Linguistic Research 42(Special Edition): 1-27
DOI: 10.17250/khisli.42..202509.001

Assessing L2 speaking in virtual worlds: Validity, mode effects, and implications for group oral proficiency testing*

Jayoung Song
(Pennsylvania State University)

Song, Jayoung. 2025. Assessing L2 speaking in virtual worlds: Validity, mode effects, and implications for group oral proficiency testing. *Linguistic Research* 42(Special Edition): 1-27. The purpose of this study is to explore the potential of virtual worlds (VWs) as a platform for assessing second language (L2) oral proficiency. Specifically, it examines VW group oral tests in comparison to face-to-face (F2F) tests in terms of concurrent validity and face validity. Data were collected from 64 ESL learners' group oral test scores, surveys on their perceptions of both testing modes, observer field notes, and interviews. Results show that while a mode effect was observed in the first administration, students' scores in the VW condition became comparable to those in the F2F condition when the test was repeated. Students' perceptions of the VW mode were also positive in relation to reduced anxiety and ease of use, although turn-taking was reported to be more complex in VW than in F2F tasks. Qualitative analyses further revealed that VWs provide several advantages, including more engaging and less stressful testing conditions as well as a stronger sense of copresence. Taken together, these findings suggest that VW-based assessments can complement traditional testing by offering more flexible and accessible options for evaluating oral proficiency, particularly in distant or online learning contexts. The pedagogical implications include the need to account for interactional challenges such as turn-taking, while also taking advantage of the unique affordances of VWs to create more engaging, scalable, and learner-centered assessment environments. **(Pennsylvania State University)**

Keywords L2 assessment, group oral, virtual worlds, ESL

* This work is based on a part of the author's doctoral dissertation completed at the University of Texas at Austin.

1. Introduction

In recent years, the exponential growth of technology has transformed communication across all domains of life. Professionals in business, medicine, science, and higher education now collaborate online with colleagues across cities and countries. Similar changes are visible in foreign language education, where online classes increasingly provide flexible access to instruction. These developments have heightened the need for oral proficiency assessments that can be administered through technology. Computer-assisted oral assessment offers clear advantages, including accessibility and scalability. However, existing research has noted its limitations in eliciting a full range of language functions and discourse strategies compared to face-to-face (F2F) testing (Shohamy 1994; O'Loughlin 2001). Although F2F speaking tests remain the ideal, their use is constrained by logistical challenges: coordinating examiners and test-takers in the same location, managing large-scale assessments, and addressing the needs of distance-learning contexts. Recent advances in virtual worlds (VWs) such as Second Life present new possibilities for overcoming these barriers. VWs provide real-time interaction among geographically distributed participants, simulating key features of in-person communication while reducing logistical constraints. Yet, despite this potential, little is known about whether VWs are a valid and reliable medium for assessing second language (L2) oral proficiency. The present study addresses this gap by examining the extent to which a VW-based group oral test can serve as a viable alternative to traditional F2F assessment.

2. Literature review

2.1 Computer-mediated assessment of oral proficiency

Technological development has enabled direct, computer-mediated speaking assessments that attempt to capture oral proficiency in real time. One prominent platform is video-conferencing (i.e., Skype, Zoom, Webex, NetMeeting, etc.), which preserves the interactional and co-constructed nature of F2F interviews while allowing test takers and examiners to connect remotely (Choi 2022; Yang 2023). A growing body of research has examined the use of video-conferencing for speaking assessment.

For example, Kim and Craig (2012) reported that classroom-based oral interviews delivered via Skype were comparable to F2F interviews in terms of accessibility, comfort, and test scores. Similarly, Nakatsuhara et al. (2017) found score comparability between video-conferencing (i.e., Zoom) and F2F IELTS speaking tests. More recently, Wang (2024) argued that combining audio and visual channels in video- conferencing (i.e., NetMeeting) can approximate the immediacy of F2F communication while fostering collaborative learning and oral competence.

Despite these promising findings, several concerns remain. First, the technical reliability of video-conferencing platforms is not guaranteed, For example, Davis et al. (2017) reported problems in all of the 25 Skype sessions that they conducted in China in which 22 sessions dropped video and 5 sessions dropped the call. Disruption in the audio or video can obviously lead to communication interference, thus making it hard for raters to make accurate judgment on test takers' actual L2 oral ability. That is, long pauses or silence might be due to technological glitches rather than the test takers' inability to carry out the conversation. Second, the quality of visual input in video-mediated interaction is limited. Previous studies found that the visual input provided by single -camera video or two-dimensional video is inadequate to support the full range of features found in interactive communication (Groen et al. 2012). Limited visual input could impact test-takers in a variety of ways. For instance, communication could be hampered by an inability to see other test takers' body language (Davis et al. 2017). Third, test security and privacy can be compromised, as video platforms expose participants to potential confidentiality risks in high-stakes testing contexts.

Taken together, research suggests that while video-conferencing enables remote oral assessment, it falls short in replicating the full interactional richness of F2F tests. This limitation points to the need for exploring alternative technological environments that may better support interactive oral assessment. Virtual worlds, in particular, offer multi-dimensional environments where participants can interact through avatars, potentially affording richer communicative resources than standard video platforms. However, empirical evidence on the validity and reliability of VW-based assessments remains scarce.

2.2 Possibilities of VWs as a form of direct computer-mediated speaking assessment

Although computer-mediated technologies have been explored as platforms for direct speaking assessments, little research has examined the use of virtual worlds (VWs) such as Second Life for evaluating L2 oral communication (Ockey et al. 2017). VWs provide 3D immersive environments where individuals, represented by avatars, can engage in real-time voice communication from multiple locations. Unlike video-conferencing tools like Skype, which often emphasize distance ("I am here" vs. "you are there"), VWs simulate real-world settings through topography, movement, and physics, fostering a strong sense of copresence and social presence (Garrison 2003). These qualities support authentic interaction, engagement, and collaboration-key components of effective oral communication (Traphagan et al. 2010).

