INVESTIGATING TEST DELIVERY MODES WITHIN VIDEO-CONFERENCED ENGLISH SPEAKING PROFICIENCY ASSESSMENT

By

Jin Soo Choi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Second Language Studies – Doctor of Philosophy

2022

ABSTRACT

INVESTIGATING TEST DELIVERY MODES WITHIN
VIDEO-CONFERENCED ENGLISH SPEAKING PROFICIENCY ASSESSMENT

By

Jin Soo Choi

Nonverbal behavior is essential in human interaction (Gullberg, de Bot, & Volterra, 2008; McNeill, 1992, 2005). For second language speakers, nonverbal features can be helpful for successful and efficient communication (e.g., Dahl & Ludvigsen, 2014). However, due to the complexity of nonverbal features, language testing institutions have conventionally employed an asynchronous format (e.g., audio-only mode), focusing on the test-taker's verbal features. Recently, the advance in technology, and most importantly, the coronavirus pandemic (COVID-19) outbreak in late 2019 and early 2020, have accelerated the use of video-conferencing applications in educational contexts, including speaking tests (Nakatsuhara, Inoue, Berry, & Galaczi, 2021). Accordingly, the synchronous test delivery mode (video-conferencing), which presents test-takers' visual information, is a timely and necessary approach to addressing the impact of nonverbal features on L2 oral communication.

In response to these issues, I investigated the comparability of different delivery modes of speaking proficiency assessment. This study aimed to understand the dimensionality of the measured speaking construct and the impact of test-takers' visual information on rater behavior. Two datasets were used to address the research goals: first dataset (i.e., dataset 1) included scores of 110 test-takers, assessed by 8 trained raters; second dataset (i.e., dataset 2) included scores of 284 prospective international teaching assistants (ITAs), assessed by 12 professional and certified raters. I collected dataset 1, and English language learning and testing program at a

Midwest University provided dataset 2. I used two quantitative techniques to analyze both datasets: confirmatory factor analysis (CFA) and Multifaceted Rasch model (MFRM) analysis. For dataset 1, I qualitatively analyzed the raters' verbal reports.

Dataset 1 had an asynchronous format; all test-takers' performances were recorded for rating. Eight trained raters gave scores to the audio-recorded and video 1 mode (test-taker and test giver were displayed). Three weeks later, the raters assessed video 2 mode (only test-taker's visual information was displayed). Within one month, raters participated in a one-on-one semi-structured interview. Dataset 2 stems from an operational testing context. This dataset has only scores, as I borrowed the data from the operational English testing program at the university. The scores were first given by examiners in a synchronous format (Live mode) and later by raters in an asynchronous format (Recorded mode).

CFA findings indicated the multi-dimensional aspect of the underlying construct of speaking for both datasets, but the high inter-correlations showed that these are associated. Findings of MFRM revealed that raters showed leniency when rating (a) video mode over audio-only mode (dataset 1) and (b) synchronous mode over asynchronous mode (dataset 2). Findings suggest that using the video-conferenced delivery mode may be beneficial. However, the degree of usefulness across video modes differed, and how the raters utilized test-takers' nonverbal behaviors (e.g., gaze) varied. Thus, I decided that further investigation is needed to sufficiently support the use of video-conferencing applications to complement the physical face-to-face delivery mode. Overall, future research is highly recommended regarding the standardization of scoring of nonverbal features about the types of video mode, which would assist with the practical and valid application of virtual speaking tests.

For my family.

ACKNOWLEDGEMENTS

This dissertation is the milestone of my five-year Ph.D. journey. Looking back, I am grateful for the knowledge I learned and the opportunities I had as I worked on this dissertation project. All the stages I went through to complete the dissertation would have been impossible without the help of many people.

My first and immense appreciation goes to my dissertation chair, Dr. Paula Winke, for the support and guidance. Paula was the most enlightening, positive, and supportive advisor a graduate student could have. She was an ideal supervisor and a mentor, offering advice and encouragement with insight and humor. I still remember my meetings with Paula during the pandemic, when social distancing was required, and all the meetings were held virtually. Paula genuinely cared about my physical and mental health, offered any help I needed, and reliably arranged our virtual meetings. I have significantly benefited from Paula's sincere consideration and was able to complete my dissertation during the pandemic period. I also learned a lot from Paula: doing and writing research, writing for funds, being open to new opportunities, expanding opportunities by working with people, and balancing work and life with family. Most of all, from her knowledge and expertise in the field of language assessment, I learned how exciting it is to do research that could bring practical implications to the field. I am grateful for and proud of my time working with Paula.

I would like to recognize invaluable assistance from my other committee members. Words cannot express my gratitude to Dr. Dan Reed, Dr. Koen Van Gorp, and Dr. India Plough. I am incredibly thankful to Dan for the advice in the early stages of developing ideas and the opportunity to experience how language tests are administered. I am grateful for Koen's

guidance in designing methodology and conducting data analyses, and India's perspective and feedback in understanding the speaking construct in language assessment.

Next, I wish to express my special thanks for the financial support from the Fulbright Scholarship Program and Language Learning Doctoral Dissertation Grant. I also thank two resources from Michigan State University: a Dissertation Completion Fellowship from the College of Arts and Letters and the Graduate College, funds from the College of Arts and Letters and the Second Language Studies program.

My special thanks go to my fellow SLS students, especially Monique Yoder and Dylan Burton, for their generous and valuable help with the item development during the piloting stage. I would like to thank Dan Isbell and Dustin Crowther, for always being available to answer my questions about assessment and speaking. I also thank Dr. Heekyoung Kim, Aaron Ohlrogge, and Ann Letson at MSU for their practical guide and advice in understanding language assessment. I appreciate the professional feedback and comment from Nicole Jess at MSU, for all the statistical questions I had.

I would like to acknowledge my research participants for their engagement, especially those who joined as raters and second-coders, for their patience, contribution, and valuable suggestions.

Lastly, my family deserves endless gratitude, whose constant love and encouragement kept me motivated and confident. I am deeply grateful to my parents for their love and unwavering support, and my brother for always being a good friend. I am also thankful to my in-laws for their continuous support and belief in me. I owe thanks to an extraordinary person, my husband, Euipyo: Thank you for standing by me through all my peaks and valleys. I wouldn't have made this far without you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

KEY TO ABBREVIATIONS

ACTFL	American Council on Teaching Foreign Languages
ANOVA	Analysis of Variance
CEFR	Common European Framework
CFA	Confirmatory Factor Analysis
ICC	Intra-class Correlation Coefficient
IELTS	International English Language Testing System
ITA	International Teaching Assistant
L2	Second Language (includes Foreign Language)
LT	Language Testing
MFRM	Multifaceted Rasch Model
OPIc	Oral Proficiency Interview-Computer
SD	Standard Deviation
SLA	Second Language Acquisition
TOEFL iBT	Test of English as a Foreign Language internet-based Test

INTRODUCTION

People use both verbal and nonverbal features to achieve successful oral communication. Nonverbal features (e.g., gaze, gestures) are part of social interaction, which critically affect ongoing interaction among people (Galaczi & Taylor, 2018; Gullberg, de Bot, Volterra, 2008; McNeill, 2005; Montero Perez, 2020; Plough, Banerjee, & Iwashita, 2018). Without nonverbal features, speech is considered incomplete (Knight & Sweeney, 2007), and comprehending the addresser's underlying intention becomes difficult (cf. Kita, Alibali, & Chu, 2017). Regardless, conventionally administered English speaking proficiency assessments have primarily focused on verbal aspects such as linguistic categories. Therefore, standardized and local English speaking tests have reflected the individualistic and psycholinguistic-oriented design of the speaking construct (Iwashita, May, & Moore, 2021) by adopting semi-direct (asynchronous) test delivery modes. In the most widely used semi-direct mode, raters generally award scores on linguistic categories (e.g., fluency, pronunciation, grammar, lexis, comprehensibility, accentedness) by listening to test-takers' audio-recorded speech samples, and test-takers' speaking proficiency is inferred based on linguistic scales.

Even in the direct (synchronous) test delivery mode such as face-to-face delivery mode, nonverbal features are disregarded from rating processes due to the complex nature of measurement (e.g., Nakatsuhara, Inoue, & Taylor, 2020). Nonetheless, researchers have empirically demonstrated that when test-takers' visual information is presented, raters' judgments may be much more dependent on nonverbal criteria than previously considered (e.g., He & Young, 1998; Jenkins & Parra, 2003; Lazaraton, 1996; Neu, 1990). Thus, assessing verbal features and disregarding nonverbal aspects to infer test-takers' true second language (L2) speaking ability has been consistently questioned by researchers (Roever & Ikeda, 2022).

In recent years, the advance in technology and the outbreak of the Corona pandemic have accelerated the use of video-conferencing applications (e.g., Zoom), which resulted in the need to reflect on nonverbal cues in rating for construct validity. In the pre-pandemic era, researchers mainly investigated the feasibility of video-conferencing applications, which display participants' visual information during speaking tests (e.g., Davis, Timpe-Laughlin, Gu, & Ockey, 2018; Nakatsuhara et al., 2017, 2020; Ockey, Timpe-Laughlin, Davis, & Gu, 2019). The outbreak of the COVID-19 global health pandemic has fastened the need to understand better the role of video-conferenced speaking tests regarding the construct validity and fairness of a test. The Corona pandemic caused social distancing in numerous professional fields, and education was no exception. Video-conferencing platforms were largely employed in local language testing contexts (e.g., ITA Speaking Test, language test for prospective undergraduate students) and standardized language testing contexts (e.g., IELTS At Home, TOEFL Speaking At Home). Without much preparation, testing institutions had to abruptly transit from their conventional test delivery modes (e.g., audio-only, face-to-face) to virtual contexts (i.e., video-conferencing).

While the feasibility of video-conferencing applications in speaking assessment has been investigated, the impact of the newly employed video delivery mode, specifically on rater behavior and how video delivery mode differs from other delivery modes (e.g., audio-only) is still under exploration. That is, whether the inclusion of participants' visual information in different types of video modes (e.g., synchronous, asynchronous, absence/presence of a test giver) will positively or negatively impact on rater behavior is barely investigated. So far, studies have mainly compared the face-to-face and audio-recorded delivery modes, and limited studies (e.g., Nakatsuhara et al., 2020, 2021) have focused on the comparability of different types of modes used in virtual settings. Thus, further investigation of comparability across different test

delivery modes is timely warranted, as different impacts of the delivery modes could lead to greater variation in ratings and the final scores test-takers receive.

One potential approach for helping local language testing centers to make more informed decisions and more valid test items, test-takers' performances, and rater behaviors from video-conferenced speaking tests during the pandemic can be investigated. With detailed knowledge of when raters show different behaviors across different delivery modes, and which aspect of visual information they consider helpful or distracting, test developers and administrators can decide which delivery mode to use and how to train raters regarding the use of nonverbal features.

This dissertation investigates the comparability of different test delivery modes (audio-only, synchronous video-conferencing, and asynchronous video-recorded) to inform video-conferenced English-speaking proficiency assessments. In the following chapters, I illustrate a project that includes two different datasets regarding video-conferenced English speaking tests.

Chapters 1 to 5 contribute to the literature review section for this project. In Chapter 1, I outline the overall structure of the literature review. Then, in Chapter 2, I review the literature on speaking construct, particularly focusing on nonverbal features, establishing a theoretical framework using video-conferenced L2 speaking proficiency tests. In Chapter 3, I provide a historical background regarding the COVID-19 global health pandemic outbreak, which influenced the local and standardized English language testing centers' and institutions' transition from the conventional audio-only test delivery mode to the use of video-conferenced delivery modes in the United States. In Chapter 4, I overview how different test delivery modes have measured the different aspects of the speaking construct. Four test delivery modes are reviewed; the audio-only mode, video-conferencing mode (synchronous), video-recorded mode (asynchronous), and face-to-face mode. Chapter 5 features, in detail, the previous studies that

compared the different test delivery modes by analyzing test scores and participants' verbal data. Chapter 6 closes the literature review with research gaps and three research questions for this study.

In Chapter 7, I briefly outline the two different datasets used in this project: dataset 1 (experimental video-conferenced speaking test) and dataset 2 (ITA Speaking Test). Chapter 8 describes the methodology ofdataset 1, with detailed information about participants, instruments, and an overview ofr the test taking and rater training procedures. The methodology of dataset 2 is provided in Chapter 9, which provides information about participants, instruments, and an overview of the rating procedure. In Chapter 10, I close the methodology section with suggested data analyses for the three research questions. Chapter 11 presents hypothesized models for the confirmatory factor analysis (CFA) used to answer the first research question.

Chapter 12 describes the overall structure of the Data Findings section with regards to the research questions and analyzed datasets. In Chapter 13, I present the results of analyses of dataset 1: data distribution, one-way ANOVAs, confirmatory factor analysis (CFA; RQ1), and Multifaceted Rasch model analysis (MFRM; RQ2). In Chapter 14, I report the results of dataset 2 at an item-level: data distribution, CFA for RQ1, and MFRM for RQ2. Chapter 15 presents the results for item types in dataset 2. In Chapter 16, I draw on interviews with eight raters (dataset 1) who assessed three test delivery modes.

Chapters 17 to 20 are dedicated to the Discussion section of this study. In Chapter 17, I briefly review the previous chapters and overview the construct of the Discussion section. In Chapter 18, I discuss the findings for dataset 1 through a critical review of the role of nonverbal behavior in video modes. Chapter 19 discusses the findings for dataset 2, also through the lens of

nonverbal behaviors and construct validity of the video-conferenced speaking test. Chapter 20 analyzes raters' verbal data regarding their perception of different delivery modes.

Finally, in Chapter 21, I close the dissertation with a discussion of room for further research, followed by broader implications for local language testing centers and test developers who plan to use video-conferencing applications for their L2-English speaking proficiency assessments.

CHAPTER 1: OVERVIEW OF LITERATURE REVIEW

In this Literature Review section, there are two overarching goals. First, I aim to provide a context considering the theoretical frameworks of the speaking construct, specifically in terms of nonverbal features, which is now becoming an increasingly important aspect within speaking assessment. Second, I develop a context of how different test delivery modes have been used to measure different aspects of L2 speaking proficiency. To do so, I review previous studies that investigated the comparability across different speaking test delivery modes; face-to-face, audio-only, video-conferencing (synchronous), and video-recorded (asynchronous). Five chapters are in this section, and below is a brief overview for each chapter.

In Chapter 2, I review the theoretical backgrounds of speaking construct, specifically considering the importance of nonverbal features. Then, in Chapter 3, I focus on the video-conferencing platform (e.g., speaking tests using Zoom), which has been increasingly used since the outbreak of the COVID-19 global health pandemic in late 2019 and early 2020. In Chapter 4, I review the extent to which different test delivery modes tap into the construct of speaking in assessment contexts. In Chapter 5, I review previous studies which investigated the comparability of different speaking test delivery modes (i.e., audio-only, face-to-face, video-conferencing, video-recorded) regarding the construct validity. Lastly, in Chapter 6, I close the literature review of this study with three research questions based on the theoretical standpoints and the previous studies reviewed these chapters.

CHAPTER 2: CONSTRUCT OF SPEAKING

Increased Awareness of Nonverbal Features

Language is multimodal; humans do not just discuss using speech. They use numerous visual articulators as well. Researchers in the second language acquisition (SLA) and language testing (LT) fields have consistently suggested that the speaking construct is multifaceted, multimodal, and includes verbal and nonverbal components (e.g., Bachman & Palmer, 2010; Canale, 1983; Canale & Swain, 1980; He & Young, 1998; Hymes, 1972; Young, 2011). Specifically, researchers have highlighted the importance of nonverbal features for decades; from Canale (1983) who noted that nonverbal components (e.g., facial expressions, eye gaze, gestures) play a salient role in oral interaction, to Montero Perez (2020), who commented that "language learning is necessarily multimodal learning" (p.654). Nonverbal elements have long been suggested as enhancing communication effectiveness.

Theoretical frameworks for the speaking construct within SLA and LT have also stated nonverbal features are part of the construct. The communicative competence framework is one that has nonverbal features as a central element, and it is the most widely used and influential theoretical framework within SLA. It was developed by Canale and Swain (cf. Canale, 1983; Canale & Swain, 1980). Canale and Swain (1980) introduced a tripartite division of the concept[1]; grammatical competence (i.e., linguistic knowledge), sociolinguistic competence (i.e., sociocultural rules of speech), and strategic competence (i.e., verbal and non-verbal communication strategies). Under the strategic competence, nonverbal behavior is explained as a feature used to compensate for breakdowns such as "momentary inability to recall an idea or

[1] Canale (1983) later added a separate fourth component, discourse competence.

grammatical form" (Canale, 1983, pp.10-11), or to enhance communication effectiveness such as "deliberately slow and soft speech for rhetorical effect" (p.10). In line with Canale (1983), Neu (1990), and Pennycook (1985) also suggested nonverbal behavior as a crucial aspect of communicative competence, and provided an example that skilled communicators employ nonverbal elements to add information during oral communication (Neu, 1990). However, it is worth noting that the major focus of this communicative competence framework was on verbal features (i.e., grammatical competence), as majority of speaking proficiency assessments reflect individualistic and psycholinguistic orientation to the L2 speaking construct (Iwashita, May, & Moore, 2021). The communicative competence framework was later adopted for language assessment by Bachman (1990, 1991; Bachman & Palmer, 1996, 2010), but at that time, non-verbal behaviors were not carried forward as strongly part of the construct in assessment contexts.

Another framework that discusses nonverbal features is interactional competence (IC), which state that oral communication is "co-constructed by all participants in an interactive practice and is specific to that practice" (He & Young, 1998, p.7). Galaczi and Taylor (2018) identified five domains of IC, and *nonverbal or visual behaviors* sits as one, alongside *topic management*, *turn management*, *breakdown repair*, and *interactive listening*. Within this framework, the speaking construct is expanded from heavy weighted focus on production features (e.g., fluency, pronunciation, grammar, vocabulary), to embodied means of nonverbal communication such as gestures, gaze, head orientations, and facial expressions to accomplish speakers' social actions (e.g., Burch & Kasper, 2016; cf. Ducasse & Brown, 2009; Galaczi & Taylor, 2018; He & Young, 1998; Kramsch, 1986; Plough et al., 2018; Roever & Ikeda, 2022; Roever & Kasper, 2018; Ross, 2018; Young, 2008, 2011). Simply put, nonverbal behaviors are

considered as social interaction skills (Riggio, 1992) that could support participants to manage their social persona and present oneself more attractively and favorably (Lippa, 1975).

Although more emphasis for nonverbal features was given in IC than in the communicative competence framework, IC's main focus also stayed on verbal aspects, particularly oral interaction, explained by using conversation analysis method. To this phenomenon, Roever and Ikeda (2022) emphasized the need to consider nonverbal aspects of IC as part of the speaking construct to be measured, because these allow participants to use a range of nonverbal (semiotic) resources (e.g., gaze, body posture, space) to achieve their goals. Regardless of the suggested importance of nonverbal features for successful oral communication, there has been few theoretical discourse about the nonverbal elements as a part of measurable speaking construct, and "very few attempts have been made to theoretically account for the fact that L2 speakers do and say different things, an L2-specific form of speech-gesture discrepancy" (Gullberg, de Bot, & Volterra, 2008, p.159).

Empirical demonstration regarding the critical role of nonverbal behavior in human communication has been found in studies from different fields (e.g., human-computer interaction, neuropsychology, psychology, social science). Among various nonverbal features, these studies emphasize gaze as one of the strongest and most extensively investigated visual cues. Gaze, within two-party interaction, sends strong social cues (Kampe et al., 2001) and is associated with variety of functions such as managing intention, social control, or highlighting a particular speech event (e.g., Ijuin et al., 2018; Palanica & Itier, 2012; Senju & Hasegawa, 2005; von Grünau & Anston, 1995). For example, participant's direct eye gaze to the addressee were reported to be more attractive, favorable, and attention-grabbing than averted gaze (e.g., Conway et al., 2008; Mason et al., 2005; Senju & Hasegawa, 2005; von Grünau & Anston, 1995), which

could affect achieving successful communication. Additionally, nonverbal cues were used as important resources for even linguistically adept adults, who relied on gaze perception to guide and interpret social behavior (e.g., Frischen, Bayliss, & Tipper, 2007).

