & Palanukulwong, T. (2016). Thai Learners' Performance on Listening Test: A Comparison of Paper-based and Web-based Testing. *Veridian E-Journal, Silpakorn University (Humanities, Social Sciences and arts)*, *9*(5), 245-257.

Plough, I. (2018). Revisiting the speaking construct: The question of interactional competence. *Language Testing*, *35*(3), 325-329.

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427-445.

Poonpon, K. (2021). Test-takers' perceptions of design and implementation of an online language testing system at a Thai university during the COVID-19 pandemic. *PASAA: Journal of Language Teaching and Learning in Thailand*, *62*, 1-28.

Qutaishat, R. S., Bataineh, A. M., & Bataineh, A. M. (2014). The effect of performance-based assessment on language accuracy of tenth grade English language students at Mafraq Borough directorate of education. *Journal of Education and Practice*, *5*(15), 97-105.

Sato, T., & Ikeda, N. (2015). Test-taker perception of what test items measure: a potential impact of face validity on student learning. *Language Testing in Asia*, *5*, 1-16.

Stricker, L. J., & Attali, Y. (2010). Test takers' attitude about the TOEFL iBT™. *ETS Research Report Series*, *2010*(1), 1-16.

Sundayana, W., Meekaeo, P., Purnawarman, P., & Sukyadi, D. (2018). Washback of English national exams at ninth-grade level in Thailand and Indonesia. *Indonesian Journal of Applied Linguistics*, *8*(1), 167-176.

Underhill, N. (1987). *Testing spoken language: A handbook of oral testing techniques*. Cambridge University Press.

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C. J. (2005). Language testing and validation. *Palgrave McMillan*, *10*, 9780230514577.

Yao, D. (2020). A comparative study of test-takers' performance on computer-based test and paper-based test across different CEFR levels. *English Language Teaching*, *13*(1), 124-133.

Zhao, H., & Gu, X. (2016). China Accreditation Test for Translators and Interpreters (CATTI): Test review based on the language pairing of English and Chinese. *Language Testing*, *33*(3), 439-446.

Zhou, Y. J. (2012). Test-takers' affective reactions to a computer-delivered speaking test and their test performance. *Working papers in corpus-based linguistics and language education*, (9), 295-310.

Zhou, Y., & Yoshitomi, A. (2019). Test-taker perception of and test performance on computer-delivered speaking tests: the mediational role of test-taking motivation. *Language Testing in Asia*, *9*, 1-19.