Research supports the feasibility of VWs for oral assessment. For example, Ockey et al. (2017) found that small-group discussions conducted via avatars successfully elicited evidence of interactive oral performance. Participants also reported heightened engagement and social presence. Recent studies extend these findings by demonstrating the potential of immersive environments, including Virtual Reality (VR) and the Metaverse, to create interactive spaces that enhance collaboration, learning, and performance outcomes. These environments enable multimodal communication through avatars, gestures, voice, and text, supporting both remote and in-person learners (Xu and Impagliazzo 2024; Silva et al. 2025).

VWs also offer benefits for test security and reliability. The anonymity afforded by avatars reduces test takers' anxiety and promotes language production by removing concerns about appearance, gender, or social roles (Horwitz 2010; Balcikanli 2012). This anonymity can also limit rater bias, as judgments are less influenced by visual characteristics irrelevant to performance (Shohamy 1989). Studies have shown that avatars' anonymity decreases foreign language anxiety and encourages greater participation and risk-taking (Hammick and Lee 2014).

Another strength of VW-based assessments lies in their authenticity. Test tasks can replicate real-life scenarios more naturally than video-conferencing or traditional face-to-face oral proficiency tests. For example, learners can participate in a restaurant role-play where the VW provides contextual cues-menus, displayed dishes, and

conversational ambience-rather than relying on written prompts. Such immersive environments also enhance motivation by allowing learners to engage in interactive and meaningful contexts beyond the classroom (Liu and Chu 2010; Wehner et al. 2011).

In terms of practicality, VWs are increasingly cost-effective and accessible. Open virtual environments often come equipped with stable voice systems and free, ready-made spaces, reducing development costs (Warburton 2009). Educators can also design custom testing spaces and simulations, ensuring secure and standardized assessment settings.

In sum, VWs might present unique opportunities to address limitations in traditional and video-based oral assessments by combining immersive realism, anonymity-driven reliability, and scalable practicality. As VR and AI technologies advance, they hold the potential to transform language assessment into more authentic, engaging, and equitable experiences for learners.

2.3 Research questions

In the absence of prior studies, this study investigates the feasibility of using virtual worlds (VW), especially Second Life, for L2 speaking assessment by directly comparing it to the traditional face-to-face (F2F) format. Since one key goal is to determine whether the VW-based test measures oral proficiency as accurately as the established F2F mode (Bernstein et al. 2010), both modes were administered to the same participants and their scores were compared. To ensure a fair comparison, the study also examines the role of familiarity training in reducing potential disadvantages associated with using a new virtual environment. By providing participants with training prior to the VW test, the study evaluates whether increased familiarity with the platform minimizes anxiety and narrows score differences between the two testing modes.

Additionally, the present study looks into test-taker experience, perception, and contextual evidence. Given the potential impact of test-takers' perceptions on language performance, it is important to appreciate how different modes of oral assessment may trigger varying emotional responses (Elder et al. 2002; Qian 2009). Assessing test-takers' perceptions can inform test developers of what amendments may be made

to make the test more acceptable or appealing to the test-takers. Such an investigation, as Weir (2005) holds, needs to be based on quantitative and qualitative evidence, thus this study gathered quantitative and qualitative data. There are three research questions that guided this study.

1) Does the VW testing mode yield comparable scores to F2F?
2) What is the impact of familiarity training on minimizing score difference between VW and F2F?
3) What are test takers' perceptions towards VW and F2F testing modes?

3. Methods

3.1 Participants

Participants included students (N = 64) enrolled in an ESL course at a language institution in the southwestern region of the United States. Twenty-two percent (n = 14) of the students had a low advanced level of English proficiency (i.e., C1), while 78% (n = 50) of the students had a high intermediate (B2) level of proficiency based on Common European Framework of Reference (CEFR). Certain concerns arose due to the slightly different proficiency levels of the participants (i.e., high intermediate vs. low advanced). However, based on the findings from a group oral test study that showed that group members' proficiency levels do not affect their group oral scores (Nakatsuhara 2011), proficiency levels were not considered when grouping students for oral tests. Their age ranged from 17 to 25, with a mean of 21. There were 29 female students and 35 male students, and students' majors were varied. Regarding VW experience, approximately 95% of the students (n = 61) had no experience in VWs, while 5% of the students (n = 3) had used VWs for fun or educational purposes. In terms of computer skills, approximately 95% of the students (n = 61) were comfortable using computers while 5% of the students (n = 3) had limited computer skills.

3.2 Data sources

The data was drawn from four sources. This data was analyzed using a convergent,

parallel mixed-methods design (Creswell and Plano Clark 2011), where quantitative and qualitative data were collected, and separately analyzed, after which findings were integrated. The two data source provided different types of information and allowed for an in-depth and comprehensive set of findings.

3.2.1 A group oral task

Construct. Drawing on Fulcher (2003)'s definition on L2 oral communication ability, the overarching construct for the group oral task is defined as the verbal use of language to communicate with others. This general ability includes interactional competence, which is an ability to actively structure appropriate speech in response to incoming information from another interlocutor in real time (Ockey and Li 2015). Given that assessing the interactional competence, a subconstruct of L2 oral ability requires the interaction of a test taker with an interlocutor, the present study incorporated a discussion task in which a small group of test takers carry out a discussion. The topics for the discussion task appear in Appendix 1.

VW group oral. During the VW speaking test, groups of 3 or 4 students were located in two computer rooms. Following a brief introduction about how to use a computer for the test, three or four participants logged into a room in a VW, Second Life, that had been chosen for this project. The space in the VW was a classroom or a library, in which students would be likely to participate in a discussion task in the real world. The students were represented by avatars (Figure 1). Female students were given female avatars, and male students were given male avatars so that avatars are consistent with the perceived identity of themselves (Segovia and Bailenson 2013). If an avatar began to speak, a signal above the avatar's head lit up, indicating who the active speaker was at that moment. As soon as the participants had entered the VW, I ensured that all students' microphones were working properly, that they were using headsets and thus able to hear clearly. Then, I showed the task prompt on the screen and began the test by saying, "You can now start speaking." The VW test was recorded with a screen-capture program called Camtasia.