Within SLA studies, when learning the target language, the positive aspects of visual information (e.g., nonverbal cues) were demonstrated more than its negative influence (Montero Perez, 2020). For example, SLA researchers have shown that from the speakers' perspective, nonverbal features do the following: (a) they support speakers' way of expressing their intended message, as well as aid in helping them formulate the linguistic structure (e.g., motion events: Choi & Lantolf, 2008; McNeill, 1992, 2005); (b) when appropriately controlled by interviewees, nonverbal features were able to help them compensate for their weaker linguistic proficiency (e.g., Gullberg, 2006; Kendon, 2004), thus giving an impression of having desired interactional competence to interviewers (e.g., eye contact, smiling or positive facial affects: Jenkins & Parra, 2003; Roever & Kasper, 2018); (c) they facilitate speakers' retrieval of words and sentence structures from memory that could reduce cognitive burden (e.g., Cassell, McNeill, & McCulloguh, 1999; Goldin-Meadow et al., 2001; Wagner, Nusbaum, & Goldin-Meadow, 2004); (d) they support speakers to fill the linguistic functions such as structural slots, referential content to deictic expressions, and modifying speech acts (e.g., Clark, 1996; Engle, 1998; Kendon, 2004; Slama-Cazacu, 1976); and (e) nonverbal features support with the retention of information regarding learners' short-term memory (e.g., Cohen & Otterbein, 1992).

Additionally, the researchers found that nonverbal features were beneficial for listeners, as these (a) facilitate listeners' comprehension (e.g., gestures: Dahl & Ludvigsen, 2014; Goldin-Meadow, 2003; lip movements: McGurk & McDonald, 1976; gestural feedback: Nakatsukasa, 2016; gestures and facial cues: Sueyoshi & Hardison, 2005; facial expressions, hand gestures:

Tsunemoto et al., 2021); and (b) assist listeners' memory for interpretation of speech (e.g., Beattie & Shovelton, 1999; Overoye, 2019). Overall, the previous studies showed that speakers composite their intended message as they deliberately "ensemble" audio and visual cues (Gullberg et al., 2008; Kendon, 2004; McNeill, 1998), and the absence/presence of visual information is a sensitive factor that could facilitate speakers' production and listeners' comprehension of speech.

Within the language testing (LT) field, the focus of this current study, researchers have also demonstrated the positive impact of nonverbal cues on rater behavior in speaking assessments. In the previous studies, raters reported that test-takers' nonverbal cues positively contributed to (a) understanding what test-takers were saying, (b) better comprehending test-takers' intended message using nonverbal means, and (c) understanding what happened during test-takers' pauses, hesitation, repetition, and awkwardness (Jenkins & Parra, 2003; Lam, 2018, 2021; May, 2011; Nakatsuhara et al., 2020; Nambiar & Goon, 1993; Neu, 1990; Roever & Kasper, 2018). In addition, raters were fond of test-takers with visual cues, as they seemed more proficient than test-takers who did not have them (cf. Gullberg, 1998; e.g., Jenkins & Parra, 2003; McCafferty, 2002). Altogether, these findings support Bachman's (1991) emphasis that "we now know that a language test score cannot be interpreted simplistically as an indicator of the particular language ability we want to measure" (p.677). In particular, studies of oral proficiency interviews (ACTFL, institutional tests; cf. Jenkins & Parra, 2003) indicated that assessment of test-takers' L2 speaking proficiency may be much dependent on nonverbal criteria than previously considered, which reflects test-takers' interaction skills employed during their test performances (He & Young, 1998; Lazaraton, 1996).

While nonverbal cues have received increased attention in speaking assessments, due to various constraints such as delivery mode, the measurement of L2 speaking ability has been limited. In other words, nonverbal features (e.g., gaze, gestures, head orientation, backchannel) weren't explicitly defined as the speaking construct, and rating was much more focused on verbal aspects such as linguistic features (e.g., fluency, lexis, grammar, pronunciation, accentedness). Semi-direct (asynchronous) has been the most widely used and favored format in large-scale L2 speaking assessments, in which practicality and reliability are of paramount importance. Audio-only mode is one common mode, which requires raters to award scores as they listen to audio-recordings of test-takers' responses. One pitfall of the asynchronous test mode is that it may lose some aspects of nonverbal components over synchronous (video-conferencing, face-to-face) modes, which may be seen more as modes that tap into social skills. Regardless of the different information conveyed by different test delivery modes (cf. van Leeuwen, 2004), raters' judgments on test-takers' speaking abilities were mostly based on verbal features defined in rubrics, as the tests are based on individualistic-psycholinguistic oriented designs. Thus, the effects of visual stimuli and different types of visual information on rater behavior is an area that has been under-investigated.

Against this background, researchers have recently employed video-conferencing platforms to measure the true and expanded construct of L2 speaking ability (e.g., Batty, 2014; Clark & Hooshmand, 1992; Craig & Kim, 2010; Kim & Craig, 2012; Nakatsuhara et al., 2017, 2020, 2021) by investigating the construct validity of different speaking test delivery modes (e.g., audio-recorded, video-conferencing, video-recorded, face-to-face) and how these modes affect rater behavior. While few studies were conducted to explore the role of video-conferenced speaking tests (e.g., Davis et al., 2018; Nakatsuhara et al., 2017, 2020; Ockey et al., 2019), the

outbreak of COVID-19 global health pandemic accelerated testing institutions' and local language testing centers' use of video-conferencing platforms. In the following, I further review about the impact of COVID-19 and its impact on the use of video-conferenced speaking tests.

CHAPTER 3: SPEAKING ASSESSMENT AND THE COVID-19

Accelerated use of Video-Conferencing Platforms in Speaking Assessments

In recent years, the effort to expand the speaking construct in assessments has been accelerated, largely due to the outbreak of the COVID-19 global health pandemic in early 2020. COVID-19 has triggered social distancing, physical proximity restriction, and remote working/studying environments, which affected the majority of domains such as education (mid/final exams taken online, remote- or hybrid-teaching classes) and professional work fields (job-interviews, business meetings). This same trend holds for the growing number of fields where technology makes remote administration possible, and L2 speaking assessment was no exception. Speaking tests were held using video-conferencing applications such as Zoom (www.zoom.us), by both standardized tests (e.g., IELTS Speaking section; Nakatsuhara et al., 2017, 2021) and local language tests (e.g., ITA Speaking Tests, Community English Language Program Online Placement Exam, University's English placement test of oral communication; for more information, see the special issue in Language Assessment Quarterly, 2021, edited by Kunnan and Ockey).

The use of a video-conferencing platform brings unique benefits to speaking assessment, largely due to the platform's potential as an online face-to-face mode that allows authentic oral communication for people who are continents apart. The synchronous video-conferencing mode also resolves the issue of semi-direct modes (e.g., audio-recordings) by providing visual information to test-takers and examiners (test givers) in real time. While some technical issues remain (e.g., sound or video quality, stability of internet connection) that could influence interaction, the video-conferencing mode is by far the only and best possible delivery mode that could complement the practical constraints on human interaction occurred by COVID-19. With

the increased use of video-conferencing platforms, LT researchers have begun to investigate the comparability of different delivery modes of speaking tests, particularly to better understand the potential of video-conferencing as a test delivery mode, also as an alternative to the face-to-face mode (e.g., Nakatsuhara et al., 2017, 2020, 2021).

While few studies have explored the feasibility of video-conferencing applications in speaking assessments pre-pandemic (e.g., Davis et al., 2018; Nakatsuhara et al., 2017; Ockey et al., 2019; Zhou, 2015), the unexpectedly swift transition to the video-conferencing mode during pandemic was a new experience for many, including language testing centers, testing institutions, and participants. Particularly, the COVID-19 pandemic has centered much language assessment on local testing contexts (e.g., English language placement testing at BYU-Hawaii: Green & Lung, 2021; Iowa State University's English placement test of oral communication: Ockey, 2021; Community English Language Program Online Placement Exam at Teachers College, Columbia University: Purpura, Davoodifard, & Voss, 2021; cf. special issue in Language Assessment Quarterly, 2021) and standardized testing contexts (e.g., TOEFL iBT At Home Testing[2], IELTS Online[3]). In the quick response to COVID-19, many centers and institutions needing test scores have created their own, in-house, video-conferenced OPI-like exams (for a comprehensive review of at-home proficiency tests, see Isbell & Kremmel, 2020). These new and wide-spread examinations are changing how the construct of speaking is defined once again.

The replacement of the conventional speaking test delivery modes (audio-recordings, face-to-face interviews) to video-conferencing applications warrants further inspection regarding

[2] https://www.ets.org/toefl/test-takers/ibt/test-day/at-home/
[3] https://www.ielts.org/news/2021/ielts-new-at-home-testing-option

the construct validity of the different speaking test delivery modes. That is, are the different delivery modes (e.g., video-conferencing, video-recorded, audio, face-to-face) comparable? Can the stakeholders and test administrators use scores of the video-conferencing speaking tests as a replacement of audio-recorded or face-to-face speaking test modes? Are the different types of video-conferencing modes comparable (synchronous vs. asynchronous, video-recorded screen displaying only test-taker vs. displaying both test-taker and examiner)?

These questions are particularly important in a time when video-conferencing platforms are widely used in our daily life communication. Soon, the use of audio- or video-recorded performances, either synchronously or asynchronously, will allow the large-scale, high-stakes, speaking assessment companies to choose among different delivery modes. For this reason, improved understanding regarding rating outcomes and rater behavior under different delivery modes is critical for research and stakeholders. Thus, investigating the construct validity among different delivery modes would be an important endeavor to better utilize speaking proficiency tests and critically enable "the adequacy and appropriateness of inferences and actions based on test scores" (Messick, 1989, p.13), that would benefit stakeholders and test-takers.

In the following, I review the purpose and the extent of the speaking construct different test delivery modes measure, and the previous studies that compared these different delivery modes of the speaking assessments.

CHAPTER 4: TEST DELIVERY MODES IN SPEAKING ASSESSMENT

Each delivery mode is suggested to tap into different aspect of the construct of speaking (cf. Nakatsuhara et al., 2020; Zhou, 2015). In this section, I discuss the extent to which each test delivery mode measures the speaking construct. I start from the mode that measures the widest aspect of the speaking construct (i.e., face-to-face) and move to the narrowest aspect of the speaking construct (i.e., audio-only). Specifically, I review the video modes regarding nonverbal cues (e.g., gaze, head orientation) that are mostly displayed in the modes of video-conferencing applications.

Face-to-Face Mode

Face-to-face delivery mode has been used in speaking tests for over a century (Weir, Vidakovic, & Galaczi, 2013), and it is the most authentic oral communication context. Researchers (e.g., Nakatsuhara et al., 2017, 2020) suggest that the speaking construct measured under this mode generally taps into linguistic, social, interactional, and nonverbal traits. For this reason, many high stakes speaking tests (e.g., IELTS Speaking, Cambridge English Exams, General English Proficiency Test in Taiwan) have employed the face-to-face mode (oral interaction between examiner and test-taker) to assess L2 learners' speaking proficiency. However, the "here-and-now" nature of face-to-face mode is the reason for low practicality and why semi-direct speaking tests, or automated speaking tests have been widely used in many tests (Nakatsuhara et al., 2017).

So far, rubrics in the face-to-face mode focus on verbal features (i.e., linguistic categories), and nonverbal aspects are not explicitly defined as part of the speaking construct in speaking assessments (Nakatsuhara et al., 2020) regardless of its critical role in real-life face-to-face communication (e.g., Vo, 2019). Considering the current assessment scales used in

institutional settings (e.g., the American Council on the Teaching of Foreign Languages, or ACTFL in the United States, and the Common European Framework of Reference for languages, or CEFR in Europe), and that speaking tests based on those scales normally lack a description of the role of nonverbal features or interactional dynamics (Salaberry & Burch, 2021), it should be pointed out that the following test delivery modes use rubrics that have construct definitions only for verbal features, particularly pronunciation, fluency, lexis, grammar, accentedness, and comprehensibility.

Video-Conferencing Mode

Video-conferencing (VC) mode is currently the most widely used mode in speaking tests across the globe. With the use of video-conferencing applications, participants are able to synchronously interact with one another, across different physical locations. In this time of era, the VC mode brings practical advantage of remotely connecting test-takers and examiners who are in different locations while preserving the co-constructed nature of face-to-face delivery mode.

Regarding the construct validity, LT researchers have found that the VC mode is comparable to the face-to-face mode and potentially taps into interactional resources (Nakatsuhara et al., 2017, 2021; cf. Berry et al., 2018; Ockey et al., 2019) because test-takers' ability to orally interact with an interlocutor is measured in real-time, as many test designers are aiming for (Butler, Eignor, Jones, McNamara, & Suomi, 2000; Fulcher, 2003; McNamara, 1996; Ockey, 2014; Ockey, Gu, & Keehner, 2017). The evidence for the comparability between the VC and face-to-face modes was demonstrated by previous studies (e.g., Craig & Kim, 2010; J. Kim & Craig, 2012; Nakatsuhara et al., 2021) that reported insignificant differences between the modes. These researchers highlighted that the VC mode and face-to-face mode may measure the

same speaking construct compared to other modes (i.e., audio-only, video-recorded modes). Identical to the face-to-face mode, the rubric used in the VC mode focus on verbal features, and nonverbal features or interactions are not explicitly defined as the speaking construct in the rubrics.

While the VC mode is promising as an alternative for face-to-face mode, I would like to note that these modes display different ranges of visual information. In the VC mode, the use of single-camera diminishes participants' visual information, compared to the face-to-face mode that shows full range of participants' visual information (Davis et al., 2018). That is, in the face-to-face mode, participants' whole-body image (e.g., physical appearance, posture, touching behavior, kinesics behavior, gaze, skin sensitivity, physical characteristics) is visible, while in the VC mode, the speakers' image of their upper torso (from head to shoulder) is usually displayed on the screen (people usually sit on a chair and have their laptops/smartphone/pads on the table). In the VC mode, it is generally speakers' gaze, facial expression, head orientation, and their backgrounds that are visible to another speaker.

Concerns regarding this limited visual information within speaking assessments have long been raised, since the early days when VC technology was not widely used. To this, Abigail Sellen (1995), who conducted pioneering studies in early VC technology era, stated that providing interlocutors visual access via VC technology does not represent the situation similar to "being physically co-present" (p. 407). Until now, researchers have stressed the limitations of using a single camera, and that VC formats may be inadequate to support the full range of features found in human communication (e.g., Groen, Ursu, Michalakopoulos, Falelakis, & Gasparis, 2012). In view of the importance of nonverbal skill as an integral aspect of social competence (Burgoon & Bacue, 2008; Feldman, Philipott, & Custrini, 1991), the different range

of visual information mediated by technology may affect how participants use and perceive the nonverbal cues to achieve communication goals (Gergle, Kraut, & Fussell, 2013). The limited visual information in the VC mode could potentially elicit different nonverbal skills and functions compared to face-to-face modes (e.g., Song & Hsu, 2021). Overall, the different range of visual information between the face-to-face and the VC modes could affect rater behavior (cf. Kendon, 1967; Senju & Hasegawa, 2005; Vuilleumier, 2002). Therefore, when administering the VC mode tests to compensate the face-to-face mode, it is important to consider the potential impact of differences in the range of visual input, as well as any diminished visual information in video modes, on rater behavior.

Video-Recorded Mode

Video-recorded (VR) mode has often been used for double-rating to achieve scoring validity (e.g., American Educational Research Association et al [AERA], American Psychological Association [APA], & National Council on Measurement Education [NCME], 2014). In the VR mode, raters are asked to watch the interaction between test-taker and examiner (test giver), or if tasks are monologic, raters watch test-takers' responses with their visual information displayed on a screen.

Within the video modes that use video-conferencing platforms, Nakatsuhara et al. (2020) demonstrated that raters in the VR mode were harsher than examiners in the VC mode, because examiners in the VC mode were given more cognitive load than raters in the VR mode. In the VC mode that is synchronous, examiners had to play a dual role (i.e., rating test-takers' performance while simultaneously responding to them) while the raters in the asynchronous VR mode were not required to interact with test-takers. Same as the face-to-face and VC modes, nonverbal features are not defined in the rubrics.

Audio-Only Mode

Lastly, audio-only mode taps into the narrowest aspect of the speaking construct, mainly verbal elements. It is the conventional test delivery mode that has been employed by testing institutions for decades, which is aligned with the prevailing institutional infrastructure of testing policies, testing standards, and testing instruments (cf. Salaberry & Burch, 2021). One major difference from the video modes is that the audio-only mode has no visual information of participants. In audio-only mode, raters are given with limited information regarding the nonverbal cues of test-takers/examiners, relationship between test-taker and examiner, or in which test-taker is situated (Nakatsuhara et al., 2020). Such lack of visual information in audio mode, as demonstrated by researchers (e.g., Conlan et al., 1994; Nakatsuhara et al., 2020), failed to make the best use of the benefits video modes offer that consequently affected rating scores. Regarding the extent to which the speaking construct can be validly assessed in the audio-only mode, therefore, is most constrained compared to the video modes. Nonetheless, the construct of speaking defined in rubrics are mostly in line with what audio-only mode captures, that is, individualistic and psycholinguistically oriented verbal features (e.g., fluency, accuracy, complexity, lexis, grammar, pronunciation).

Altogether, different test delivery modes have been used in speaking assessments depending on various reasons, such as practicality and cost efficiency issues. While the audio mode may be most cost- and time-efficient, it measures the narrowest part of the speaking construct. The video modes (face-to-face, VC, VR), on the other hand, measure the wider aspect of the speaking construct, but have been less investigated and administered in speaking assessments due to the complexity brought about by visual information and actual interactions in speaking tasks (e.g., Fulcher, 2003; Galaczi, 2008, 2014; Plough et al., 2018; Roever & Ikeda,

2022; Roever & Kasper, 2018; Weir, 2005; Young, 2011). While it looks appealing to administer VC speaking tests because they could tap into wider aspects of speaking construct, the addition of visual information has resulted in concerns about the construct validity of using them as a replacement for conventional audio-recorded or face-to-face delivery modes. Thus, fundamental questions regarding test delivery modes warrant further investigation, such as whether and how the test delivery medium changes the underlying construct of speaking being measured.

In light of the different test delivery modes tapping into different extent of the speaking construct and newly emerging delivery mode in speaking tests, in the following chapter, I review previous studies that compared different delivery modes (e.g., video-conferencing, face-to-face, audio-recorded, video-recorded) in L2 speaking assessment domain. Although there have been relatively few studies conducted especially for the VC mode, the findings of previous studies could be a steppingstone for obtaining fuller understanding of the construct validity of recent speaking test delivery mode.

CHAPTER 5: COMPARING DIFFERENT SPEAKING TEST DELIVERY MODES

Studies that Compared Different Speaking Test Delivery Modes

Since the late 1980s, researchers have investigated the construct validity of speaking tests across different delivery modes. The majority of the arguments came from the integration of quantitative (e.g., functional use of language, score data) and qualitative data (e.g., verbal reports, written comments), to give a more comprehensive basis for conclusions (cf. O'Loughlin, 2001; van Lier, 1989; Weir, 2005; Zhou, 2015). Scores (quantitative data) are useful evidence for comparing tests and supporting sound conclusions. However, researchers such as Nakatsuhara et al. (2021) warned that scores are "one lens for gathering evidence" (p. 371). Thus, qualitative data may be indirect evidence for comparison, but it brings important insights into perceptions of test-takers, raters, and examiners, which provides fuller picture of the delivery mode, such as the usability and stability of VC technology (Berry et al., 2018; Davis et al., 2018; Ockey et al., 2019). Thus, in this sub-section, I focus my review of the previous studies in terms of score comparison across modes and analysis of verbal reports or written commentaries.

Face-to-Face Mode vs. Video-Conferencing Mode

The video-conferencing (VC) mode has a shorter history in speaking assessments than that of the face-to-face mode; however, the use of the VC mode is not new. The earliest study that compared the VC mode and the face-to-face mode was in 1992, an exploratory study conducted by Clark and Hooshmand at the Defense Language Institute Foreign Language Center (DIFLC) in the United States. The technical developments at the Foreign Language Center at the DIFLC enabled the use of satellite-based video technology, which was mostly used for language instruction and was incorporated in assessment settings. The researchers called VC mode as "screen-to-screen testing" which broadcasted and received test-takers' audio and visual

information in real time. Two languages (Arabic, Russian) were tested in both delivery modes. The researchers reported no statistically significant differences for scores of Arabic proficiency test across modes, although face-to-face ratings were higher than VC ratings. For Russian proficiency test, scores were identical between the modes.

Clark and Hooshmand also gave short feedback questionnaires to test-takers and examiners. From the responses, test-takers preferred to be tested on a face-to-face mode than VC mode, while examiners reported no preferences for either test modes. Particularly, test-takers commented that VC mode had several distracting elements such as audio cut-outs, frozen image, and background noise. Despite such distractions, however, test-takers responded that they were able to concentrate to their speaking tasks.