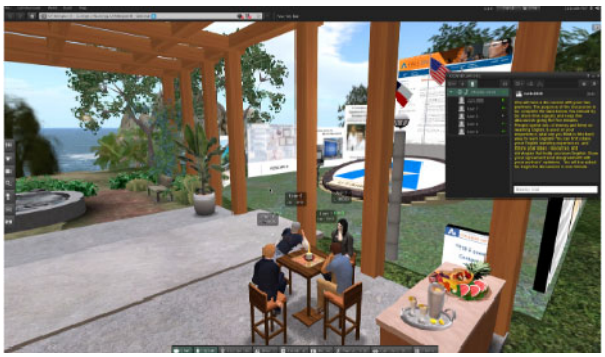

Figure 1. VW group oral

Figure 2. Face-to-face group oral

F2F group oral. The F2F test format was the same as that of the VW test, with the only difference being the location wherein the actual conversation took place (Figure 2). The researcher did not participate in the discussion and sat outside of the group, giving the cue for participants to begin the test by saying, "You can now start speaking." The F2F test was recorded with camcorders.

Scoring rubric. Ockey's (2009) scoring rubric served as a basis and provided ratings on a four-point scale for each of the four oral communication subscales: delivery (pronunciation, intonation, and fluency), language use (use of morphology, complexity of syntax, and range of vocabulary), topic development (sustained response, coherence of speech, and overall topic development) and interactional competence (awareness of other interlocutors, smooth turn-taking, topic initiation, and use of communication strategies). (Appendix 2)

Rater training. All four raters had native-English speaking competence, graduate degrees in ESL teaching or related field, experience in teaching a similar population of students, and experience in rating similar group oral tests. Prior to the rater training, both the researcher and an assistant independently rated four samples of F2F or VW group oral tests and reached a consensus on the scoring. Then, the researcher provided training to the raters with three purposes: (a) familiarizing the raters with the assessment tasks, testing modes, and the rating criteria descriptions; (b) establishing a shared understanding of differentiating test takers' oral communication abilities according to the rating criteria descriptions; and (c) practicing scoring and discussing cases. Raters were asked to conform to the already agreed upon ratings. The video recordings of the group oral test were randomly distributed to the four raters, with two raters scoring the half the test takers (n=32) while the other two raters scoring the remaining half (n=32). The inter-rater reliability of the first two raters was r=.90 and the other two raters was r=.94, which showed high internal consistency.

3.2.2 A questionnaire about perceptions of the two testing modes

A questionnaire consisting of modified questions from Van Moere's model (2006) and Ockey et al.'s study (2017) was distributed to ascertain participants' perceptions of the VW and F2F speaking tests. All the participants completed the questionnaire after each testing period (i.e., F2F and VW). The questionnaire consisted of 14 five-point Likert scale items in four categories: (1) efficacy (α = .58), (2) anxiety (α =. 74), (3) ease of test-taking (α =. 82), and (4) ease of turn-taking (α = .73). The example questions include: (1) Basically, I showed my true level of English conversational ability in this test (efficacy), (2) I was nervous while I was doing the task in this mode (anxiety), (3) I could clearly hear other speakers in this mode (ease of test-taking), (4) It was easy to take turns when I was doing the task (ease of turn-taking). Additionally, there were 6 open-ended questions which asked test takers' preference and opinions about each testing mode.

3.2.3 Interview

The students who had indicated their willingness to participate in the interview were invited to do so. The 30-minute interviews were conducted face to face after they

had completed the second test. The questions were focused on their experience using Second Life for assessing speaking. The interviews were audio-recorded and transcribed.

3.2.4 Observer's field notes

The researcher was present at the testing site and took field notes. I observed test takers' behaviors under both test modes and recorded any particular behavior that the test takers showed in each mode.

3.3 Experimental design

The 64 students were randomly assigned to control group (n=30) or experimental group (n=34) and completed the group oral tasks in both F2F and VR modes over the two days. The control group completed both F2F and VW on a single day, and another set of F2F and VW on the other day. On the other hand, the experimental group took both F2F and VW tests on a single day. Two days later, they received a ninety minutes of VW training, which included watching a short YouTube video clip regarding how to use Second Life, navigating with their avatars in Second Life. After the training, they completed the remaining F2F and VW tests. In order to avoid topic and practice effects, the topics of the test tasks were counterbalanced; half of the participants received one combination of topics (Topic 1 and Topic 2) first while the other half received the other topic combination (Topic 2 and Topic 1) first. Additionally, in order to avoid practice effects, half of the students took a face-to-face test first while others took a virtual-world test first (Table 1 and 2).

Table 1. Pretest topic and mode

Test order	Group A (n=15)	Group B (n=15)	Group C (n=15)	Group D (n=19)
1	Topic 1-F2F	Topic 2-F2F	Topic 1-VR	Topic 2-VR
2	Topic 2-VR	Topic 1-VR	Topic 2-F2F	Topic 1-F2F

Table 2. Posttest topic and mode

Test order	Group A (n=15)	Group B (n=15)	Group C (n=15)	Group D (n=19)
1	Topic 3-F2F	Topic 4-F2F	Topic 3-VR	Topic 4-VR
2	Topic 4-VR	Topic 3-VR	Topic 4-F2F	Topic 3-F2F

3.4 Data analysis

For the first question, the concurrent validity of the VW group oral test in comparison to F2F test was analyzed by means of repeated-measures ANOVA and repeated-measures MANOVA with four analytic categories of the group oral scores as dependent variables. The second research question, impact of familiarity training on test takers' scores in the VW group oral test was analyzed by means of repeated-measures ANOVA. The third research question discusses the face validity of VW group oral tests from test taker's perspective. Test takers' answers to the perception questionnaire were analyzed by means of repeated-measures ANOVA for four categories (i.e., efficacy, anxiety, ease of test-taking, and ease of turn-taking).

In order to answer the third question, all written comments in the perception questionnaire and transcribed interview data were compared across the F2F and VW testing modes. Once all analyses were completed by the research assistant, thematic categories and coded information were presented to another researcher, and the emerged themes and coding accuracy across different data sources were confirmed. The results obtained in the analyses of test-takers' scores, observer's field notes, test takers' perception questionnaire responses, and interviews were triangulated to explore and give detailed insights into how the VW testing mode is different from or similar to F2F mode.