Two decades later, Craig and Kim (2010) and Kim and Craig (2012) compared the face-to-face and VC mode with L2 English learners whose L1 was mostly Korean. Ratings were focused on global and analytic categories (fluency, functional competence, accuracy, coherence, interactiveness) for both modes. Before and after the assessment, test-takers responded to questionnaires regarding their "anxiety" in the two modes. The researchers reported that both global and analytic scores showed no statistically significant differences between the modes. Regarding the "anxiety" questionnaire that was operationalized by asking about "nervousness" and "comfort" when interacting with an examiner, test-takers responded they were comfortable in both modes Kim and Craig (2012). However, test-taker anxiety was significantly higher before the face-to-face mode than VC mode and they were more comfortable in the VC mode than the face-to-face mode. The researchers concluded that VC mode has several beneficial aspects including reliability and construct validity.

A recent study is by Nakatsuhara et al. (2017), who used a convergent, parallel mixed-methods design to investigate the comparability of IELTS Speaking test between the internet-based VC mode and face-to-face mode. The researchers' focus was on criterion-based validity to tap into the construct of speaking by comparing "different versions of the same test and into equivalences of parallel test versions" (p. 4). They analyzed scores of global and analytic scales (i.e., fluency, lexis, grammar, pronunciation), and four different resources of qualitative data (i.e., observers' field notes, examiner's written notes, examiner's feedback questionnaires, and examiners' verbal report on their rating of test-taker performances). Nakatsuhara et al. reported similar test scores and comparable range of language functions (*asking for clarification*, *comparing*, *suggesting*, etc.) for the modes, however, Nakatsuhara et al. observed differences in test-takers' functional output and examiners' behavior. For example, examiners reported they tended to slow their speech rate and articulate more clearly to ensure that test-takers understood them, which was possibly to "mitigate any perceived technical challenges (e.g., transmission delay or poor sound quality)" (p.12).

In terms of scores, the findings confirmed the equivalence between face-to-face and VC modes, which corroborated the findings by Clark and Hooshmand (1992). Craig and Kim (2010), and Kim and Craig (2012). Nakatsuhara et al. stated that although not statistically significant, VC mean scores were lower than face-to-face mode. From the comparability of scores, Nakatsuhara et al. concluded that VC mode could be a parallel alternative to a standard face-to-face mode.

It is important to note that such findings of score comparability can be interpreted from the unidimensionality perspective, the most investigated construct in language assessment (speaking tests: Huang, Bailey, Sass, & Change, 2020; Kim & Crossley, 2020; Liu, Aryadoust,

& Foo, 2022; Sawaki & Sinharay, 2017; Yan, Cheng, & Ginther, 2019). In research that compares speaking test delivery modes, Nakatsuhara et al. (2017) is one of the few studies that explicitly mentioned unidimensionality, and who reported that from their results of 4-facet MFRM analysis (*test-taker ability*, *rater severity*, *test version*, *test mode*, *rating scale*), the lack of misfit across rating scales can be interpreted as indirect evidence of unidimensionality that "both modes are in fact measuring the same construct" (p. 10).

Nakatsuhara et al. (2021) conducted a follow-up study with larger sample size (99 test-takers, 10 examiners; note, however, 30 test-takers' performances were selected for the analysis of language functions) in a counter-balanced design. The researchers investigated the comparability of face-to-face and VC modes of operational IELTS tests in China. Nakatsuhara et al. highlighted that although scores between the modes are comparable, further investigation, such as language use, must be held, to resolve the "doubts on the equivalence of the construct measured in the two modes" (p. 371). The researchers used a modified version of O'Sullivan, Weir, and Saville's (2002) observation checklist for the analysis of language function *(i.e., informational functions, interactional functions, managing interaction functions*).

From the analysis of 4-facet (*test-takers*, *test versions, examiners*, *test delivery modes on each rating scale*) and 5-facet (*test-takers*, *test versions*, *examiners*, *test delivery modes*, *rating scale*) MFRM analyses, Nakatsuhara et al. (2021) reported the marginally lower scores in the VC mode than the face-to-face mode, which do not affect test-takers' final band scores. Their findings corroborate the previous research (Clark & Hooshmand, 1992; Kim & Craig, 2012; Nakatsuhara et al., 2017), that the VC mode resulted in statistically non-significant differences but marginally lower scores across analytic categories. Findings of language function indicated a slight change in the construct of oral communication ability, based on one significantly higher

use of functional output (i.e., *asking for clarification*) in the VC mode. Nakatsuhara et al. suggested more frequent use of such language use could be a signal of how people deal with communication breakdowns during the interaction in VC mode, showing that they are actively engaging in the communication (for more detailed discussion, see p.382). They concluded by highlighting the importance of further research focused on interactional competence in the VC mode.

Semi-Direct (Audio) vs. Direct (Face-to-Face) Modes

For more than 30 years, researchers have investigated the comparability of semi-direct and direct modes, because these modes have been widely used in both standardized and local speaking assessments. While the face-to-face mode is preferred because it represents most authentic oral interaction context, the practicality issues (e.g., cost-effectiveness, hiring examiners, time-efficiency) caused the use of semi-direct mode (i.e., raters assess audio-recordings of test-takers' performances) as an alternative by many testing centers and institutions (e.g., ACTFL's Language Testing International, Educational Testing Service, Center for Applied Linguistics, ITA Speaking tests at universities). While the studies summarized in Table 1 are all worth reviewing, I focus on a study by Zhou (2015), which is directly related to the purpose of this current study. Prior to reviewing Zhou's study, I first briefly review the studies described in Table 1.

Previous studies that compared semi-direct (absence of an examiner, usually audio-recordings are rated; *hereafter* audio) and direct modes (a face-to-face interview format between examiner and test-taker; *hereafter* face-to-face) reported lower scores in semi-direct (audio) mode than direct (face-to-face) mode (Conlan et al., 1994; Larson, 1984; Nambiar & Goon, 1993; O'Loughlin, 2001; Qian, 2009), except for Kenyon and Tschirner (2000) and Shohamy

(1994), who reported the equivalent scores across semi-direct (SOPI) and direct (OPI) modes. These mixed findings are different from the previous studies (e.g., Clark & Hooshmand, 1992; Craig & Kim, 2010; Kim & Craig, 2012; Nakatsuhara et al., 2017, 2020) that compared scores between face-to-face and VC mode, which showed equivalence in scores.

One possible reason for the score difference across the audio and face-to-face modes is that each mode taps into different range and aspects of the construct of speaking. For example, certain examiners/raters were likely to take more account of nonverbal features when assessing speech samples in the face-to-face mode than other raters. Such additional information of test-takers' nonverbal cue could have affected rater behavior, that led to higher scores in direct than semi-direct mode. While the focus of previous studies was on score differences across delivery modes, less has been investigated regarding whether visual information leads to greater variation in raters' scores.

So far, Zhou (2015) is the only recent study that compared (computer-delivered) audio mode and face-to-face mode regarding construct validity aspect. Zhou's basis for conducting exploratory factor analysis (EFA) was based on the hypothesized comparability between the modes, that is, test-takers' performance in semi-direct mode may not reflect their ability measured in face-to-face mode in which test-takers and examiner co-construct discourse through interaction (cf. Chapelle & Douglas, 2006; Lazaraton, 1996; McNamara, 1997). Zhou randomly assigned a total of 79 Japanese L1 students to two groups in a counterbalanced design, with an interval of seven to ten days between taking face-to-face mode (Group A) and the computer-delivered mode (Group B). The focus of rating was on analytic categories (grammar, vocabulary, fluency, pronunciation).

Zhou reported that from the one-way multivariate analysis of variance (MANOVA), the results showed no significant impact of mode on rating. Then, Zhou conducted EFA to investigate whether the audio (computer-delivered) and face-to-face modes measured common components. The 16 variables (4 analytic scales x 2 monologic tasks x 2 delivery modes) loaded highly on the factor (range from .78 to .88) which suggested that the analytic scales of tasks across the modes contributed similarly to the factor, that is, psychometric properties of speaking. Overall, Zhou concluded that the results of EFA proved the unidimensional factor structure across the two modes. However, Zhou also emphasized that this is an unexpected finding compared to the findings of previous research (e.g., Kenyon & Tschirner, 2000; Shohamy, 1994), potentially because Zhou used EFA while previous studies used paired *t*-tests. Zhou suggested future research to conduct confirmatory factor analysis (CFA) for (a) rigorous comparison of the monologic tasks across modes, and (b) further demonstration of the unidimensionality of the speaking construct in assessment contexts.

Audio-Only vs. Video-Recorded Modes

The earliest study that investigated the comparability of different recording modes was by Styles (1993), who had 30 IETLS test-takers and 3 examiners. In his study, the correlations for inter- and intra-rater reliability showed that noticeably higher values were observed in audio rating than video-recorded rating mode. However, as noted by Styles himself and Nakatsuhara et al. (2020), the findings should be interpreted with caution because the sound quality was poor in audio-recordings, and independent-measures research meant that test-takers with different abilities may have been assigned to the audio- and video-groups.

Recently, Beltrán (2016) investigated the comparability between audio and video-recorded delivery modes. Specifically, Beltrán's goal was to explore systematic effect of audio-

only and video-recorded modes (i.e., audio-with-video) on raters' behavior, such as their consistency and severity. By adopting a "quasi-experimental repeated measures single-group" and mixed-methods design, a total of 7 adult ESL learners' oral responses to monologic tasks were rated by 25 graduate students. The raters first assessed audio-only speech samples. After weeks of interval time, they rated the video-recordings. Beltrán focused on analytic categories: fluency, pronunciation, vocabulary, grammar, and meaningfulness. Once the rating was completed, raters responded to the questions about their perceptions of the two rating modes.

From the analysis using a paired samples *t*-test, Beltrán reported no statistically significant differences in the mean scores between the two rating modes, and concluded that the visual stimuli did not influence raters' scoring in a systematic way. The raters' responses to the questionnaire indicated that 76% of the raters preferred video mode, largely due to the ease of comprehension by watching the nonverbal elements (e.g., body language, facial expressions, attitudes and feelings). One rater preferred audio-only ratings because for them, nonverbal features were considered as a source of distraction, which made it more difficult for the rater to focus on the test-takers' performance. Although it was only one rater, nonverbal elements could be a source of potential bias that could affect scoring. Three raters had no preference.

While studies reviewed so far mostly focused on the comparison of two rating modes, specifically audio versus audio-with-video (i.e., the presence or absence of visual input in the rating process), two experimental studies in the following compared three different delivery modes; Lavolette (2013) compared video and two types of audio modes; Nakatsuhara et al. (2017) compared audio and two types of video modes. I further review each study below.

Video-Recorded vs. Audio-Only vs. Audio-Extracted from Video Modes

Lavolette (2013) examined the ratings of three types of speech samples: audio-recordings, video-recordings, and audio-extracted-from-video-recordings in the context of formative assessment (i.e., informal, learning activities with feedback provided to learners as instruction). Thirty-nine ESL learners' six speech samples (2 webcams, 2 with microphones only, 2 video track extracted from the webcam recordings) were rated by 15 teachers. Lavolette reported that the results of repeated-measures ANOVAs indicated different rating behavior across modes; raters gave significantly higher scores in audio files stripped from the video mode than audio-only and video rating modes. Such finding was in line with Kenyon and Malabonga's (2001) study, in which the authors reported raters' bias against video modes, but contrasted the findings of Nambiar and Goon (1993) who found lower scores in audio mode than face-to-face mode.

Lavolette's further investigation of test-takers' and raters' preferences for different modes revealed that for test-takers who preferred recording audio, they received significantly higher scores in audio extracted from the video mode than other modes. However, test-takers who preferred video recording had non-significant score differences across modes. Lavolette reported that for raters, their preference of rating audio recordings showed significantly higher scores than the video mode. Raters' preference for rating video modes didn't show significant differences across modes. Lavolette concludes that raters' bias was detected in her study. Specifically, when raters preferred assessing audio recordings, test-takers received significantly higher scores in audio files stripped from the video mode than the corresponding video mode. She further notes that because the recordings are test-takers' same performance, raters were biased against the visual stimuli. However, she stated that possible reasons for such biased behavior was unclear.

Audio-Only vs. Video-Recorded vs. Live (Face-to-Face) Modes

Lastly, Nakatsuhara et al. (2020), which is of particular interest in this current study, employed a convergent, parallel-mixed methods design to investigate the construct validity across different delivery modes. They compared rater behavior and raters' perceptions across three different delivery modes including audio and two different video modes (video-recorded, face-to-face). The goal of their study was to investigate the validity of double rating in which "a live examiner in the test and then a second rater who raters the recorded performance post hoc" (p. 2). Specifically, they investigated as to which aspects of test-takers' performance is more suitably assessed via different rating modes (audio-/video-recording formats, live rating).

Six trained IELTS examiners assessed 36 test-takers' performances and wrote justifications for their ratings. All examiners were assigned to sufficiently overlap with one another; one examiner carried out all types of rating (audio-recording, video-recording, face-to-face), two examiners assessed in face-to-face and audio rating modes, one examiner rated audio- and video-recordings, and another examiner rated only video-recordings. Once rating was completed, four examiners participated in verbal report sessions and watched four test-takers' audio- and video-recordings in two phases (phase 1: listen/watch the entire audio/video speech sample without pausing, phase 2: listen/watch the entire audio/video speech sample pausing whenever necessary). During each phase, examiners made general comments about test-takers' performance using stimulated recall methodology and gave scores.

Results of a 6-facet Multifaceted Rasch Model (MFRM) analysis (*test-taker*, *test version*, *examiner*, *test part*, *rating mode*, *rating criterion*) revealed that raters were harsher in the audio mode than in the face-to-face (live) and video-recorded modes. Nakatsuhara et al. interpreted that the lower score in the audio mode was observed because the audio-only rating condition

impose limitations (e.g., absence of visual information) on the assessment of test-taker performance. Regarding the finding of score outcome comparability across the face-to-face and video-recorded rating modes, Nakatsuhara et al. noted that "a remote video option may be acceptable in a context where a live face-to-face speaking test is not possible" (p.19). However, they highlighted the differential outcome for *Fluency*, where scores in face-to-face mode were slightly higher than the video-recorded mode. In sum, Nakatsuhara et al. concluded that the speaking construct measured in video-recorded mode is closer to the face-to-face mode than the audio mode, which assesses a narrower construct than video modes. The absence of visual information in the audio rating mode questions the construct validity, which limits the construct measured by semi-direct test formats.

The findings of examiner comments and verbal reports revealed that except for the face-to-face rating condition, examiners in two recorded formats noticed similar number of negative performance features but only led to lower scores in the audio rating mode. *Fluency* showed differential results in commentaries as well; under the video-recorded condition, there were slightly more negative features than under the audio condition. Lastly, examiners gave more negative comments in the recorded modes than the face-to-face mode. Nakatsuhara et al. interpret such difference by noting that in the recorded modes, examiners have no time pressure, no need to multitask compared to the face-to-face mode where the examiner plays a dual role (for further discussion, see p.19). In turn, examiners could attend their focus on negative features that they might have missed when serving as both interlocutor and rater. Overall, the visual information could potentially either serve as positive or negative source towards rater behavior and the final scores. The researchers highlighted the importance of standardizing the ways to

interpret visual information, and examiner training such as increasing the awareness about how to use verbal information.

Table 1.

Summary of empirical studies that compared different delivery modes of L2 speaking tests

	Authors (year)	Journal/research report	Tests	Participants	Data analysis method	Results (score comparisons only)
Semi-direct (audio) vs. direct (face-to-face)	Larson (1984)	Foreign Language Annals	German/Spanish: direct test (pronunciation, read-aloud, structured interview), semi-direct test (pre-recorded structured interview-type)	29 intermediate German students, 20 intermediate Spanish students	Correlation analysis, *t*-tests	Most students scored higher on the direct test than the semi-direct test
	Nambiar & Goon (1993)	RELC Journal	10-12 min. interview, a negotiation task	87 undergraduates	Independent t-test, paired t-test, correlation	Scores in audio rating was significantly lower than face-to-face rating
	Styles (1993)	A report on a project conducted at the British Council center in Brussels	IELTS Speaking test	3 examiners, 30 IELTS test-takers	Raw score data with Classical Test Theory (CTT) analysis	Audio mode produced lower mean score than the live scores
	Conlan, Bardsley, & Martinson (1994)	Unpublished study commissioned by the International Editing Committee of IELTS	IELTS Speaking test	3 examiners, 27 IELTS test-takers	CTT analysis & retrospective verbal reports from examiners	Audio recording had a band lower score than the live mode

Table 1 (cont'd)

Kenyon & Tschirner (2000)	Modern Language Journal	German OPI and SOPI	6 raters, 20 students	Spearmen rank order correlations, Pearson product-moment correlations, *t*-tests	Equivalent scores between OPI and SOPI
Kenyon & Malabonga (2001)	Language Learning and Technology	OPI, SOPI, computerized OPI (COPI) for Arabic, Spanish, Chinese	55 test-takers, no rater info.	Survey questionnaires on test-takers' opinions on modes	Test-takers preferred face-to-face OPI to SOPI and COPI, lower scores in face-to-face OPI than on a SOPI or COPI
O'Loughlin (2001)	Studies in Second Language Testing	The Australian Assessment of Communicative English Skills	20 test-takers (10 tape-mediated versions, 10 live version), no rater info.	Non-parametric factorial analysis	Lexical density was higher in the live version than the tape-based version
Shohamy (1994)	Language Testing	Hebrew OPI and SOPI		Paired *t*-tests for linguistic/discourse features	No significant differences
Qian (2009)	Language Assessment Quarterly	IELTS (direct), GSLPA*** (semi-direct)	186 university students	semi-structured survey, no statistical analysis	Most students preferred direct mode than semi-direct mode
Zhou (2015)	Language Testing in Asia	Monologic task type (narrative and opinion tasks)	5 raters, 79 Japanese students	One-way MANOVA, exploratory factor analysis (EFA)	EFA results: no differences in the underlying factor structure of the two modes

Table 1 (cont'd)

Face-to-face vs. video-conferencing (VC)	Clark & Hooshmand (1992) ****	System	Arabic and Russian tests administered as a regular end-of-course proficiency test at Defense Language Institute Foreign Language Center	16 instructors (examiners), 32 Arabic learners and 32 Russian learners	Paired *t*-tests, short questionnaires for both test-takers and examiners	Arabic: face-to-face rating was higher than screen-to-screen rating (non-significant) Russian: identical scores between two modes
	Craig & Kim (2010)	Multimedia Assisted Language Learning	Oral interview test	2 interviewers, 2 raters, 42 undergraduate students	Descriptive (score means), paired *t*-test (for survey data)	No significant difference between the modes
	Kim & Craig (2012)	Computer Assisted Language Learning	Oral interview test	2 interviewers, 2 raters, 40 undergraduate students	Paired *t*-test	No significant difference between the modes
	Nakatsuhara, Inoue, Berry, & Galaczi (2017)	Language Assessment Quarterly	IELTS Speaking test	2 examiners, 32 test-takers	CTT analysis, 4-facet and 5-facet MFRM analysis	No significant differences between rating modes
	Nakatsuhara, Inoue, Berry, & Galaczi 2021)*	Assessment in Education	IELTS Speaking test	10 examiners, 99 test-takers	4- and 5-facet MFRM analysis, language function analysis	Same score outcomes from face-to-face and video-conferencing modes
Audio vs. video-recordings	Styles (1993)	A report on a project conducted at the British Council center in Brussels	IELTS Speaking test	3 examiners, 30 IELTS test-takers	Raw score data with Classical Test Theory (CTT) analysis	Audio rating mode is noticeably higher than video-recorded rating mode

Table 1 (cont'd)

	Beltrán (2016)	Columbia University Working Papers in TESOL and Applied Linguistics	Final achievement speaking test (monologic tasks) at an adult ESL program at Teachers College, Columbia University	25 raters, 7 test-takers	Paired samples *t*-test, raters' feedback questionnaires	No significant difference between audio and video recordings
Audio vs. video-recordings vs. audio extracted from video	Lavolette (2013)	The International Association for Language Learning Technology	TOEFL iBT Test Independent Speaking prompts	20 graduate students (raters), 39 ESL learners	Repeated-measures ANOVAs	Audio-extracted-from-video rating was significantly higher than audio- and video-recordings
Face-to-face (live) vs. video-recording vs. audio-recording	Nakatsuhara, Inoue, & Taylor (2020)	Language Assessment Quarterly	Two retired versions of IELTS Speaking test	6 examiners, 36 test-takers	6-facet MFRM analysis (test-taker, test version, examiner, test part, rating mode, rating criterion), examiners' written commentaries	Scores in audio mode were significantly lower than scores in live and video modes

Studies within each category are listed in a timely manner.
*This study was conducted after the outbreak of COVID-19.
**SOPI: semi-direct performance-based speaking test that emulates the OPI in a tape-recorded format (Kenyon & Tschirner, 2000)
***GSLPA: Graduating Students' Language Proficiency Assessment-English, a performance-based English proficiency test developed at the Hong Kong Polytechnic University (Qian, 2009)
****This study compared face-to-face mode and screen-to-screen mode (i.e., examiner saw two test-takers' images of "face-on and from about the waist up" using the video camera transmitting. Test-takers were able to see themselves on screen; see p.296 for full description)

CHAPTER 6: RESEARCH GAP AND RESEARCH QUESTIONS

Research Gap

Within the L2 speaking assessment, ample researchers, with their studies reviewed so far, have investigated the comparability of the semi-direct (audio-recordings) and direct (face-to-face) delivery modes, while there have been relatively few research studies into comparing the different modes (i.e. audio, video ratings) of video-conferenced speaking tests. The recently accelerated use of video-conferenced platforms in local and standardized speaking tests warrant thorough investigation of whether different video-conferenced delivery modes (e.g., synchronous/asynchronous, displaying only test-takers or both test-takers and examiners in the video-recorded mode) leads to raters' behavior oriented to more positive or negative aspects of test-takers' oral performances related to the rubric (analytic, holistic). The comparison of different modes will provide empirical support of using the delivery mode that best captures test-takers' accurate L2 speaking ability.