4. Results

4.1 Research question 1

Considering the purpose of the first research question, which was to examine the differences and similarities between the test-takers' group oral scores in both F2F and

VW modes, only their scores at time 1 were compared using repeated-measures ANOVA. A repeated-measures ANOVA with one between-subject factor (group) and one within-subjects factor (testing mode) indicated that the main effect of the testing mode was statistically significant at the .05 level, F (1,62) = 9.03, p = .004, partial eta-squared = .13, with test-takers achieving higher scores when they were tested in F2F mode.

In order to examine in which section differences occurred in test-takers' F2F and VW group oral scores, repeated-measures MANOVA was conducted, with four analytic categories of the group oral scores as dependent variables. As shown in Table 3, differences were observed in all sections of the group oral scores between the two testing modes, with better performances in delivery (D), language use (U), topic development (T), and interactional competence (IC) when they were tested in the F2F mode. A repeated-measures MANOVA revealed a significant multivariate main effect for mode, Wilks' λ = .851, F (4,59) = 2.6, p = .046, partial eta-squared = .149. Given the significance of the overall test, the univariate main effects were examined. Significant univariate main effects for mode were obtained for topic development, F (1, 62) = 7.74, p = .007, partial eta-squared = .111. This result indicated that there were no statistical differences between the test-takers' F2F and VW group oral scores in delivery, language use, and interactional competence sections but they did occur in topic development. Pairwise comparison with the Bonferroni adjustment indicated that the students displayed .20 higher scores in topic development, on average, when they were tested in the F2F mode than in the VW, t (62) = 2.77, p =.007. In conclusion, test-takers performed better overall when they were tested in the F2F than in the VW testing mode, with markedly better performances in terms of topic development in the F2F mode.

Table 3. Mean and standard deviations of test takers' group oral scores in the face-to-face and virtual-world testing mode at time 1

	F2F (D)	VW (D)	F2F (U)	VW (U)	F2F (T)	VW (T)	F2F (IC)	VW (IC)
	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)	M (SD)
Group1	2.62(.62)	2.58(.60)	2.62(.58)	2.50(.57)	3.17(.83)	2.98(.83)	2.78(.83)	2.82(.73)
Group 2	3.04(.79)	3.03(.65)	2.94(.70)	2.88(.67)	3.28(.61)	3.06 (.71)	3.11(.60)	2.80(.65)

Note: N = 30 for group 1 (control), and N = 34 for group 2 (experimental). The possible score for each section was from 0 to 4.

4.2 Research question 2

Given that the students showed better performance in the F2F mode when they were tested with the each mode first time, the purpose of this research question was to investigate if familiarization with the VW environment through training could help decrease the differences in students' scores between the two modes. The descriptive statistics of the test-takers' scores in each testing mode and at each time point are presented in Table 4. The primary finding of interest was that there was a statistically significant time by mode interaction (Wilks' λ = .84, F (1,62) = 11.01, p =.002, partial eta-squared = .151) across the control and experimental groups. This indicated that the mean differences between the F2F and VW modes were statistically different at the first trial. However, when the students were tested again, the differences in their scores on the F2F and VW tests were not statistically significant, t (62) =.58, p =.522.

Table 4. Mean and standard deviations of test takers' group oral scores in the face-to-face and virtual-world testing modes at two time points

Group	1st test		2nd test	
	F2F Mean (SD)	VW Mean (SD)	F2F Mean (SD)	VW Mean (SD)
Group 1 (Control)	11.18 (2.45)	10.88 (2.42)	11.37 (2.60)	11.36 (2.52)
Group 2 (Experimental)	12.38 (2.15)	11.77 (.2.17)	12.34 (2.11)	12.32 (2.18)

Note: N = 30 for group 1 and N = 34 for group 2. The possible score range was from 0 to 16.

Another significant finding was that there was no three-way interaction between group, testing mode, and time. It was hypothesized that the score differences between the two testing modes at two time points would be different for the experimental and control groups; that is, it was expected that students in the experimental group would exhibit improvements in their VW group oral scores after training, making their VW test scores comparable to their F2F scores, while students in the control group would show no improvements in their VW scores at the second trial as they had received no training. The results, however, indicated that ESL students could demonstrate their abilities regardless of testing mode when they had experience in the new testing mode. This suggests that students can easily be familiarized with the VW by simply taking a test in that environment.

4.3 Research question 3

Students' answers to the questionnaire, consisting of four factors (i.e., efficacy, anxiety, ease of test-taking, and ease of turn-taking) were analyzed. For each factor, repeated-measures ANOVA with one between-subjects factor (group) and two within-subjects factors (time, reference to administrations 1 and 2, and testing mode) were conducted.

Efficacy. A one-way repeated-measures ANOVA revealed a significant multivariate main effect for mode (Wilks' λ = .84, F (1,62) = 12.08, p =.001, partial eta-squared = .163), indicating higher self-efficacy in the F2F mode across time. There was also a significant multivariate time by mode interaction (Wilks' λ = .90, F (1,62), p =.011, partial eta-squared = .100). The testing mode contrasts for each time with the Bonferroni adjustment showed that ESL students' efficacy scores were .53 higher, on average, when they were tested via the F2F mode than via the VW mode, t (62) = 4.07, p = .000, at the first trial. However, when the students were tested again, their self-efficacies in the F2F and VW modes were not statistically significant, t (62) = .27, p =.763. This suggests that test-takers believed they could demonstrate their actual competence in speaking whether they were tested via the F2F or the VW mode once they became accustomed to the novelty of the environment.

Anxiety. A one-way repeated-measures ANOVA revealed a significant multivariate main effect for time (Wilks' λ = 2.70, F (1,62) = 4.50, p = .038, partial eta-squared

= .068), indicating higher anxiety at Time 1 across the two testing modes. There was also a significant multivariate time by group interaction (Wilks' λ = 4.72, F (1,62) = 7.88, p =.007, partial eta-squared = .113). The results showed that students' anxiety levels changed at the second trial. Pairwise comparison with the Bonferroni adjustment showed that the anxiety levels of the control group did not change at the second trial while the experimental group's anxiety levels decreased significantly at the second trial (t (62) = 3.6, p =.001) The result lends support to the effectiveness of familiarity with VW, achieved through training, in terms of lowering anxiety.