In addition, raters' behavior across different delivery modes needs further inspection, because the newly used video-conferenced delivery modes include visual information which includes different ranges of information from the face-to-face mode, and have additional visual information that audio modalities do not include. Building on the findings of earlier studies that compared different delivery modes, I designed this current study to fill these gaps by investigating rater behavior and measurement of underlying L2 speaking ability in depth.

Specifically, I aim to investigate different video-conferenced speaking test delivery modes, to better understand how added visual information (a) brings variation to rater behavior and (b) represents the measurement of underlying L2 speaking ability. This study has the potential to demonstrate that the choice of the speaking test delivery modes should be carefully

employed, with the consideration of theoretical and practical consequences, because modes provide their unique benefits while these inevitably produce certain limitations.

This Study

Given the apparent lack of studies investigating (a) construct validity in video-conferenced speaking tests, and (b) how raters behave across different test delivery modes, it is important to understand to what extent currently administered video-conferenced speaking tests can provide information about test-takers' L2 speaking ability to gauge the degree of possible improvements needed. In this current study, I investigated the construct validity and comparability of different video-conferenced speaking test delivery modes based on previous studies (e.g., Nakatsuhara et al., 2017, 2020, 2021; Zhou, 2015) to ascertain the degree to which such delivery modes reflect test-takers' oral communication ability. I used two different datasets to address the following research questions through quantitative (RQ1 and RQ2) and qualitative (RQ3) analyses:

RQ1. What latent structure (i.e., delivery modes) of the video-conferenced speaking test best represent test-takers' oral performances? (datasets 1 and 2)

RQ2. Are there any differences in raters' scores when they rate test-taker oral performance under different rating conditions? (datasets 1 and 2)

RQ3. How do raters perceive test-taker performance under different rating conditions? (dataset 1)

CHAPTER 7: OVERVIEW OF METHODOLOGY

In this study, I used a convergent, parallel mixed-methods design (Creswell & Plano Clark, 2018). The quantitative data (based on dataset 1 and dataset 2) and qualitative data (based on dataset 1 only) were collected in two parallel strands, and each data strand is analyzed separately. I integrated the findings from the two strands for an in-depth and comprehensive understanding of the different delivery modes and rater behaviors. Figure 1 shows the structure and components of this research design, representing how I analyzed dataset 1 and dataset 2. Dataset 2 only contributed to the top half of the research design, as dataset 2 only comprised quantitative data. Figure 1 illustrates the data sources and analysis strands that were relevant to the research questions.

As mentioned above, this current study includes two different datasets which differ in context. Dataset 1 is based on an experimental context (a low-stakes testing situation). Each test-taker joined as a participant to experience and prepare their video-conferenced speaking tests in the future. Participants (N = 145) who contributed to dataset 1 were international students at American universities, whose participations were voluntary. Each student virtually met with me via Zoom[4], a video communication platform, and took a Zoom-administered speaking test. The participants' responses were audio- and video-recorded with their consent, and their speech samples were rated by 12 trained raters who were graduate students majoring in TESOL and

[4] Zoom (https://zoom.us) is a cloud-based video communication application that allows the users to set up virtual video and audio communication including conferencing, webinars, live chats, and screen-sharing. Pending the corona-virus crisis, such video communication platform can be downloaded and used in laptops and smartphones, which has now become the global standard for connecting with people virtually across countries. An issue regarding virtual communication such as non-verbal cues can be further checked in this New York Times review article, published in year 2020: https://www.nytimes.com/2020/04/29/sunday-review/zoom-video-conference.html

applied linguistics. Dataset 1 was collected a year after the outbreak of pandemic, that is, in spring 2021.

Dataset 2 stems from an actual testing context (a high-stakes testing situation). The test-takers (N = 285) were prospective international teaching assistants (ITAs) at a Midwest University, and they had to take a video-conferenced speaking test offered by an English language learning and testing program at the university in order to prove their English was high enough to qualify them as ITAs. The 12 raters in this dataset had professional experience rating ITA speaking tests either from face-to-face test modalities or audio-recorded ones; however, assessing speech performances in video-conferencing (VC; synchronous) or video-recorded (VR; asynchronous) modes were relatively new to them. Dataset 2 includes scores gathered for a year, which began right after the pandemic. To be clear, I collected the data from dataset 1, whereas the data from dataset 2 were ones that I borrowed from the operational testing program at the university's English language learning and testing program. This explains why only dataset 1 has corresponding qualitative data: I was able to interview raters and test-takers who partook in my data collection, but I did not have the opportunity to collect additional data from the test-takers and raters who contributed to the English language learning and testing program's data comprising dataset 2.

This study is unique in a sense that the two datasets include different test delivery modes, and the test scores from the datasets differed in terms of their real-world consequences and stakes for the test-takers. They also differed in that dataset 1 incorporated audio-recorded and video-recorded rating modes, whereas dataset 2 includes different formats of video modes (i.e., VC and VR modes). Three chapters are dedicated for findings based on analyses from each

dataset (Chapters 13,14, and 15), and qualitative findings from dataset 1 are reported in Chapter 16.

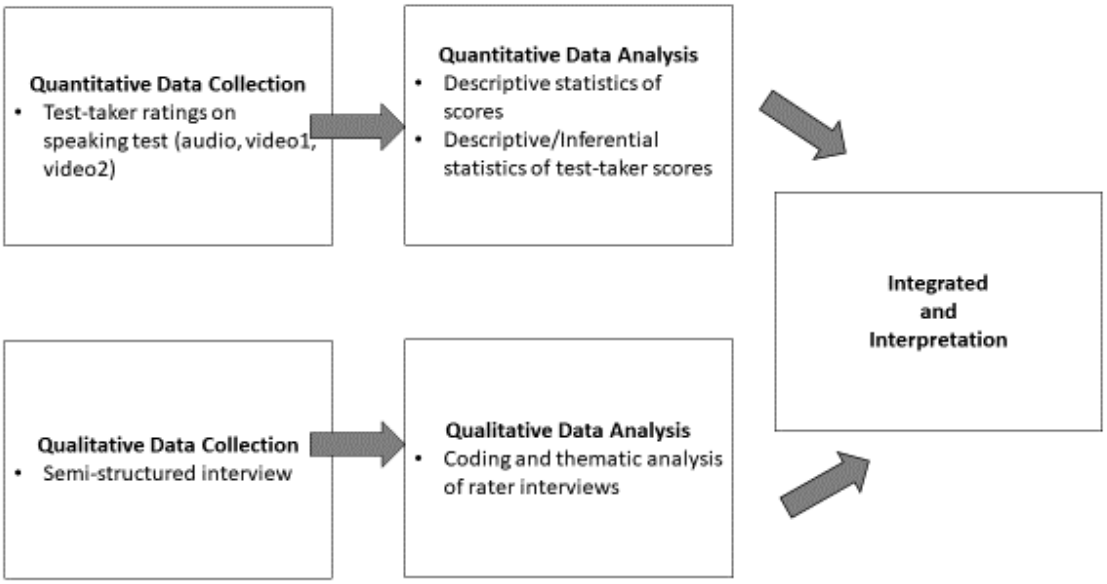


Figure 1. Overview of the data sources and analysis strands for dataset 1

CHAPTER 8: METHODOLOGY OF DATASET 1

Participants

Test-takers

After I piloted my materials and data collection protocol on five participants (and after I revised the procedures minimally based on the piloting), a total of 140 international students participated as test-takers in this study. Prior to rating sessions, I excluded 10 test-takers' data due to technical issues (e.g., internet connection instability which hindered the recording of a test-taker's visual information). After two rounds of rating, I excluded another 20 test-takers' data four raters who rated their speech showed inaccurate rating behaviors (i.e., extremely high/low scores for low/high proficiency test-takers, awarding the same scores across different proficiencies) that could potentially contaminate data analysis. As I will explain later, this deletion of participant data is actually the deletion of unreliable raters' performances (the scores they awarded); rather than the deletion of the participants from the study themselves. Nonetheless, a total of 110 test-takers' (male: n = 47, female: n = 27.28) responses were rated reliably, and I report those 110 test-takers' demographic information in this section.

Test-takers' mean age was 27.28, who were either undergraduate or graduate students studying in the United States. Their ages ranged from 18 to 51 (median = 27). According to the background questionnaire responses, more than half of the test-takers had lived in the United States for at least a year (3 years or more: n = 38; 1-2 years: n = 38), and more than two-thirds of the test-takers were graduates studying in the United States (n = 81). Most test-takers responded that they had studied English for at least 6 years before studying in the United States (n = 92). Based on their survey responses, I expected the participants to be comfortable communicating in English, their L2 (see Figure 2). About half of the students self-assessed their English

proficiency to be 4 out of 5 in perception skills (listening, reading), while they considered themselves less proficient in the production skills (speaking, writing). Note that each pie chart displays the proportion of test-takers' responses to L2 English ability.

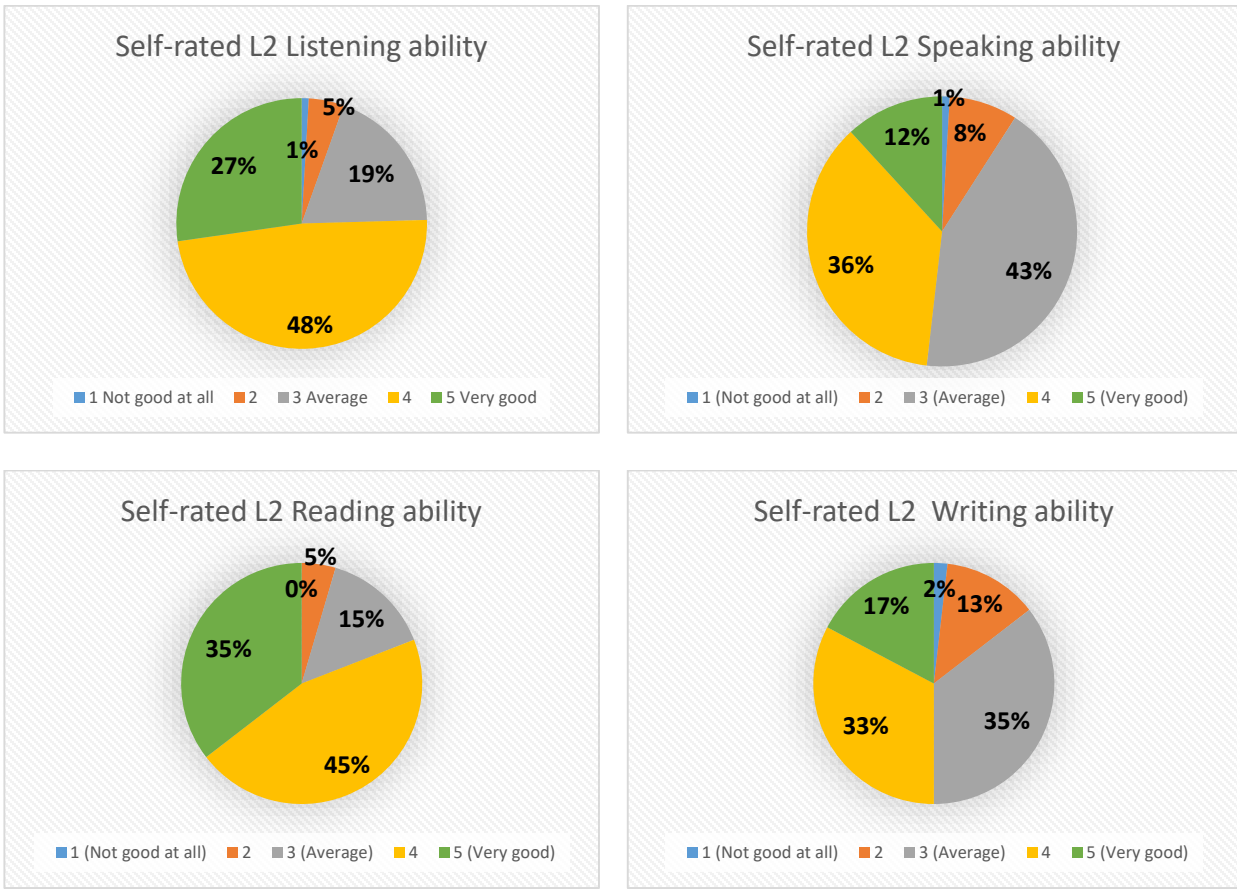


Figure 2. Pie charts of test-takers' self-rated proficiency of L2 English (N = 110)

Raters

Initially, a total of 12 raters participated in this study. The raters were graduate students majoring in applied linguistics who were enrolled in a language testing course at a Midwest university. While all twelve students voluntarily participated as raters, as I mentioned above, four raters' scores had to be excluded due to the raters' inaccurate rating behaviors. The exclusion was based on: (a) the average measure of ICC, (b) inter-rater correlation matrix (i.e., raters who showed correlation below 0.4 were considered as showing poor agreement with other raters), and

(c) 'Cronbach's alpha if item deleted' option from SPSS (i.e., if there was an increase in overall Cronbach's alpha if a rater was deleted, then it means that this rater showed poor agreement with other raters. Note, however, the rater in this case is not considered as showing poor rating performance). Thus, I present the remaining eight raters' demographic information, and these are the eight raters whose ratings were used for data analyses. The raters' first language varied, including native speakers of English (*n* = 3), Chinese (*n* = 1), Indonesian (*n* = 1), Korean (*n* = 1), Spanish (*n* = 1), Turkish (*n* = 1). Most of the raters were experienced English teachers; five raters had years of teaching English for 2 years, 4 years, 5 years, 6 years, and 10 years, respectively. I also asked about the raters' experience with technology, specifically video-conferencing platforms. Seven raters used computer for more than 30 hours per week. With regards to the purpose of using video chat, 6 raters responded for schoolwork, while 2 raters responded video chatting with friends and families.

Table 2.

Rater background characteristics (*N* = 8)

Gender		Age range		Teaching ESL/EFL		Hours of using computer in a week		Frequency of using computer to video chat*	
Male	Female	22-30 years old	31-45 years old	< 1 year	> 1 year	20-29 hours per week	30 hours or more per week	More than once a week	Every day
n = 3	*n* = 5	*n* = 5	*n* = 3	*n* = 2	*n* = 5	*n = 1*	*n = 7*	*n* = 5	*n* = 3

Note. *: In the questionnaire, provided examples were Skype, Zoom, Google Hangouts, WhatsApp, etc.

Materials

Background Questionnaire

In developing the background questionnaire survey, I partially adapted the survey questionnaires from the study by Nakatsuhara, Inoue, Berry, and Galaczi (2016) and added

questionnaires aimed for the test-takers and the raters of this study. The questionnaire was delivered via Qualtrics, an online platform. Test-takers were provided with questions about their first language, language learning experience (where, how much), educational background (degrees, majors), language-use experience (living abroad), test experience (type of English speaking test), and technology familiarity background (frequency of online learning and video-conferencing, reluctancy on using video-conferenced platforms) (for questionnaires, see Appendix A). Raters were also provided with the survey, with additional questions soliciting information concerning their teaching and rating experiences (Appendix B).

Test-Taker Speech Samples

All test-takers responded to a total of 8 speaking tasks, which I adapted from open-source ACTFL OPI speaking test items (see Appendix E for samples; sources can be found here: https://www.languagetesting.com/pub/media/wysiwyg/manuals/opi-examinee-handbook.pdf). The eight tasks were distributed in a counterbalanced design (Table 3). That is, four tasks with differing difficulty levels (i.e., intermediate, advanced, superior on the ACTFL (2021) Proficiency scale[5]) were provided within each test format. The tasks were estimated at those levels of difficulty by topic familiarity and test-takers' ability to justify their opinions and communicate about abstract ideas. Further description is presented in Appendix E.

Test formats were also counterbalanced, for example, if Format A was given in video-on mode (test-taker could virtually see the interlocutor), then Format B was given in video-off (audio-only) mode (test-taker could not virtually see the interlocutor). In this line, test delivery modes were naturally counterbalanced as well. If test-takers' IDs with odd numbers first

[5] https://www.actfl.org/sites/default/files/guidelines/ACTFLProficiencyGuidelines2012.pdf

responded to speaking tasks in a video-on mode, then, I turned off the screen and proceeded to the video-off mode. The order was in the reverse order for the test-takers with even numbers.

When test-takers were responding to the tasks on Zoom, I manually recorded their responses by clicking the "Record" button. Once I stopped recording, test-takers' recorded responses were automatically saved into two different formats (audio- and video-recorded).

For the video-on test format, two different modes were generated: (a) audio-recorded mode (sound files) and (b) video-recorded files that display the test-taker and me. For the video-off test format, two different modes were also generated: (a) an audio-recorded mode (sound files) and (b) a video-recorded files that displayed only the test-taker. Thus, each test-taker had a total of 16 recorded files (i.e., 4 tasks x 2 modes x 2 test formats). To prevent rater fatigue, however, I used only two tasks (i.e., intermediate-level and superior-level tasks) from each test format. That is, each test-taker's speech samples consisted of two video-on modes and two video-off modes. A total of 440 speech samples (i.e., 4 tasks x 110 test-takers) comprised dataset 1, which I used for analyses to answer the research questions in this study.

Table 3.

Example of counterbalanced design for two test modes

	Test-takers' IDs with odd numbers (ID 001, 003, …. 109)	Test-takers' IDs with even numbers (ID 002, 004, …. 110)
Format A (k = 4)	Video-on	Video-off (audio only)
Format B (k = 4)	Video-off (audio only)	Video-on

Note. Each format includes intermediate (k = 1), advanced (k = 2), and superior (k =1) level tasks.

Procedures

Interview with Test-Takers

Once the test-takers completed the speaking test, I virtually met each test-taker (one-on-one) for a 30-minute semi-structured interview over Zoom. All test-takers were asked the same questions in the same order (for interview questions, see Appendix C). Note, however, due to the

focus of this study to investigate mode differences and rater behaviors, test-takers' verbal data was not analyzed. Figure 3 below displays an overall summary of process test-takers went through.