Ease of test-taking. Students responded that, generally speaking, it was relatively easy to take a group oral test both in the F2F and the VW. The results from a repeated-measures ANOVA showed a significant multivariate main effect for mode (Wilks' λ = .847, F (1,62) = 11.87, p =.001, partial eta-squared = .153), and a significant multivariate time by group interaction (Wilks' λ = .834, F (1,62) = 12.36, p =.001, partial eta-squared = .166). The results showed that, while students initially found it easier to take a group oral test in the F2F mode, their perceptions changed over time, with the experimental group reporting an increase in perceived ease of use in both modes on the second test, t (62) = 3.1, p =.003. As had been the case with the anxiety score at the second trial, the control group students reported no significant increase. The results again indicate the effects of training on students' perceptions of the VW testing mode.

Ease of turn-taking. The results from repeated-measures ANOVA showed a significant multivariate main effect for mode (Wilks' λ = .314, F (1,62) = 28.36, p =.000, partial eta-squared = .314), showing higher scores in the F2F mode across groups. There was no significant two-way interaction or three-way interaction. Unlike other categories in which students' perceptions changed in terms of efficacy, anxiety, or use of the testing mode after the first test, it seemed that their perceptions toward the ease of turn-taking did not change drastically.

Benefits and challenges of the two test modes. Qualitative analyses revealed several themes regarding students' perceptions of the F2F and VW group oral tests. Most students preferred the F2F mode, citing three main advantages: access to nonverbal cues, instant responses, and the natural flow of conversation. About 77% (n=50) favored F2F because they could use body gestures, facial expressions, and eye movements to aid understanding, support opinions, and manage turn-taking. As one participant noted, "I felt more comfortable speaking F2F because I could read faces

and express myself better with gestures and eye contact." Students also appreciated the immediacy of responses during F2F interaction, which helped them feel assured and understood. Additionally, 25% felt conversations were more natural in the F2F mode, as it reflected real-world communication. However, a few students (11%) reported initial anxiety, finding it awkward to speak with unfamiliar people while being directly observed.

In contrast, the VW mode offered unique advantages, particularly in reducing anxiety. About 60% of students reported feeling more relaxed in VW environments due to the anonymity provided by avatars. One student shared, "In F2F, everyone looked at me, but in VW, I felt like I was sitting in my apartment. It was more comfortable." The novelty and interactivity of the VW also appealed to students. Sixty percent described VW testing as more interesting than F2F, enjoying the avatars and immersive settings. Some participants even reported a strong sense of copresence, feeling as though they were in the same space as other test-takers despite being remote: "After the training session, I felt like I was with the others in the same place and could concentrate more." Despite these benefits, VW testing introduced its own challenges. For students unfamiliar with virtual environments or gaming, the new interface initially caused anxiety. One participant admitted, "The VW looked complicated, and I was nervous about taking a test in that environment." However, familiarity training significantly reduced these concerns, as reflected in the quantitative results.

Overall, students perceived F2F testing as more natural and supportive due to visible nonverbal cues and instant responses, while VW testing fostered comfort, engagement, and innovation through anonymity and immersive environments. Yet, VW's unfamiliarity initially posed barriers for some learners, suggesting the need for orientation sessions to maximize its potential.

5. Discussion

The first and second research question aimed to verify the concurrent validity of the VW group oral test. Results showed that the VW testing mode demonstrated concurrent validity when students took the test for the second time. However, when students took the test for the first time, a mode effect was observed, with students

achieving higher scores in the F2F mode. This may be explained by the additional cognitive demands of the VW environment. Robinson (2001) argued that task complexity, consisting of several dimensions, significantly affects language production, and Skehan and Foster (1997) also noted that task conditions shape learners' perceptions of role and status during performance. In this study, the VW mode may have been more cognitively challenging than the F2F mode, which students had repeatedly experienced in classroom contexts. Even though students were not allowed to modify avatar characteristics in order to reduce construct-irrelevant variance (Ockey et al. 2017), they were still required to monitor avatars, manage locations, and adjust audio settings prior to the test. These extra steps may have demanded additional mental resources, which limited the attention available for oral production and affected performance during the first administration.

Further analysis of scoring categories reinforces this interpretation. Differences between the two modes were most evident in the topic development category. Students' delivery scores (e.g., pronunciation and fluency) did not differ significantly across modes, but topic development scores were weaker in the VW. Because topic development is cognitively demanding (Sternberg 1977), adapting to a novel testing environment may have reduced the resources available for elaborating ideas. Thus, the quantitative findings suggest that novelty effects in the VW condition likely contributed to weaker performance in developing topics, even while other categories remained stable.

One interesting finding was that there was no statistically significant difference between the two modes in interactional category. Field notes revealed that students adapted their interactional strategies to the affordances of each mode. In the F2F condition, they relied on nonverbal cues such as eye contact and body language to manage turn-taking, agreement, and disagreement. In the VW, however, they compensated for the absence of visual cues by employing explicit verbal strategies, such as "Can I start?" or "I agree with you." Students also opened turns to the group at large or requested permission to speak, while others verbally confirmed their agreement. These adaptations indicate that although VW environments lack some nonverbal resources, students were able to strategically adjust their interactional behavior to maintain communicative effectiveness.

The second research question specifically focused on whether students' performance in the two modes would converge after repeated testing. Findings showed

that although there was a mode effect during the first administration, no significant differences were observed between F2F and VW scores during the second administration. This suggests that as students became more familiar with the VW environment, their performance stabilized and became comparable to their performance in the F2F mode. This finding aligns with Nakatsuhara et al. (2021), who noted that newly developed test formats can yield valid results when carefully designed and implemented. Similarly, the present study indicates that VW-based oral tests can serve as a viable alternative to F2F tests when test takers are provided with sufficient practice and preparation.