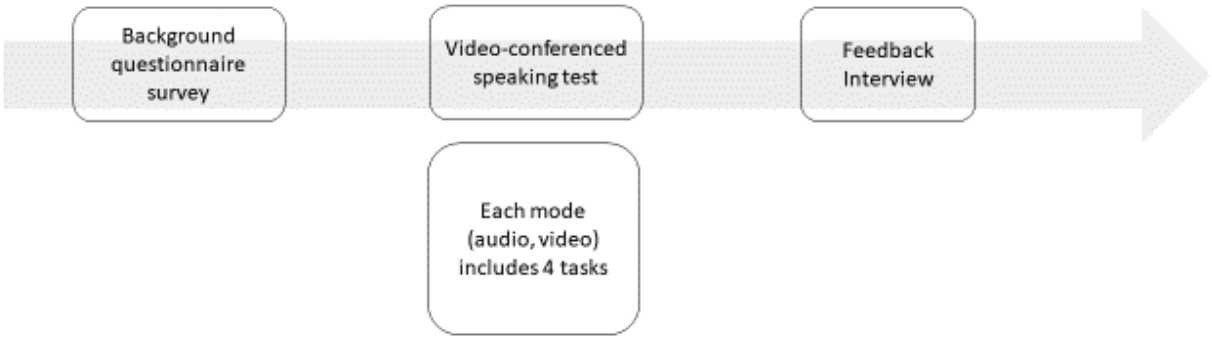


Figure 3. A visual summary of test-taking procedure

Rater Training and Rating Sessions

Raters were trained before they went through two rounds of rating. For training, I joined as a co-trainer of these participating raters in a language testing class they were taking. (Not all students in the class were volunteering for this research, but all in the class were trained in rating speech samples as part of the class.) For the rater participants in this study, the first three months were spent training and norming. The rest of the semester-long research project (which was approximately five months total) was used by the raters to complete two rounds of rating (Figure 7) with a three-week interval between the ratings of different modalities (audio and video versus audio only).

According to Eckes (2015), the purpose of training was to achieve raters' common understanding of (a) the construct being measured, (b) the levels of performance the assessment is aiming at, (c) the criteria and the associated descriptors that represent the construct of each performance level, (d) the categories of the rating scales, and (e) the overall difficulty level of the

tasks to which examinees are to respond (for detailed discussion about rater training, see p.41). Specifically, I used the training to guide the raters' assessment of the speech samples of different delivery modes; audio and video (video1, video2).

Once the raters were familiar with the rubric, and norming sessions were successfully completed, they underwent two rounds of rating (see Figure 4 for rating matrix). Note that during norming, the raters rated approximately 20 samples one-by-one, and we discussed the ratings openly after each rating, so that we could, over time, come to a high consensus on ratings. The consensus was achieved for individual speech samples through open discussion, which established a mind-share on how to rate and how to interpret and use the rating criteria and scale. Note that the crossed-off raters in Figure 4 were excluded from the analysis

Each rater was then assigned to operationally rate 30 test-takers' speech samples (k = 170). Raters were paired for the same set of test-takers' speech samples, which stayed the same throughout the rating sessions. While it is better to have a staggered rating matrix, I had the raters assess the same test-takers with the aim to see their changes in rating behaviors when only the test delivery mode differed.

The raters accessed the speech recordings via Qualtrics, and gave scores while listening or watching the recordings. The raters were not allowed to go back and change their previous ratings. In the first round of rating, raters awarded scores to video 1 delivery mode and audio-only delivery mode. In video1 mode, a screen displayed both a test-taker and an interlocutor's visual information with a proportion of 50:50 (Figure 5). From the raters' point of view, the test-taker was presented on the left rectangular part of the screen, and the interlocutor was on the right side. The raters were able to see both verbal and non-verbal interaction to some extent. The audio-only mode included the audio recording of a test-taker's response.

As briefly mentioned above, the second round of rating was conducted after three weeks. I did this in order to prevent raters from remembering the scores they gave to particular test-takers. This time, the video2 delivery mode was given, which displayed only test-taker's visual information. The interlocutors' screen was blacked out. The screen display was the same as in video 1 mode; 50:50 ratios. An example of is in Figure 6. Both screenshots (Figures 5 and 6) are used under the test-takers' permission.

The purpose of conducting two rating rounds was two-fold: (a) to examine whether the raters behaved differently depending on whether they were rating audio-only or video-recordings, and (b) within video-recordings, whether the raters showed different rating behaviors across different forms of video-recordings (i.e., with or without the interlocutor displayed on the screen).

		Round 1				Round 2	
		Audio (k = 40 for unique, k = 20 for anchor)		Video1 (k = 40 for unique, k = 20 for anchor)		Video2 (k = 40 for unique, k = 10 for anchor)	
Unique	Test-taker ID 001- ID 020	Rater A	Rater B	Rater A	Rater B	Rater A	Rater B
	~~Test-taker ID 021-ID 040~~	~~Rater C~~	~~Rater D~~	~~Rater C~~	~~Rater D~~	~~Rater C~~	~~Rater D~~
	Test-taker ID 041-ID 060	Rater E	~~Rater F~~	Rater E	~~Rater F~~	Rater E	~~Rater F~~
	Test-taker ID 061- ID 080	Rater G	~~Rater H~~	Rater G	~~Rater H~~	Rater G	~~Rater H~~
	Test-taker ID 081-ID 100	Rater I	Rater J	Rater I	Rater J	Rater I	Rater J
	Test-taker ID 100-ID 120	Rater K	Rater L	Rater K	Rater L	Rater K	Rater L
Anchor	Test-taker ID 121-ID 130	All raters					

Figure 4. Example of rating matrix for dataset 1

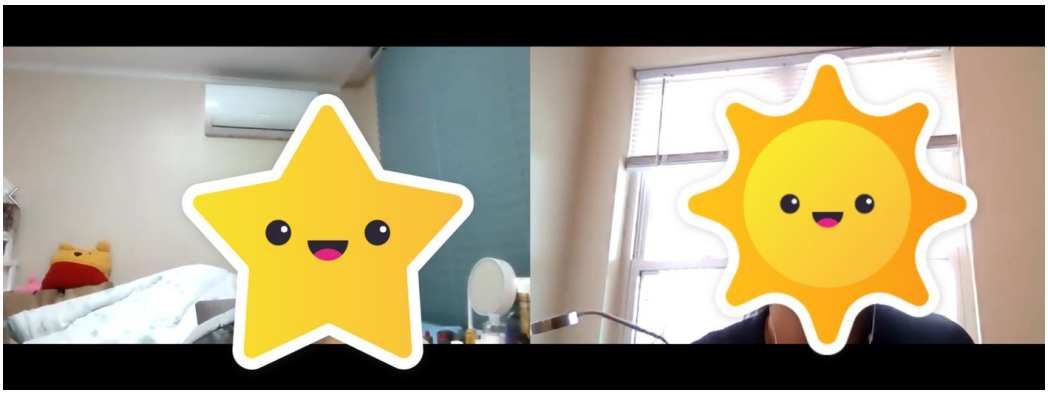


Figure 5. Screenshot of video1 delivery mode

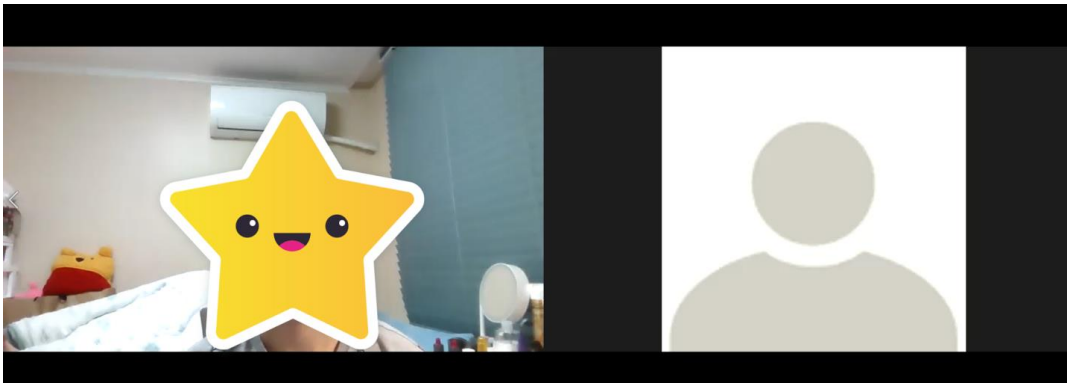


Figure 6. Screenshot of video2 delivery mode

Interview with Raters

Once two rating rounds were completed, I virtually met with each rater on Zoom for a 15-minute semi-structured interview. I asked four guided questions to the raters (see Appendix D). Their responses were video-recorded.

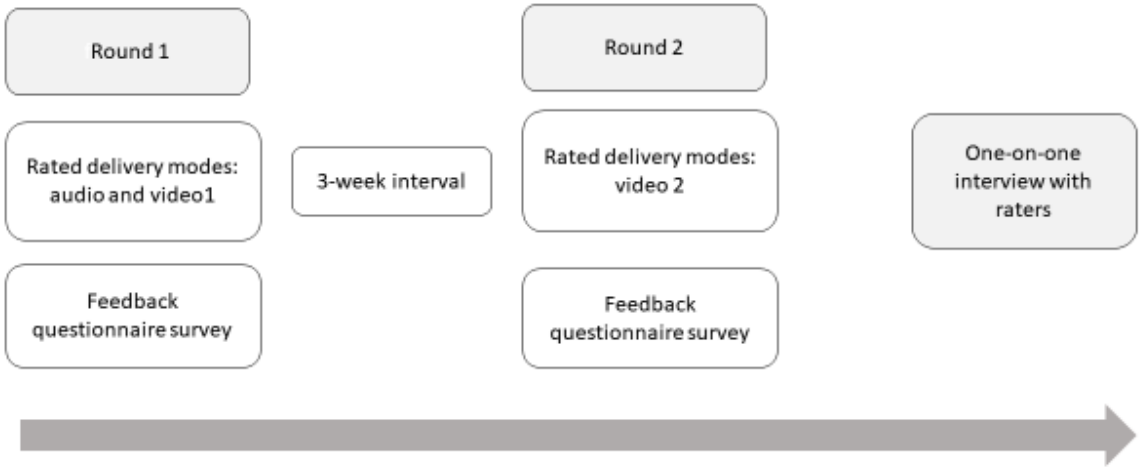


Figure 7. *A visual summary of the process of two rating rounds*

CHAPTER 9: METHODOLOGY OF DATASET 2

Testing Context

The second dataset was provided by an English language learning program (the unit in charge of providing language instruction and testing) at a Midwest University. This ITA speaking test makes three decisions according to the scores test-takers receive. First is pass, that graduate students with high enough scores on ITA speaking test (above 45 out of 50) or TOEFL are decided to teach as ITAs. Second is to have an appeal meeting with test-takers who are on boarder line, that is, score of 45 in ITA test. A test-taker can either receive a waiver, where there could be restrictions (e.g., only to do recitations, only TA in conditions where the TA is not in charge of teaching new content), or a full waiver, to teach as an ITA. Third, for test-takers with low scores, they are recommended to take English language classes that could improve their English language ability, particularly their speaking skills.

The dataset in this current study comprises scored item responses from 284 ITA perspectives who took an ITA speaking test in the midst of the pandemic. The test-takers were from 32 different majors, and approximately 236 were majoring in STEM fields (science, technology, engineering, math). Due to the protection of test-takers' personal information, no further information was provided.

Speaking Tasks

A total of five different forms, with 12 items for each form, were used for the test. The twelve items were designed to represent 8 different item types (see Table 4). The number of item types differed across test forms. For example, three test forms included item type 3 (Hypothetical) while two test forms did not have item type 3. In addition, in some forms, three items ($k = 3$) were from item type 1, while other forms had four items ($k = 4$) of item type 1.

Table 4.

Description of eight item types

	Description
Item type 1	Comparison
Item type 2	Supported opinion
Item type 3	Hypothetical
Item type 4	Comparison
Item type 5	Role play
Item type 6	Description/explanation
Item type 7	Graph presentation
Item type 8	Classroom announcement

Rating

Two different delivery modes were used in this dataset, namely the video-conferencing (*hereafter* VC) and the video-recorded (*hereafter* VR). Test-takers attended the synchronous live-Zoom ITA speaking test (VC mode), and interacted with one examiner who both administered and scored the exam. During the speaking test, the examiner read the script while zooming with the candidate and rated when the candidate responded. The examiner also video-recorded the candidates' oral responses by clicking 'Record' from Zoom and an additional rater later watched the video recording and provided a second score set. Thus, one test-taker had scores from two delivery modes, one from an examiner who gave scores as the candidate responded, and one from an examiner who assessed the candidate's oral response by watching the recorded video.

Figure 8 shows the rating matrix for dataset 2. Different from dataset 1, the 12 raters in this test were professional trained and certified, and they were randomly assigned to rate different test-takers. For example, Rater A participated in a VC test delivery mode and assessed a test-taker's (ID 001) oral performance as an examiner, and assessed another test-taker's (ID 007) recorded oral performance in a VR delivery mode as a rater. When the score difference between two raters in these two modes for each test-taker showed a discrepancy of 7 or higher, a third

rater was assigned. Note that each test-taker had responses for 12 tasks, and only the holistic scores are reported. For the final data analysis, only two pairs of raters with a discrepancy lower than 7 were selected.

	VC	VR
ID 001	Rater A	Rater B
ID 002	Rater C	Rater D
ID 003	Rater E	Rater F
ID 004	Rater G	Rater H
ID 005	Rater I	Rater J
ID 006	Rater K	Rater L
ID 007	Rater B	Rater A
ID 008	Rater D	Rater C
ID 009	Rater F	Rater E
ID 010	Rater H	Rater G
ID 011	Rater J	Rater I
ID 012	Rater L	Rater K
….		

Figure 8. Example of rating matrix for dataset 2

CHAPTER 10: DATA ANALYSIS

In light of the theoretical standpoints, previous studies, and the two datasets, three research questions guided this study. I analyzed datasets 1 and 2 to answer research questions 1 and 2 using statistical analyses methods. For dataset 1, I further explored raters' perspectives toward test delivery modes by answering research question 3. Below are the research questions for each dataset.

Dataset 1

Data Analysis for Research Question 1

To examine the latent structure of L2 oral communication ability in this speaking test, I used confirmatory factor analysis (CFA). I used CFA to confirm the relationships among measurable variables (i.e., observable variables or indicators, such as scores) and their latent variables (i.e., factors; here, delivery modes) (Brown, 2015; Kline, 2016) by testing three proposed hypotheses regarding the relationships: that the variables form one-factor (Model A); comprise two factors, (Model B), or three (Model C). The three models are explained in full after I present the research questions and in section "Hypothesized models".

I chose to use CFA because the tasks for this speaking test were based on Oral Proficiency Interview-Computer (OPIc), which was developed for test-taking situations where an examiner (an avatar) asks questions, and a test-taker responds to them. Regardless of delivery mode (audio prompts and audio recordings of test-takers' speech, or video-conferenced test delivery formats), or rating mode (interlocutor as rater, or raters who listen to or watch and listen to recordings), the construct is considered to be a single one. This assumption, that no matter the format, an L2 oral communication test measures, quite simply, L2 oral communication, does not consider any differences to the measured construct, even when the test is a video-conferenced

speaking test that requires the presence of an examiner (note, however, tests like *TOEFL iBT Home Edition* do not require an examiner)[6]. The general assumption is that a test-taker will obtain the same scores regardless of testing modality if the L2 oral communication tasks are the same (single-factor model; see Figures 9, 11, and 13). Thus, as one of the next steps to confirm the validity argument (i.e., proposed interpretation and use of test scores), I assessed the construct validity by associating test-task scores with underlying factors (i.e., delivery modes) in two other ways (correlated two-factors models; see Figures 10, 12, and 14; correlated three-factors models; see Figure 11). The examination of this test's three proposed underlying structures could provide evidence or counterevidence for a statement that the video-conferenced speaking test measures L2 oral communication ability equally in three test delivery modes (audio, video1, video2).

Another reason for using CFA is to provide convergent and discriminant validity by modeling an underlying structure using latent factors, that is, delivery modes (Kane, 2006; M. Kim & Crossley, 2020). Convergent evidence is provided if the observed scores of the underlying structure are expected to load on the appropriate same latent factor. On the contrary, discriminant validity is provided if the observed scores that represent distinct characteristics load on different latent factors. Again, for further clarity, the hypothesized models are explained in full after the research questions.

To conduct CFA, I used a statistical software package R (R Core Team, 2021) and *lavaan* packages (Rosseel, 2012). In the models, latent variables (i.e., delivery modes) were displayed in ovals, and observed variables (i.e., scores) in squares. The latent variables were fixed at 1.0 when

[6] Further information of each test can be found here: OPIc: https://www.languagetesting.com/test-delivery-logistics and *TOEFL iBT Home Edition*: https://www.ets.org/toefl/test-takers/ibt/take/at-home

evaluating the model, to compare the factor loadings for each indicator variable. In other words, estimates of the indicator variables on the latent variables were examined.

To evaluate the model fit statistics for selecting the optimal model, I checked the following model fit indices:

(a) Santorra-Bentler chi-square (SBχ^2): mean-adjusted chi-square

(b) Comparative fit index (CFI): the region of .90 but values above .95 reflect a good model fit (Hair et al., 2010)

(c) Root mean square error of approximation (RMSEA): recommended to be around .07 or less

(d) Standardized root mean residual (SRMR): same as RMSEA, recommended to be around .07 or less

(e) Akike's information criterion (AIC): measure of comparative fit. This index is closely related to BIC. χ^2 is used to compare the non-nested models (Kline, 2016).

(f) Bayesian information criterion (BIC): also measure of comparative fit.

It should be noted that the fit guidelines are dependent on the complexity of the model (how parsimonious the model is) and the sample size (Hair et al., 2010).

Data Analysis for Research Question 2

Data Analysis:

To understand the role of delivery modes and test conditions, I calibrated the scores of audio, video1, and video2 rating conditions with the Multifaceted Rasch model (MFRM) using Facets 3.83.6 (Linacre, 2021). I first examined an overall picture of the score results and then conducted a bias/interaction analysis. MFRM is a way to assess the internal structure of the measurement of a construct that is unidimensional, which here is L2 oral communication ability;

that is, all items/prompts on the test are proposed to measure the same underlying construct of L2 oral communication. Similar to regression analysis, MFRM can estimate how much various fixed facets that should not impact the speaking test scores (like the individual rater; raters should be interchangeable; or the individual test modality; they should be interchangeable) actually do impact scores.

For an overall picture of the score results and the facets that are part of the scoring system, I performed a rating scale model analysis rather than partial credit model, because the 9 bands across the analytic scales of the rubric (IELTS Speaking band descriptor) were designed to be comparable with the original IETLS Speaking used for holistic rating (cf. Nakatsuhara et al., 2020, Taylor & Falvey, 2007). In this current analysis, I used five facets as potential sources for score variance: *test-taker* (S001-S110), *test delivery mode* (audio, video1, video2), *rater* (A-H), *task difficulty* (intermediate, superior), *rating criterion* (fluency, lexis, grammar).

Then, I dummied out the *delivery mode* facet and conducted bias/interaction analysis between the *delivery mode* and *rater* facets. The purpose of this analysis was to examine whether and to what extent the raters interact with the three delivery modes. For example, whether rater X had a bias when rating video modes as detected by the scores and *p*-values. Dummying the *delivery mode* facet anchors the logit value at 0, that is, this facet was not used for estimation. Followed pairwise comparisons were run by Facets with the residuals from those interactions.

Data Analysis for Research Question 3

Data Analysis:

To complement the findings of inferential statistics, I adopted both deductive and inductive coding approaches: I partially adopted a coding scheme from previous research

(Nakatsuhara et al., 2017, 2020), and inductively coded the themes that emerged from the raters' verbal report. I first transcribed eight raters' verbal data and then generated a code book.

In a preliminary coding, I first transcribed and iteratively extracted the emerging themes and sub-themes. The final coding scheme included the themes from previous studies (e.g., negative and positive experience, Nakatsuhara et al., 2020) and themes that uniquely emerged from the current raters' verbal reports. The final coding scheme included 5 main themes with a total of 14 sub-themes. The code book is presented in Appendix F.

Using a code book, a second coder (a graduate student studying in the United States) did a second coding. I first went through the code book with the second coder, and we discussed the final codes to work through. Once the second coder completed a practice coding and got familiar with using the codebook, the second coder coded all the transcripts. Two cases of disagreements were resolved in the second round of coding: (a) when there were more than two codes emerged within the same sentence, we treated them as different codes, and (b) when the highlighted range for each code differed, the second coder and I adjusted the extent to which the segment should be coded. The inter-coder reliability (92.97%) was examined using the most sophisticated option available from MAXQDA (i.e., minimum code overlapping of x% at the segment level option is used).

Dataset 2

For this dataset, I analyzed two different types of data; analyses for scores at item-level, and analyses for scores of item types. Two research questions were answered for each data type.

Data Analysis for Research Question 1

RQ1. What latent structure (i.e., delivery modes) of the ITA Speaking Test best represents test-takers' oral performances?

Data Analysis:

Identical to RQ1 in dataset 1, I conducted CFA to answer this question.

Data Analysis for Research Question 2

RQ2. Are there any differences in examiners'/raters' scores when they rate test-taker oral performance under VC- and VR-rating conditions?

Data Analysis:

Identical to RQ2 in dataset 1, I conducted MFRM to answer this question.

CHAPTER 11: Hypothesized Models for CFA

While research questions 1 and 2 are based on the model-to-data fit approach, CFA particularly requires hypothesized models that are based on previous studies in terms of theoretical standpoint. Hence, I present the hypothesized models for each dataset below.

Dataset 1

For dataset 1, I conducted three competing hypothesized CFA models to determine which model would best represent the latent structure of the speaking test. I briefly discuss each model below.

Single-Factor Model

The observed variables (fluency, grammar, vocabulary) were specified as loading on a single factor (L2 English oral communication ability). As presented in Figure 9, this model indicates that this speaking proficiency test is unidimensional (i.e., no distinction is made among the three delivery modes concerning the construct being assessed). This model is the most predominate one in the field of language testing, and has been discussed and proposed by researchers who investigated the construct validity of different types of speaking test (e.g., TOEFL Junior Speaking test for adolescents: Huang, Bailey, Sass, & Cheng, 2020; CEFR-based Examination for the Certificate of Competency in English: Kim & Crossley, 2020; Michigan English test: Liu et al., 2022; TOEFL iBT test: Sawaki & Sinharay, 2017; ITA Speaking test: Yan et al., 2019). Note, however, that these studies have focused on different language skills (e.g., writing, listening, speaking, reading) as latent variables. While language skills assessed have been investigated using structural equation modeling, no studies have investigated the dimensionality of the underlying speaking construct in terms of test delivery modes.

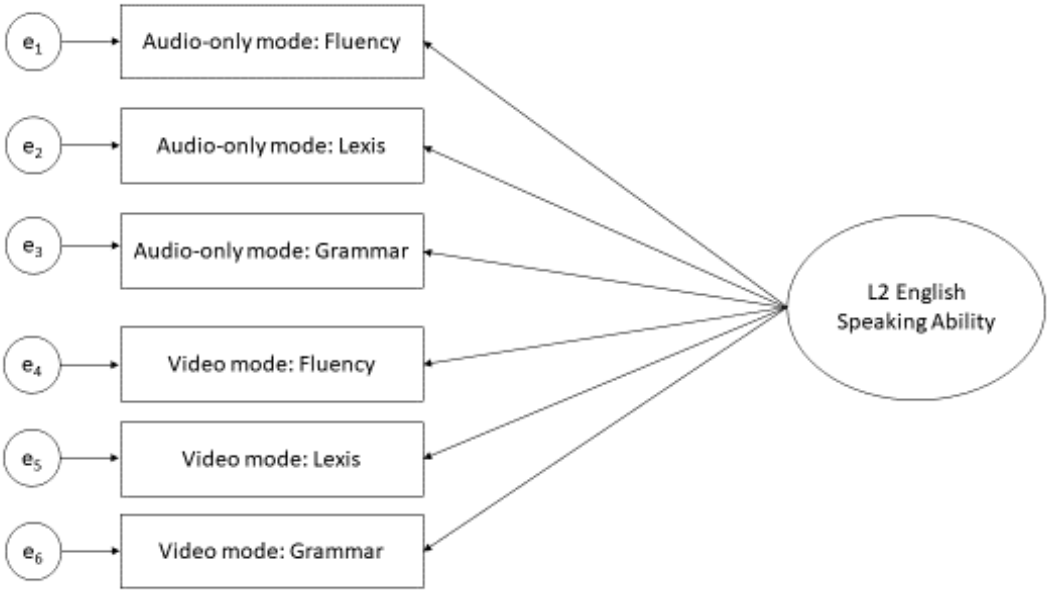


Figure 9. Hypothesized single-factor model

Correlated Two-Factor Model

Two correlated and potentially distinct factors – audio and video delivery modes – are specified in this model (Figure 10). This model is partially based on previous research works (e.g., Nakatsuhara et al., 2020) which suggested that audio and video test delivery modes show statistical differences in scores, but are unidimensional when it comes to the underlying construct. However, it should be noted that most previous studies used multifaceted Rasch model analysis (MFRM) to investigate the underlying structure of the measured speaking construct, and so far, no studies have used CFA to investigate the underlying speaking construct, specifically regarding the different test delivery modes. The hypothesis for this model is to suggest multidimensionality of the speaking test: a test-taker's English speaking ability in the audio and video delivery modes are distinct from one another and result in distinct scores.

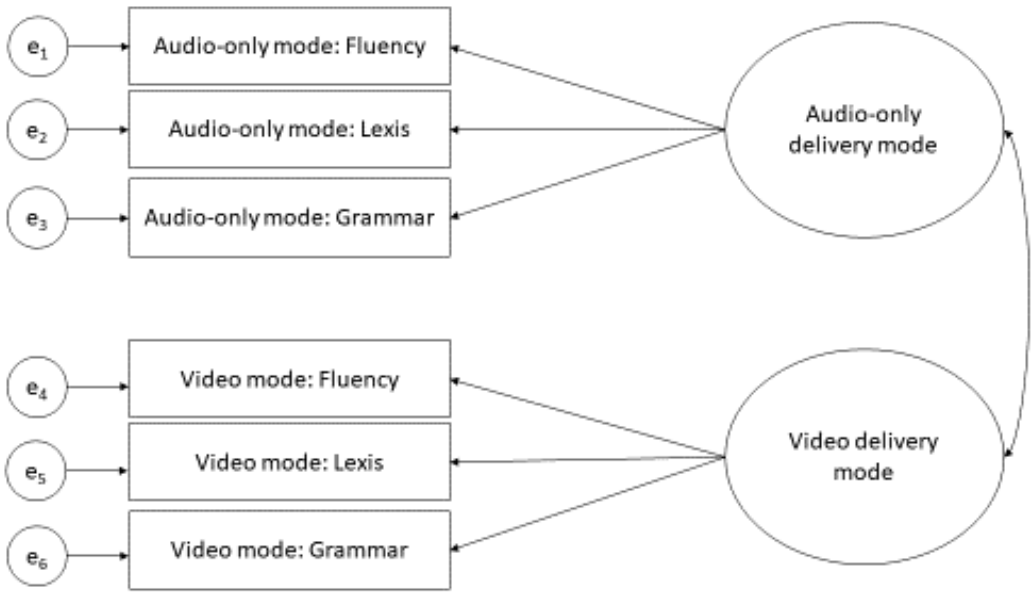


Figure 10. Hypothesized correlated two-factor model

Correlated Three-Factor Model

In this model, three correlated and potentially distinct factors – audio, video1, and video2 delivery modes – are specified (Figure 11). The observed variables were specified as loading on each factor. This model supports partial multidimensionality in that three delivery modes of an L2 oral communication ability test are correlated but distinct from one another. To confirm this hypothesized interrelationship, the factor inter-correlations should be below 0.80 (Kline, 2016). It is important to note that, for the purpose of this study, which is to test arguments (or hypotheses) that support the construct validity of this test, factor inter-correlations are expected to be above 0.80 with poor discriminant validity. That is, the latent factors (delivery modes) should not be distinct with one another.

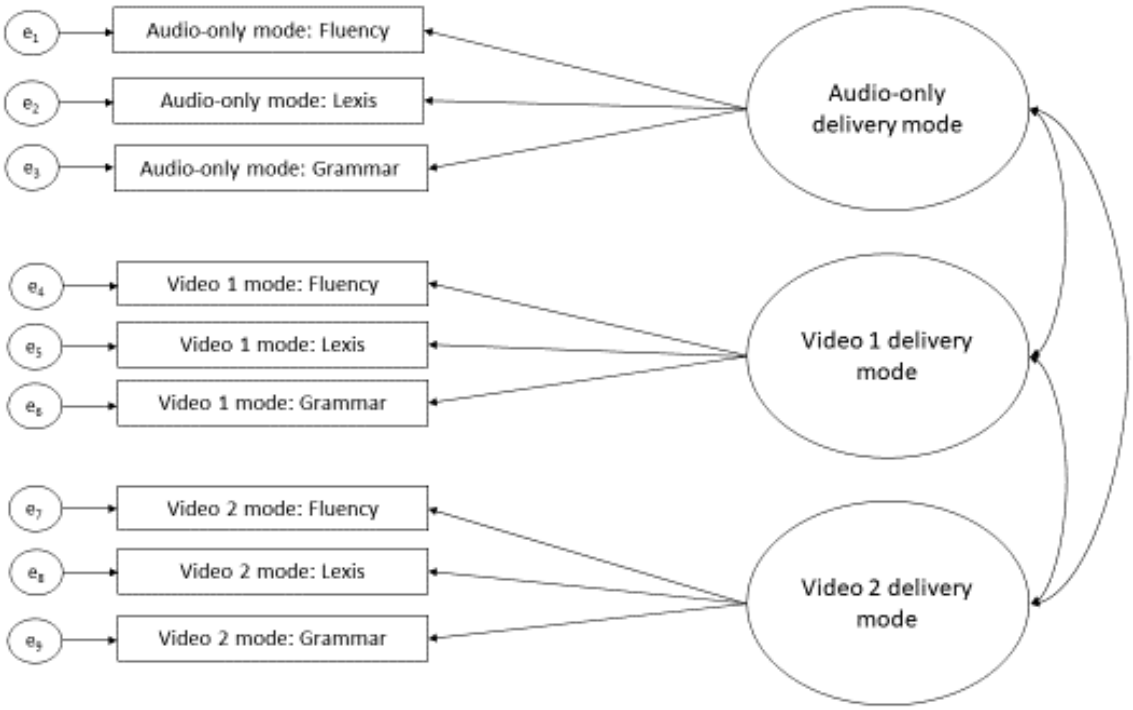


Figure 11. Hypothesized correlated three-factor model (dataset 1)

Dataset 2

Items

Single-Factor Model

The observed variables (holistic scores of 12 items) were specified as loading on a single factor (L2 English oral communication ability). This model indicates that this speaking proficiency test is unidimensional (i.e., no distinction is made among the two delivery modes concerning the construct being assessed). The model is presented in Figure 12. Note that for all hypothesized models in this dataset, VC indicates video-conferencing (synchronous) delivery mode, and VR indicates video-recorded (asynchronous) delivery mode.

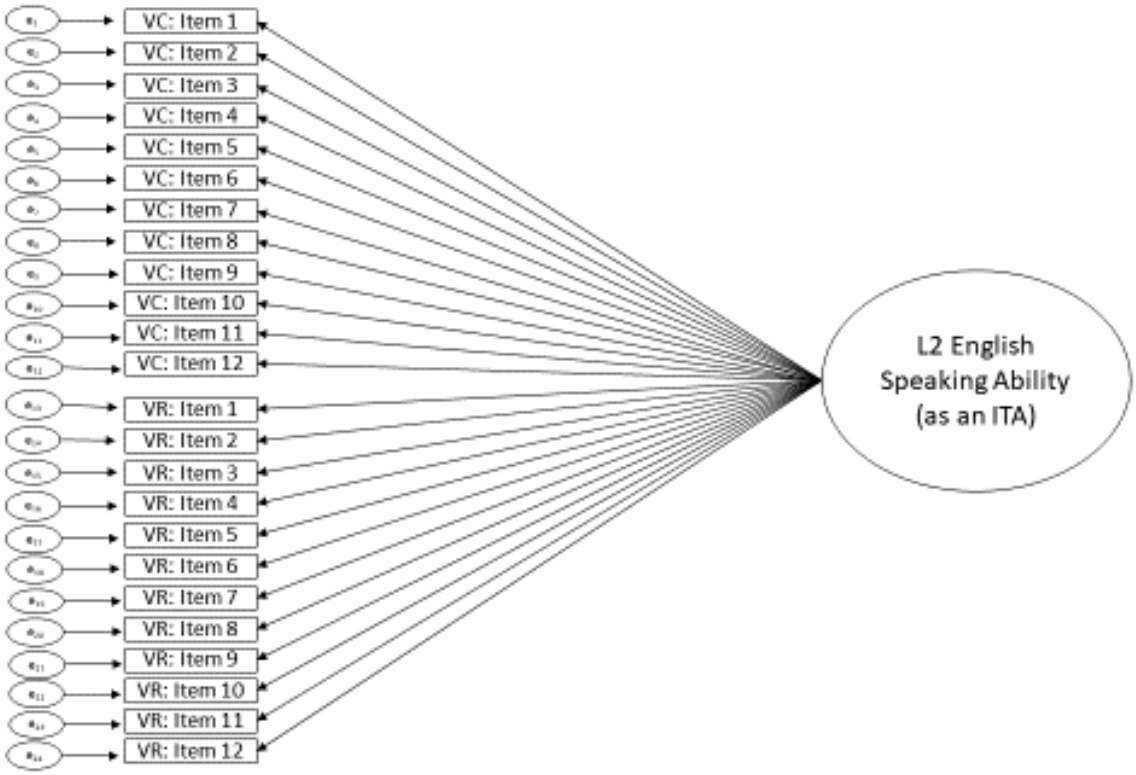


Figure 12. Hypothesized single-factor model (item, dataset 2)

Correlated Two-Factor Model

Two correlated and potentially distinct factors – VC and VR delivery modes – are specified in this model. The observed variable (holistic scores of 12 items) was specified as loading on each factor. The hypothesis for this model is to suggest multidimensionality of the speaking test: VC and VR delivery modes are distinct from one another. Identical to dataset 1, the factor inter-correlations are expected to be above 0.80 for the purpose of this study.

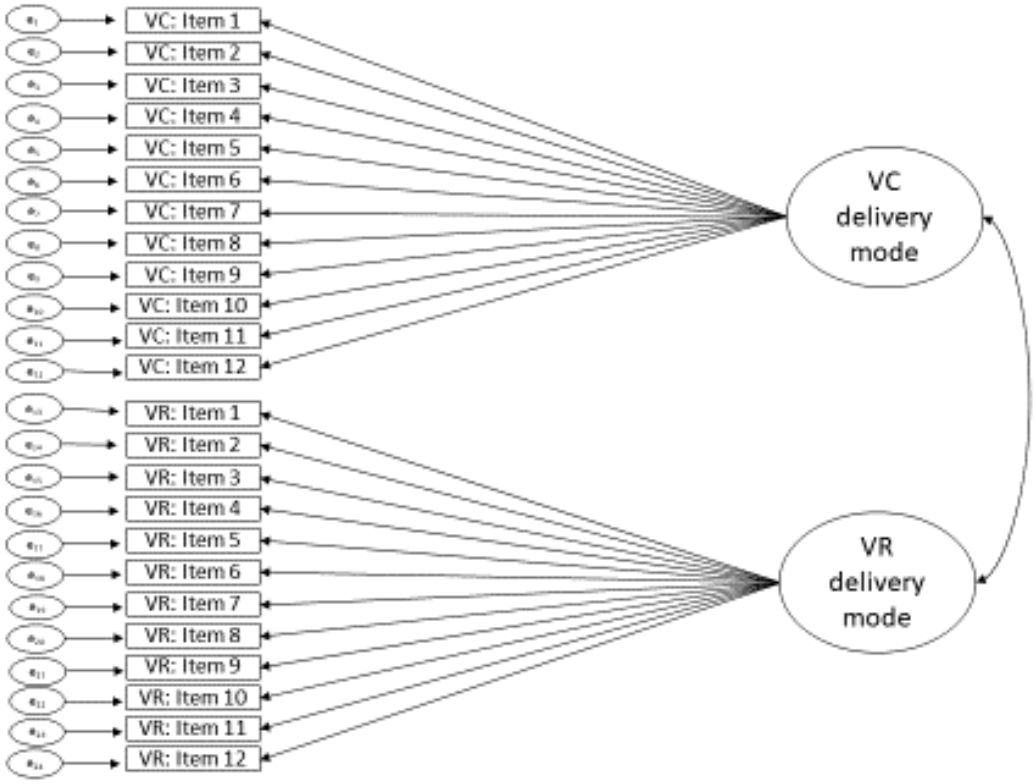


Figure 13. Hypothesized correlated two-factor model (item, dataset 2)

Item Type

Single-Factor Model

The observed variables (holistic scores of 8 item types) were specified as loading on a single factor (English oral communication ability). This model's estimates indicate that this

speaking proficiency test is unidimensional (i.e., no distinction is made among the three delivery modes concerning the construct being assessed). The diagram is presented in Figure 14.

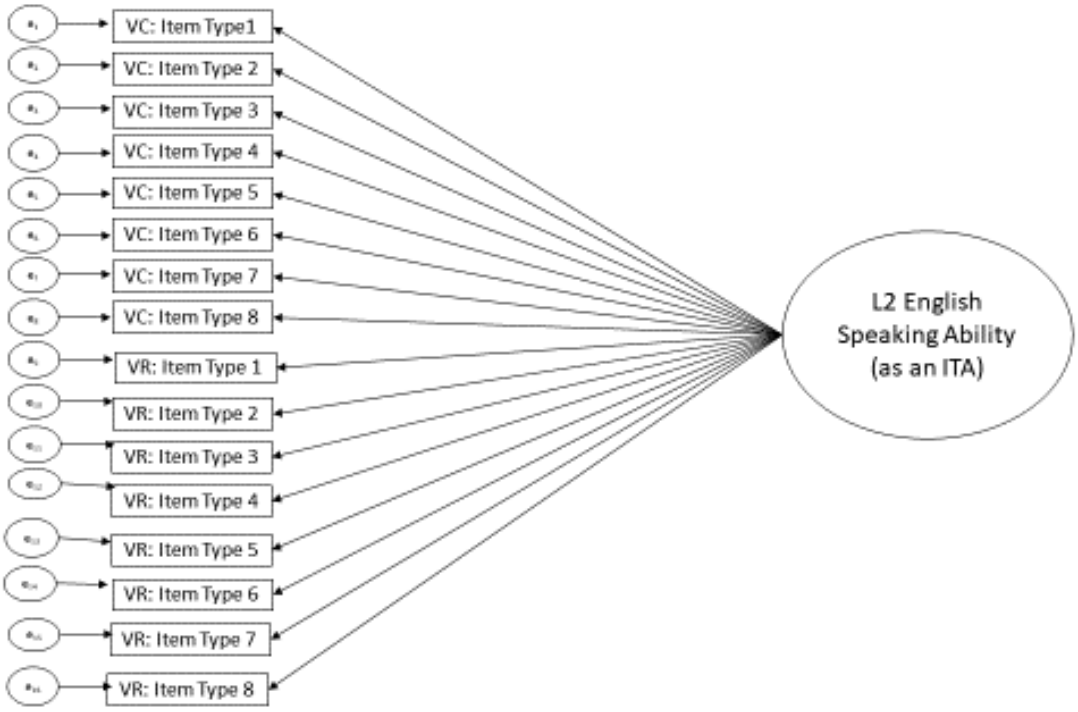


Figure 14. Hypothesized single-factor model (item type, dataset 2)

Correlated Two-Factor Model

Two correlated and potentially distinct factors – VC and VR delivery modes – are specified in this model. The observed variable (holistic scores of 8 item types) was specified as loading on each factor. The hypothesis for this model was confirmed and suggest multidimensionality of the speaking test: VC and VR delivery modes are distinct from one another.