The third research question addressed the face validity of the VW testing mode by examining students' perceptions. Results showed that most students preferred the F2F testing mode, a finding that is consistent with prior studies comparing new test formats to established ones (Kenyon and Malabonga 2001; Qian 2009). In this study, all participants reported having taken F2F proficiency interviews in the past, while 96% had never used a VW environment for learning or entertainment, and none had previously taken a test in a VW. Given this lack of familiarity, students favored the F2F mode because it allowed them to better anticipate test demands and feel more competent. In contrast, the novelty of the VW environment may have raised doubts about its reliability as a testing medium. Despite this preference, an important finding is that many students indicated they would be willing to take a VW test if it offered greater accessibility. In particular, they expressed that they would choose VW testing if it enabled them to take oral proficiency exams from home instead of traveling to a testing center. This finding highlights the potential of VW-based testing to expand access and convenience, even if initial preferences lean toward traditional F2F testing.

Results also showed that perceptions of efficacy, confidence, and ease of use significantly improved after the second administration, suggesting that challenges associated with using VW environments can be reduced with repeated exposure. Students demonstrated that they could perform with comparable levels of confidence in both modes once they had experience in the VW. However, perceptions related to turn-taking did not change after repeated testing. Unlike other variables, which improved with familiarity, turn-taking remained a persistent challenge due to the lack of visual cues in VW environments. This limitation suggests that while VW testing can become a valid and accepted mode with experience, the absence of nonverbal

signals for turn management continues to constrain interaction in ways that differentiate it from F2F testing.

6. Conclusion

The research has some limitations including the relatively small number of students (N=64), convenient sampling, a single type of task (i.e., group discussion), and a sample without varied language proficiency. Studies which incorporate groups of other proficiency levels or different types of tasks may be fruitful areas for future research. There has been a clear need for further studies into the development of a new testing format in which test-takers can synchronously discuss topics using a computer, particularly with regard to overcoming the practical limitations of an F2F group oral test and the limitations of computer-mediated assessments that are performed at the cost of actual human interaction (Ockey 2009; Qian 2009; Newhouse and Cooper 2013; Ockey et al. 2017). The current study responds to this need by examining the VW group oral test's potential, in terms of its reliability, concurrent validity, and face validity.

This study has practical implications for test developers or language practitioners. If test designers are interested in minimizing the effect of the testing mode, it appears appropriate to simplify the VW test as much as possible. A less complex task, in which test takers are simply required to speak with a predetermined avatar in a designated place, may allow them to conserve their mental resources for completing the task. Furthermore, the degree of influence exerted by lack of familiarity with the VW testing mode may be manipulated by providing test takers with a practice session or brief training materials. The study has also two key implications for rater training. Rater training might include guidance on how to measure interactional competence especially in the VW assessment. Raters can observe test takers' body language to make judgments regarding their abilities to manage the conversation and work together cooperatively in the F2F testing mode. In the VW assessment, however, raters are obliged to consider other signals for interactional effectiveness, including explicit cues such as, "I agree with you," and, "Can I start?" Using meaningful descriptors, which guide them in making decisions, raters can decrease the complexity of the decisions they are required to make.

Decisions regarding which of the two tests is "better" or "more appropriate" should be based on several criteria, such as the purpose of the tests and the use of the test results (Shohamy 1994). For example, in situations where testers wish to ensure that the test is a representative sample of real-life conversation, it is necessary to consider what is denoted by "real-life" conversation. Sacks, Schegloff, and Jefferson (1974) and van Lier (1989) defined conversation as an F2F interaction that has not been planned ahead, the outcome and sequence of which is unpredictable. If testers adopt the aforementioned definitions of conversation, F2F group oral testing may be the better assessment mode as it measures F2F interaction. However, in the digital era, it seems that online communications, in which individuals cannot be identified, have become increasingly common mode of communication (Nakatsuhara et al. 2021; Inoue et al. 2024.). Given that younger generations are "digital natives", whose way of communication differs fundamentally from that of their predecessors as a result of their immersion in new technology (Prensky 2001), online communication is becoming the part of the target language use domain. Thus, assessing L2 oral communication ability via online virtual interaction maybe appropriate. While there are many additional factors that guide decision makers, it seems most important to know exactly what a test measures and whether a particular testing mode best assesses how well someone speaks a second language for a given context and purpose.

References

Bachman, Lyle F. 2000. Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing* 17(1): 1-42.

Balcikanli, Cerm. 2012. Language learning in second life: American and Turkish students' experiences. *Turkish Online Journal of Distance Education* 13(2): 131-146.

Bernstein, Jared, Alistair Van Moere, and Jian Cheng. 2010. Validating automated speaking tests. *Language Testing* 27(3): 355-377.

Birjandi, Parviz and Saeed Rezaei. 2010. Developing a multiple-choice discourse completion test of interlanguage pragmatics for Iranian EFL learners. *ILI Language Teaching Journal* 6(Special Issue): 43-58.

Brinkmann, Svend and Steinar Kvale. 2015. *InterViews: Learning the craft of qualitative research interviewing*. Thousand Oaks, CA: Sage Publications.

Chapelle, Carol A. and Dan Douglas. 2006. *Assessing language through computer technology*.