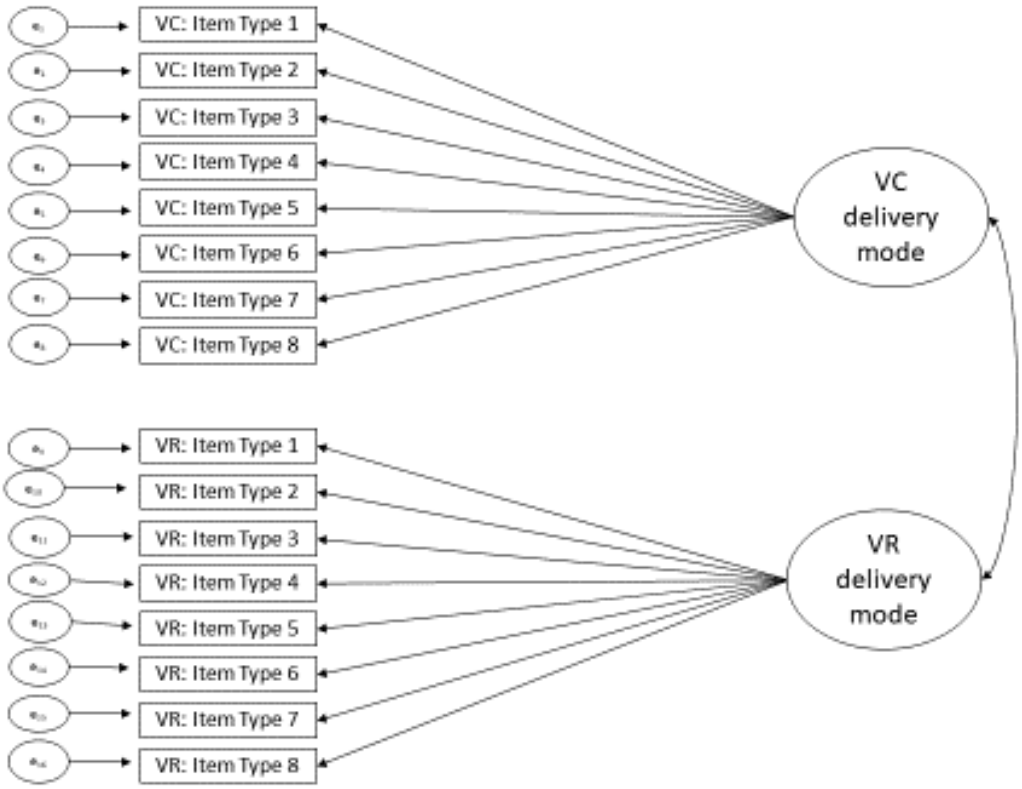


Figure 15. Hypothesized correlated two-factor model (item type, dataset 2)

CHAPTER 12: RESULTS OVERVIEW

The following results provide information on the distribution of the experimental video-conferenced speaking test (dataset 1) scores and ITA Speaking Test (dataset 2) scores, reliability of scores, variation and covariation of scores, and detailed Rasch analysis summary statistics of test facets. The information reported are considered as evidence to inspect the measurement invariance for test delivery modes in a video-conferenced speaking assessment.

The results are largely divided into three chapters. In Chapter 13, I report on the data results of the first dataset. Then, in Chapter 14 I report on the results of the second dataset (ITA Speaking Test), focusing on the analysis at item-level. In Chapter 15, I report the results of item type from dataset 2. Lastly, in Chapter 16, I present the findings of verbal report from dataset 1. In each chapter I outline explicitly which research questions are being addressed by the chapter.

Table 5.

Overview of statistical results for two datasets

	Research question	Data analyzed	Statistical method
Chapter 13	Do the tasks (intermediate-level, superior-level; *hereafter* I-L and S-L) of different test delivery modes (audio, video 1, video2) measure the same construct (L2 oral communication ability) and have equivalent relationships to this construct in the three modes of video-conferenced oral English proficiency test?	Dataset 1 (experimental data)	CFA
	What are the effects of test delivery mode (audio, video1, video2) and test conditions (test-taker ability, rater severity, rating categories, task difficulty) on the ratings of the speaking test?		MFRM
Chapter 14 (item-level), Chapter 15 (item type-level)	Do the items of different test delivery modes (VC, VR) measure the same construct (L2 oral communication ability) and evidence equivalent relationships to this construct in the two modes of ITA Speaking test for which the measurement will be used?	Dataset 2 (ITA Speaking test data)	CFA
	What are the effects of test delivery mode (VC, VR) and test conditions (test-taker ability, rater severity, item difficulty) on rating of the ITA Speaking test?		MFRM

Table 6.

Overview of qualitative analysis results for dataset 1

	Research question	Data analyzed	Qualitative method
Chapter 16	What are the raters' perceptions toward the test delivery modes when rating the speech samples?	Dataset 1	Thematic coding

CHAPTER 13: RESULTS OF DATASET 1

Results

Prior to computing inferential statistical analyses, I examined the distribution of the scores using means, skewness, and histograms. The mean values in Table 7 show that for all rating categories (*fluency*, *vocabulary*, and *grammar*), the video2 delivery mode (i.e., screen displaying only test-taker) had the highest scores, followed by the video1 delivery mode (i.e., screen displaying both test-taker and examiner) and the audio delivery mode. Within each task type, I-L tasks had *fluency* as the highest score for all three delivery modes, while the video2 mode showed the same means for both *fluency* and *grammar* (M = 6.30). For S-L tasks, *grammar* had the highest means. In addition, I-L tasks had Lexis as the lowest means while S-L tasks had *fluency* as the lowest means, with the video1 having the same mean values for both *fluency* and *vocabulary*.

Followed by this was the examination of skewness and kurtosis (Table 8). A general rule-of-thumb was used to interpret the data distribution: if the skewness was between -2 and +2, and the kurtosis is between -7 and +7 (Byrne, 2010; Hair et al., 2010), I considered the data as normally distributed. Specifically, the skewness near zero (between -0.5 and +0.5) indicate that the distribution is approximately symmetric. Table 8 shows that all three test delivery modes have both skewness and kurtosis within the suggested range. Thus, I considered the test scores of each item type within each mode as normally distributed. I generated histograms for visual inspection (Figures 16, 17, and 18). Since histograms are strongly influenced by sample sizes when determining the data shape, I added normality curves (thin blue lines).

One-Way ANOVA

I conducted one-way ANOVA to examine whether there were any statistically significant differences across the three delivery modes (audio, video1, video2) in terms of the averaged scores of fluency, vocabulary, and grammar.

The results showed there was a statistically significant effect of test delivery modes on *fluency* ($F(2, 327) = 5.43$, $p = .01$). A Scheffe post-hoc analysis revealed the significant difference between the audio and the video1 delivery modes ($p = 0.05$), the audio and the video2 delivery modes ($p = 0.01$). Compared to the mean score of fluency in the audio delivery mode ($M = 5.91$), the two video delivery modes showed higher fluency scores (video1: $M = 6.27$, video2: $M = 6.37$).

Second, the findings of *lexis* also showed the significant impact of test delivery modes ($F(2,327) = 5.51$, $p = 0.004$). Post-hoc analyses indicated a significant difference ($p = 0.007$) between audio and video 2 modes, that audio mode had significantly lower score ($M = 5.87$) than video2 delivery mode ($M = 6.34$).

Lastly, test delivery mode had significant impact on *grammar* ($F(2,327) = 5.12$, $p = 0.006$) as well, with post-hoc results that showed scores in audio delivery mode ($M = 5.95$) significantly lower than video2 delivery mode ($M = 6.40$).

In sum, the results of one-way ANOVA indicate that audio delivery mode had a significantly lower score than video delivery modes in all three linguistic categories, and the major difference occurred between the audio and video2 delivery modes.

Table 7.

Descriptive statistics for video-conferenced oral proficiency test

	Audio mode				Video1 mode				Video2 mode			
	I-L		S-L		I-L		S-L		I-L		S-L	
	N	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)	*N*	*M* (*SD*)
Fluency	110	5.88 (1.09)	110	5.93 (1.15)	110	6.26 (1.20)	110	6.29 (1.11)	105	6.30 (1.15)	105	6.42 (1.20)
Lexis	110	5.75 (1.04)	110	6.00 (1.07)	110	5.99 (1.20)	110	6.29 (1.13)	105	6.21 (1.08)	105	6.45 (1.08)
Grammar	110	5.82 (0.98)	110	6.08 (1.10)	110	6.16 (1.18)	110	6.31 (1.11)	105	6.30 (1.06)	105	6.49 (1.14)

Note. The rating scale is from minimum score of 0 to maximum score of 9. I-L and S-L each indicate intermediate-level speaking task and superior-level speaking task. The linguistic categories are presented in one keywords due to the limitation of space. Fluency denotes “fluency and coherence”, Lexis denotes “lexical resource”, and grammar denotes “grammatical range and accuracy”.

Table 8.

Distribution of scores

	Audio mode				Video1 mode				Video2 mode			
	I-L		S-L		I-L		S-L		I-L		S-L	
	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis
Fluency	0.17	-0.83	0.1	-0.47	-0.15	-0.41	0.16	-0.25	0.14	-0.41	-0.14	-0.16
Lexis	0.20	-0.61	0.23	-0.56	0.17	-0.40	0.35	0.09	0.21	0.15	0.10	-0.35
Grammar	0.33	-0.34	0.13	-0.52	0.15	-0.54	0.30	-0.09	0.26	-0.06	0.00	-0.24

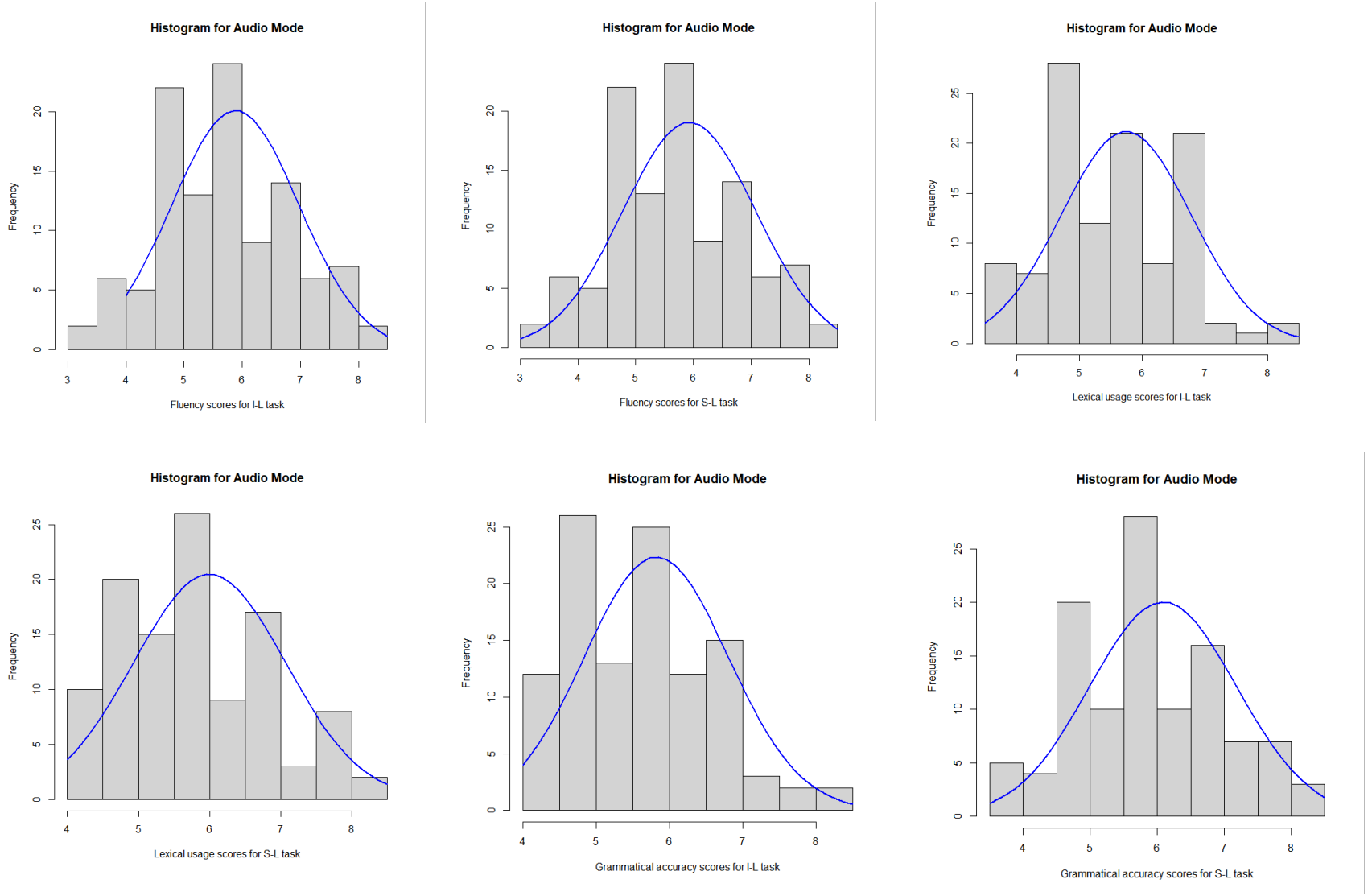


Figure 16. Histograms representing the distribution of the scores from audio delivery mode

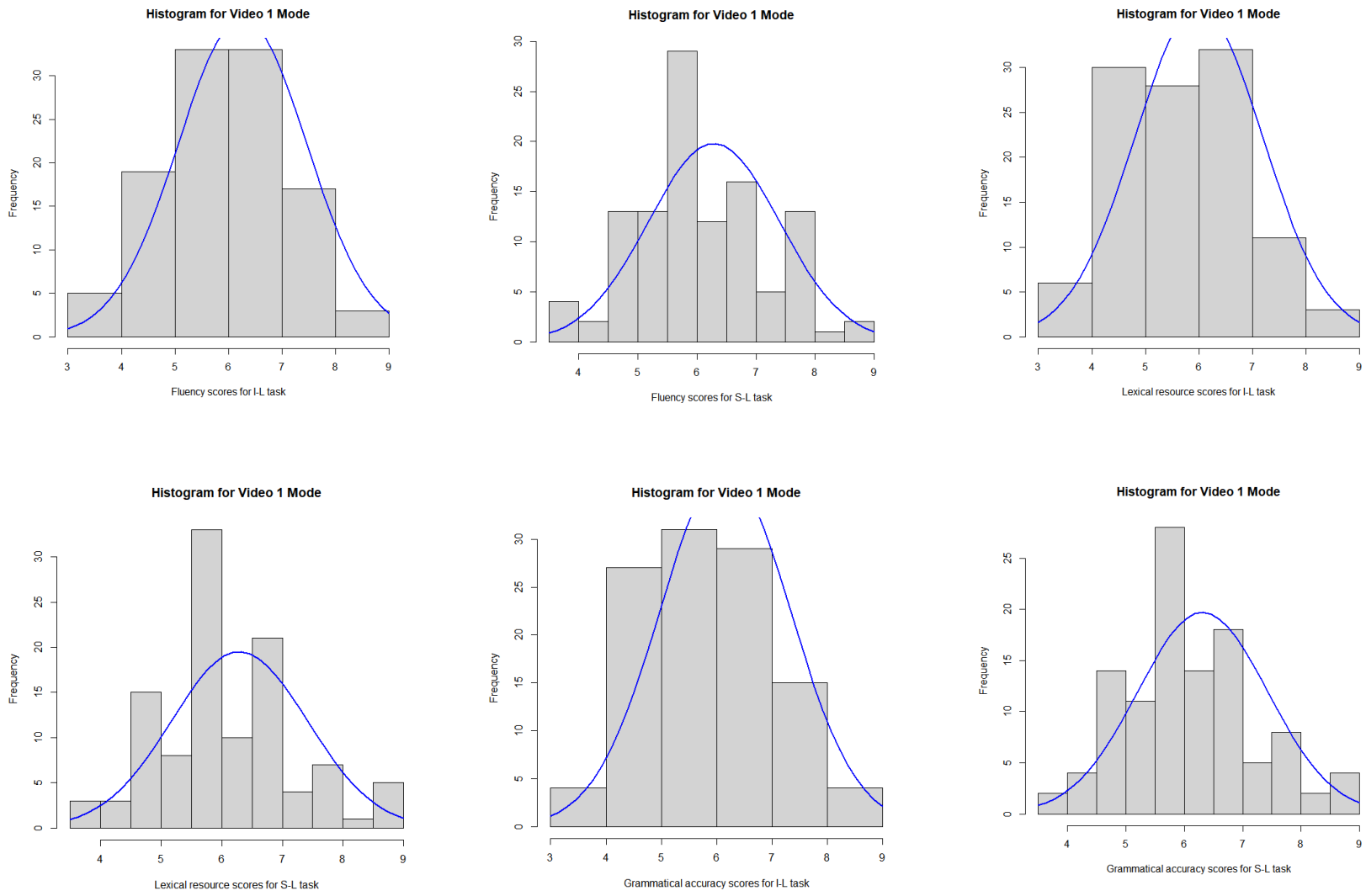


Figure 17. Histograms representing distribution of score categories from video1 delivery mode

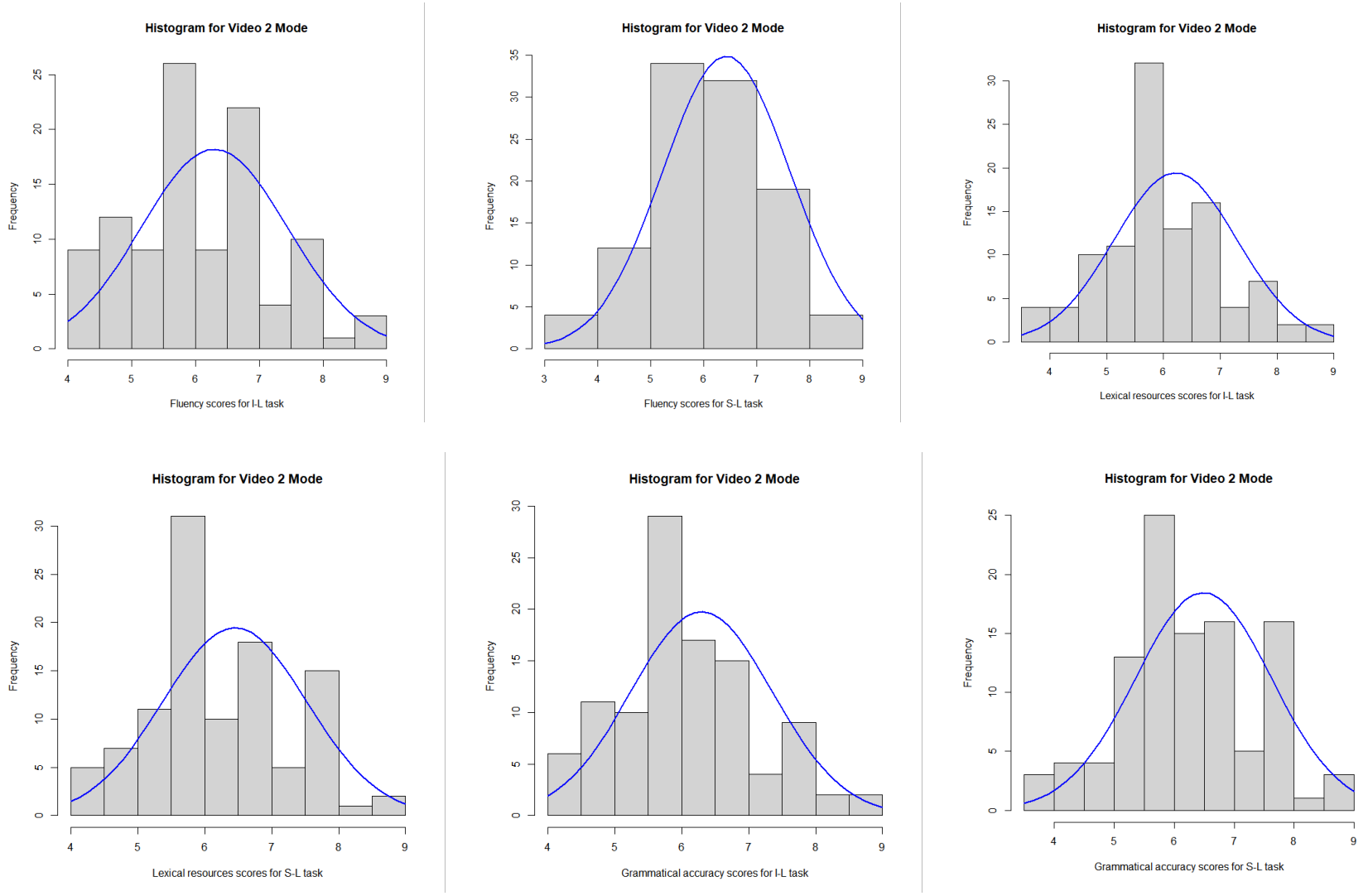


Figure 18. Histograms representing distribution of score categories from video2 delivery mode

Intra-Class Correlation Coefficient

Following the descriptive statistics, I used SPSS version 27 to check internal consistency of all speech ratings, using average measure intra-class correlations (ICCs) separately by test delivery mode (audio, video1, video2). The ICCs values of the video modes showed excellent reliability. The audio delivery mode showed the lowest but good reliability, as displayed in Table 9.

Table 9.

Inter-rater reliability for speech ratings across 8 raters by test delivery mode

Audio-only mode		Video1 mode		Video2 mode	
ICC	95% CI	ICC	95% CI	ICC	95% CI
0.89	[0.78, 0.94]	0.91	[0.83, 0.95]	0.94	[0.89, 0.97]

Confirmatory Factor Analysis (CFA)

The purposes of the CFAs were to compare the rated linguistic categories (*lexis*, *grammar*, *fluency*) within the video-conferenced speaking proficiency test and across different test delivery modes (audio, video1, video2). In this sub-section, I report the correlation matrices for the indicator variables and fit statistics for the hypothesized models.