Cambridge: Cambridge University Press.
Choi, Jin Soo. 2022. *Investigating test delivery modes within video-conferenced English speaking proficiency assessment*. Phd Dissertation. Michigan State University.
Clark, John L. and Dariush Hooshmand. 1992. Screen-to-screen testing: An exploratory study of oral proficiency interviewing using video teleconferencing. *System* 20(3): 293-304.
Creswell, John W. and Vicki L. Plano Clark. 2011. *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, CA: Sage Publications.
Davis, Larry, Veronika Timpe-Laughlin, Lin Gu, and Garry Ockey. 2017. Face-to-face speaking assessment in the digital age: Interactive speaking tasks online. In John M. Norris, John McE. Davis, Margaret E. Malone, Todd H. Mckay, Young-A Son (eds.), *Useful assessment and evaluation in language education*. Washington, D.C.: Georgetown University Press.
Ekbatani, Glayol V. 2010. *Measurement and evaluation in post-secondary ESL*. New York: Routledge.
Elder, Catherine, Noriko Iwashita, and Tim McNamar. 2002. Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing* 19(4): 347-368.
Fulcher, Glenn. 2003. *Testing second language speaking*. Harlow: Pearson Education.
Fulcher, Glenn and Fred Davidson. 2007. *Language testing and assessment*. London: Routledge.
Galaczi, Evelina D. 2010. Face-to-face and computer-based assessment of speaking: Challenges and opportunities. In Luisa Araújo (ed.), Computer-based assessment of foreign language speaking skills. 29-51. Luxembourg: Publications Office of the European Union.
Garrison, Randy. 2003. *E-learning in the 21st century: A framework for research and practice*. London: Routledge Falmer.
Groen Martin, Marian Ursu, Spiros Michalakopoulos, Manolis Falelakis, and Epameinondas Gasparis. 2012. Improving video-mediated communication with orchestration. *Computers in Human Behavior* 28(5): 1575-1579.
Hammick, Jin K. and Moon J. Lee. 2014. Do shy people feel less communication apprehension online? The effects of virtual reality on the relationship between personality characteristics and communication outcomes. *Computers in Human Behavior* 33: 302-310.
Horwitz, Elaine K. 2010. Foreign and second language anxiety. *Language Teaching* 43(2): 154-167.
Inoue, Chihiro, Fumiyo Nakatsuhara, Vivien Berry, and Evelina D. Galaczi. 2024. *Video-conferencing speaking tests: An investigation of context validity related to test administration*. Cambridge: Cambridge University Press and Assessment.
Jucker, Andreas H. 2009. Speech act research between armchair, field and laboratory: The case of compliments. *Journal of Pragmatics* 41(8): 1611-1635.
Kenyon, Dorry M. and Valerie Malabonga. 2001. Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology* 5(2): 60-83.
Kim, Juntae and Daniel A. Craig. 2012. Validation of a videoconferenced speaking test. *Computer*

Assisted Language Learning 25(3): 257-275.

Liu, Tsung-Yu and Yu-Ling Chu. 2010. Using ubiquitous games in an English listening and speaking course: Impact on learning outcomes and motivation. *Computers & Education* 55(2): 630-643.

Luoma, Sari. 1997. *Comparability of a tape-mediated and a face-to-face test of speaking: A triangulation study*. Unpublished Licentiate Thesis. University of Jyvaskyla. Available at: http://urn.fi/URN:NBN:fi:jyu-1997698892.

McNamara, Timothy. 2000. Language testing. Oxford: Oxford University Press.

Miles, Matthew B. and Michael Huberman. 1994. *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage Publications.

Nakatsuhara, Fumiyo. 2011. Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing* 29(4): 553-573.

Nakatsuhara, Fumiyo, Chihiro Inoue, Vivien Berry, and Evelina D. Galaczi. 2017. Exploring the use of video-conferencing technology in the assessment of spoken language: A mixed-methods study. *Language Assessment Quarterly* 14(1): 1-18.

Nakatsuhara, Fumiyo, Chihiro Inoue, Vivien Berry, and Evelina D. Galaczi. 2021. Video-conferencing speaking tests: Do they measure the same construct as face-to-face tests? *Assessment in Education: Principles, Policy & Practice* 28(4): 369-388.

Newhouse, Paul and Martin Cooper. 2013. Computer-based oral exams in Italian language studies. *ReCALL* 25(3): 321-339.

Nguyen, Thi Hoang Huong, Bui Thi Thanh Nguyen, Giang Thi Lan Hoang, Nguyen Thi Hoa Pham, and Dang Thi Thu Cuc. 2024. Computer-delivered vs. face-to-face score comparability and test takers' perceptions: The case of the two English speaking proficiency tests for Vietnamese EFL learners. *Language Testing in Asia* 14(6): 1-30.

Ockey, Garry. 2009. The effects of group members' personalities on a test taker's L2 group oral discussion test scores. *Language Testing* 26(2): 161-186.

Ockey, Garry and Zhi Li. 2015. New and not so new methods for assessing oral communication. *Language Value* 7(1): 1-21.

Ockey, Garry, Lin Gu, and Madeleine Keehner. 2017. Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly* 14(4): 346-359.

O da Silva, Maria M., João M. Teixeira, Francisco F. Peres, and Cátia R. Maurício. 2025. Learning in the Metaverse: Reflections on Potential Benefits, Possibilities and Challenges. *Digital Society* 4(2): 1-14.

Ogiermann, Eva. (2018). Discourse completion tasks. In Andreas Jucker, Klaus Schneider and Wolfram Bublitz (Eds.), *Methods in pragmatics*. Berlin: Mouton de Gruyter.

O'Loughlin, Kieran J. 2001. *The equivalence of direct and semi-direct speaking tests* (Vol. 13). Cambridge: Cambridge University Press.

O'Sullivan, Barry and Cyril J. Weir. 2011. Test development and validation. In Barry O'Sullivan

(ed.), *Language testing theories and practices*. Basingstoke, UK: Palgrave Macmillan.

O'Sullivan, Barry Cyril, J. Weir, and Nick Saville. 2002. Using observation checklists to validate speaking-test tasks. *Language Testing* 19(1): 33-56.

Prensky, Marc. 2001. Digital natives, digital immigrants part 1. *On the Horizon* 9(5): 1-6.

Qian, David. 2009. Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly* 6(2): 113-125.

Robinson, Peter. 2001. Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics* 22(1): 27-57.

Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* 50(4): 696-735.

Schroeder, Ralph. 1996. *Possible worlds: The social dynamic of virtual reality technologies*. Boulder, CO: Westview Press.

Segovia, Kathryn Y. and Jeremy N. Bailenson. 2013. Identity manipulation-What happens when identity presentation is not truthful. In Yair Amichai-Hamburger (ed.), *The social net: Understanding our online behaviour*. Oxford: Oxford University Press.

Shohamy, Elana. 1994. The validity of direct versus semi-direct oral tests. *Language Testing* 11(2): 99-123.

Shohamy, Elana, Chambers Gordon, Dorry M. Kenyon, and Charles W. Stansfield. 1989. The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. *Bulletin of Hebrew Higher Education* 4(1): 1-20.

Silva Douglas Pereira, Sousa Cristina Pinto de, Vieira Eduardo Simão de Souza, Pedroso Maria Thereza Macedo, Silva Júnior Luiz Honorato da and Mazzafera Bernadete Lema. 2025. From the field to the virtual: A systematic review of the use of the metaverse in agriculture. *Editora Impacto Científico*: 59-75.

Skehan, Peter, and Pauline Foster. 1997. Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research* 1(3): 185-211.

Sternberg, Robert J. 1977. *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Hillsdale, NJ: Erlbaum.

Strauss, Anselm, and Juliet Corbin. 1994. Grounded theory methodology. In Norman Denzin and Yvonna Lincoln (eds.), *Handbook of qualitative research*. 273-285. Thousand Oaks, CA: Sage.

Traphagan, Tomoko Watanabe, Yueh-hui Vanessa Chiang, Hyeseung Maria Chang, Benjaporn Wattanawaha, Haekyung Lee, Michael Charles Mayrath, Jeongwon Woo, Hyo-Jin Yoon, Min Jung Jee, and Paul E. Resta. 2010. Cognitive, social and teaching presence in a virtual world and a text chat. *Computers & Education* 55(3): 923-936.

van Lier, Leo. 1989. Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly* 23(3): 489-508.

van Moere, Alistair. 2006. Validity evidence in a university group oral test. *Language Testing* 23(4): 411-440.

van Moere, Alistair. 2012. A psycholinguistic approach to oral language assessment. *Language Testing* 29(3): 325-344.

Wang, Yuping. 2004. Supporting synchronous distance language learning with desktop video conferencing. *Language Learning and Technology* 8(3): 90-121.

Warburton, Steven. 2009. Second Life in higher education: Assessing the potential for and the barriers to deploying virtual worlds in learning and teaching. *British Journal of Educational Technology* 40(3): 414-426.

Wehner, Amy K., Andrew W. Gump, and Steve Downey. 2011. The effects of second life on the motivation of undergraduate students learning a foreign language. *Computer Assisted Language Learning* 24(3): 277-289.

Weir, Cyril J. 2005. *Language testing and validation*. Hampshire: Palgrave Macmillan.

Xu, Xiaofei and John Impagliazzo. 2024. Metaverse services in computing and engineering education. *Frontiers of Digital Education* 1(2): 132-141.

Yang, Huijin. 2023. Assessing nonverbal and verbal interactional competence in a video-mediated oral test. *Linguistic Research* 40(Special Edition): 171-205.

Appendix 1. Group oral task topics

You will have a discussion with your two partners. The purpose of the discussion is to complete the task below. You should try to share time equally and keep the discussion going for five minutes. You will be asked to begin the discussion in one minute.

Topic 1 People spend lots of money and time on learning English. Based on your experience, what do you think is the best way to learn English? Share your English learning experiences and discuss ideas, resources, and strategies that help you learn English. Show your agreement and disagreement with your partners' opinions.

Topic 2 Teachers play an important role in students' learning. Based on your experience, what do you think are the qualities of a good teacher? Share your experiences with your teachers and discuss what makes a good teacher or what qualities a good teacher has. Show your agreement and disagreement with your partners' opinions.

Topic 3 What do you want in a spouse - someone who is intelligent, someone who has a sense of humor, someone who is physically attractive, someone who is rich, or someone who is reliable? Describe the type of person you would like to be in a relationship with or to marry. Discuss why these qualities are important in a marriage. Show your agreement and disagreement with your partners' opinions.

Topic 4 Some people say classmates are a more important influence than parents or teachers on a teenager's success in school. Do you agree or disagree? Share your experiences and discuss with your partners who you think is the most influential in a teenager's success in school. Show your agreement and disagreement with your partners' opinions.

Appendix 2. Scoring rubric

Score	Delivery	Language Use	Topic Development	Interactional Competence
0	Fragments of speech that are so halting that conversation is not really possible. Sounds incomprehensible.	Cannot produce a sentence.	Topic is not developed at all.	Shows no awareness of other speakers; may speak but not in a conversation-like way.
1	Consistent pronunciation and intonation problems cause considerable listener effort and frequently obscure meaning. Delivery is choppy, fragmented. Speech contains frequent pauses and hesitations.	Produces very basic sentence forms. Overall, turns are short, structures are repetitive, and errors are frequent.	Limited relevant content is expressed. The response lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task.	Does not initiate interaction, produces monologue only; Shows some turn-taking, may say, "I agree with you," but not relate ideas in explanation; too nervous to interact effectively.
2	Speech is clear at times though it exhibits problems with pronunciation, intonation or pacing and so may require significant listener effort.	Primarily uses basic sentences; more complex-structures are absent or contain significant errors. Vocabulary sufficient to discuss topic, but generally simple. Errors are common.	The response is connected to the task; though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed with limited elaboration.	Response to others without long pauses to maintain interaction; shows agreement or disagreement between others' opinions.
3	Speech is generally clear with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation or pacing and may require some listener effort. Overall intelligibility remains good.	Produces a mix of short and complex sentence forms, typically uses shorter forms. Vocabulary is adequate to discuss topics at length. Errors in grammar and vocabulary are sometimes noticeable.	Response is coherent and sustained and conveys relevant ideas. Overall development is somewhat limited.	Generally confident responds appropriately to others' opinions. Shows ability to negotiate meaning quickly and naturally.

4	Speech is clear, fluid and sustained. It may include minor difficulties with pronunciation. Pace may vary at times. Overall intelligibility remains high.	Makes use of longer sentences and a variety of structures. Uses a range of vocabulary; words are precise. Errors remain but not distracting.	Response is sustained and sufficient to the task. It is generally well developed and coherent; relationships between ideas are clear.	Turn-taking is very smooth. Can initiate discussion and conclude the discussion. Shows agreement and disagreement with the interlocutors.

Jayoung Song
Assistant Professor
Department of Asian Studies
Pennsylvania State University
201 Old Main, University Park,
State College, PA, U.S.A.
E-mail: jayoung.song@psu.edu

Received: 2025. 08. 28.
Revised: 2025. 09. 13.
Accepted: 2025. 09. 19.