Correlation Matrices

Table 10 shows the correlation matrices among indicator variables (i.e., scores for each rating category of each mode) for CFA. The correlations of rating categories for each test delivery mode are stronger than other modes. For example, linguistic categories of the audio mode (A-F, A-L, A-G) have correlations above 0.9, while correlations with other test modes are less strong. Overall, all indicator variables showed strong correlations with each other with coefficients ranging from 0.79 to 0.96.

Table 10.

Correlation matrices for indicator variables (N = 110)

	1	2	3	4	5	6	7	8
1 A-F	1							
2 A-L	0.94	1						
3 A-G	0.95	0.96	1					
4 V1-F	0.86	0.87	0.84	1				
5 V1-L	0.83	0.83	0.81	0.94	1			
6 V1-G	0.84	0.86	0.82	0.96	0.96	1		
7 V2-F	0.82	0.81	0.83	0.81	0.80	0.79	1	
8 V2-L	0.83	0.84	0.86	0.84	0.84	0.84	0.94	1
9 V2-G	0.83	0.82	0.86	0.84	0.83	0.83	0.95	0.97

Note. All correlation coefficients are significant at $p < .001$. The first initials represent the test delivery mode; A (audio), V1 (video1), V2 (video2). Followed acronyms are: F (fluency), L (lexis), and G (grammar). For example, A-F stands for fluency scores in audio mode; V2-L stands for lexis scores in Video2 mode.

Model Fit Statistics

Next, I computed CFA to test the overall model fit of the three hypothesized models. As presented in Table 11, the results of the CFA indicated the single-factor model (Figure 19) had poor fit (low CFI and high RMSEA), while the correlated two-factor and correlated three-factor models had excellent fit. However, it should be noted that for the correlated two-factor model, the CFI value of 1 and RMSEA value of 0 do not indicate that the model has a perfect fit. The CFI value of 1 happens when χ^2 is less than its expected value (the *df*). Also, when the sample size is small ($n < 200$) as in this dataset, RMSEA is stated to be positively biased (Curran et al., 2003).

In the correlated three-factor model (Figure 21), the three latent variables (audio, video1, video2) showed correlations that do not go over 0.9 (Kline, 2016): (a) audio and video1 (r = .85), (b) audio and video2 (r = .87), and (c) video1 and video2 (r = .84). However, the correlated two-factor model (Figure 20) showed correlations over 0.9 (audio and video: r = .91). While inter-correlation value of 0.8 indicate that the model has no discriminating function (Kline, 2016), which is the case for both models, the correlated three-factor model is chosen as the best fitting model. Regarding the inter-correlation values, it is important to note that the inter-correlations across latent variables above 0.8 imply indicate that the model has weak discriminating function (Kline, 2016).

Table 11.

Fit statistic for the three models

Model	SBχ^2	*df*	CFI	SRMR	RMSEA	AIC	BIC
Single-factor	1387.657	15	0.852	0.054	0.452	752.163	800.771
Correlated two-factor (audio, video)	1387.657	15	1.000	0.006	0.000	546.332	597.641
Correlated three-factor (audio, video1, video2)	2020.725	36	0.994	0.011	0.070	986.706	1067.720

Note. All models are standardized

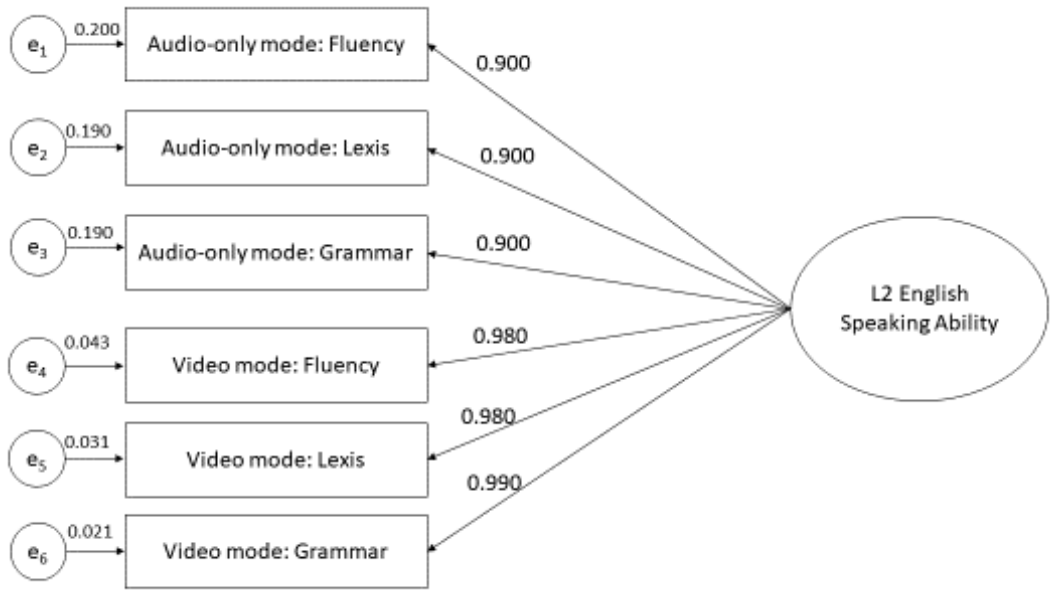


Figure 19. Single-factor model

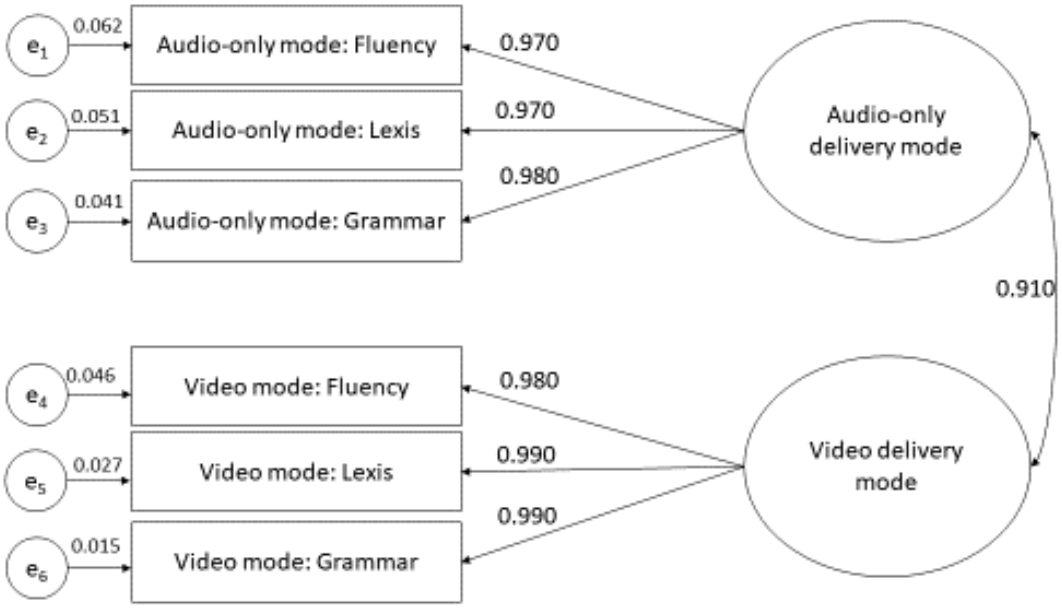


Figure 20. Correlated two-factor model (audio, video)

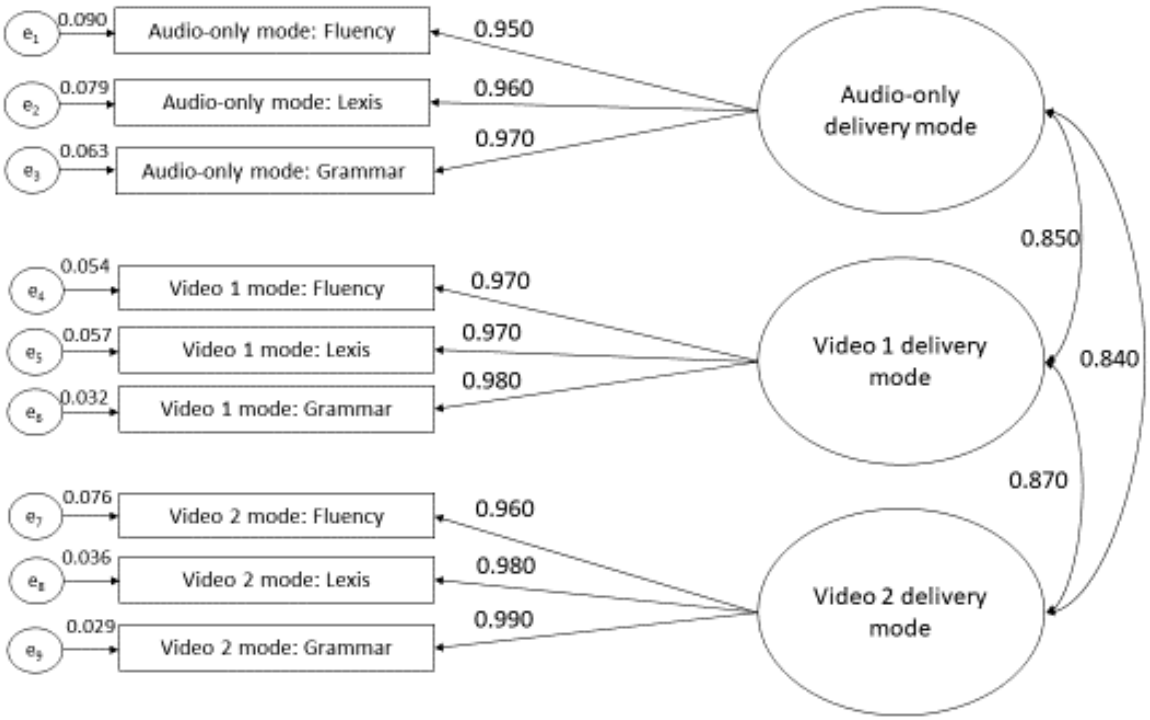


Figure 21. Correlated three-factor model (audio, video1, video2)

Description of the Final Model

The final model, correlated three-factor model (Figure 21), has audio, video1, and video2 as latent variable constructs. The oral communication ability within each test delivery mode were measured using three linguistic categories with a scale of 1 to 9 that load on each latent factor. The factor inter-correlations are above 0.80 (audio and video: $r = 0.85$, video1 and video2*:* $r = 0.87$, audio and video2: $r = 0.84$), which indicates poor discriminant validity (Kline, 2015). In other words, the delivery modes of this speaking test are not clearly distinct from one another, thus this model does not support the unidimensionality of the test. Linguistic category error variances are indicated by parameters $\varepsilon 1$- $\varepsilon 9$ in a small circle.

Results of Multifaceted Rasch Model (MFRM) analysis

MFRM was conducted using FACETS software program (Linacre, 2021). The analysis was performed with 2,316 measurable responses. From the data summary, the mean standard

residual (Resd) and the mean standardized residual (StRes) were 0.00, and the standard deviation (SD) was 0.99, which indicate that the estimation was successful (Linacre, 2021). The current data's variance explained by Rasch measures was 69.72%.

In this section, I first interpret the FACETS variable map (i.e., the Wright map, which displays measures graphically), followed by facets summaries and summary statistics for each facet. Then, I report empirical and expected item characteristics curves (ICCs) and category probability curves (CCCs) for the speaking test indicators. Lastly, I report the bias/interaction between test delivery modes and raters.

Description of the FACETS Variable Map

I first describe the FACETS variable map (*Figure 22*), which presents variables and key information of the data analyzed. From left to right of the map:

(1) The *first* column ("Measr") displays the logit scale, a reference for all facets. The scale ran from 6 logits to -5 logits.

(2) The *second* column ("Test-takers") displays the 110 students' estimated speaking proficiency. Each star represents one student (* = 1). The test-takers' tendency of receiving low or high scores across the four facets (*raters, speaking tasks, test delivery modes*, and *rating categories*) are summarized using a single number on the logit scale. This is possible because "test-takers" is the only non-centered facet that does not have a local origin. That is, the measure is relative to the origins of other facets which allow the test-taker facet to "float" relative to other facets. The "+" on the column heading represents that the more score implies more measure. The higher-scoring test-taker is placed on the top whereas the lower-scoring test-takers is on the bottom of the column. (range: 6.39 to -4.84 logits)

(3) The *third* column ("Raters") shows the 8 raters' severity or leniency in rating the test-takers' oral responses. The raters' tendencies to award higher or lower scores on average could be estimated because more than one rater scored each test-taker's response. In contrary to *test-takers* facet, the "-" on the column heading specifies that more severe raters (e.g., rater G) appear higher in the column, and more lenient raters (e.g., rater H) appear on the lower end of the column. The narrower distribution of rater severity measures (range = 2.13) than the test-taker proficiency measures (range = 11.23) indicates that the raters' individual difference had small impact on test-takers' scores. (range: 1.14 to -0.99 logits)

(4) The *fourth* column ("Test Delivery Mode") compares three test delivery modes (audio, video1, video2) in terms of their relative difficulty. Test-takers had more difficulty receiving high ratings in test modes appearing higher in the column (i.e., audio) compared to test modes appearing lower in the column (i.e., video2). In other words, raters tended to award lower scores for test-takers' responses in audio mode whereas raters scored higher in video modes. (range: 0.43 to -0.32 logits)

(5) The *fifth* column ("Rating Category") compares three rating categories (fluency, vocabulary, grammar) about their difficulties. It was in the same level of difficulty for the test-takers to receive higher or lower scores for three rating categories. (range: 0.11 to -0.08 logits)

(6) The *sixth* column ("Task") compares two task types (intermediate-level, superior-level) regarding their difficulties. The test-takers had more difficulty receiving higher scores in intermediate-level task type while it was easier for them to receive higher scores in superior-level task type. (range: 0.20 to -0.20 logits)

(7) The *seventh* column (“Scale”) shows the 9-point rating scale used by the raters to score the test-takers’ speech samples. The horizontal line (“---") across the column indicates a .5 score point. This indicates the likelihood of a test-taker receiving next higher score exceeding the likelihood of receiving the next lower score for a task in a particular test delivery mode. Approximately, proficiency measures between about -5 logits and -3 logits were more likely to receive a rating of 4 than any other rating. Likewise, proficiency measures between about -3 logits to -1 logits were more likely to receive a rating of 5; those between -1 logit to 1 logit more likely to receive a rating of 6; those between 1 logit to about 3 logits more likely to receive a rating of 7; those between about 3 logits to 5 logits more likely to receive a rating of 8; those above about 4 logits more likely receiving a rating of 9, the perfect score.

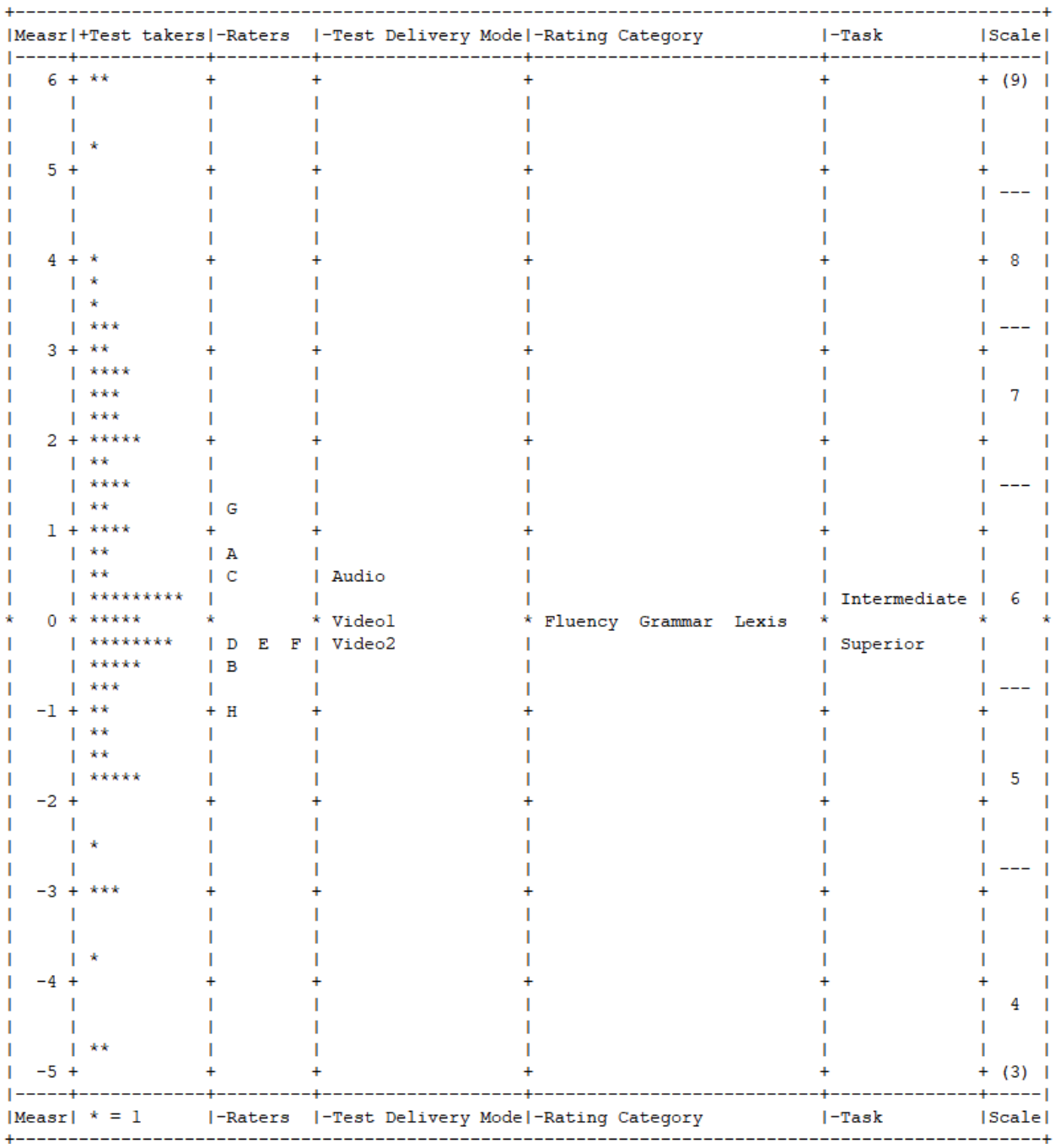


Figure 22. Variable map from the FACETS analysis of the dataset 1

Facets Statistics Summary

To better understand the data-to-model fit for dataset 1, summary statistics for all facets and each facet are presented below. In Table 12, I present the overall facets statistics summary. Then, I report summary statistics for each facet, starting with test-taker facet, followed by raters,

test delivery mode, rating category, and task facets. Prior to summary statistics presentation, I describe the key terminologies (cf. Bond & Fox, 2015).

(1) *Fair average*. In Rasch analysis, measures are reported in logits. To make sure that the measures are represented in a familiar way, "Fair average" is used. It shows what the measures designate as scores for a *standard* person assessed by *standard* rater on a *standard* task. Here, "standard" refers to "an imaginary element with the average measure of the elements of the facet" (Linacre, 2012). When there are missing data, as in current dataset, the "Fair average" adjusts for the missing data whereas the "Observed average" does not.

(2) *Difficulty measure* (in logits) refers to the estimation (or calibration) of a person ability or the item difficulty.

(3) *Model S.E*. refers to the measurement precision (also understood as *noise* to the results). That is, the standard error (S.E.) of the model provides information about how exactly the measure is located on the latent variable. *Model S.E.* is about precision, which means how reproducible the location of the measure on the latent variable is with this type of data (Linacre, 2012).

(4) *Infit mean-square* and *Outfit mean-square* indicate the measurement accuracy, which are the quality-control fit statistics. The Infit statistics is inlier-pattern sensitive; it is sensitive to the patterns in the targeted responses. The Outfit statistics is outlier sensitive, the outlying responses. These parameter estimates show how the measure corresponds to an external standard (i.e., Rasch-model ideal of invariant measure additivity). The rule-of-thumb with Outfit and Infit statistics is that if mean-square (MnSq) is above 1.5, then it is

large enough to be distorting the data-to-model fit. Generally, high mean-squares are a serious problem (MnSq > 2.0) because this means that the measurement is inaccurate.

(5) *RMSE* refers to "root mean square error". This is a statistical average of the standard errors of the measures, which shows the measure of how spread out the residuals are.

(6) *Strata* means the number of statistically distinguishable strata among the measures. This calculation is held on the context that the concept of the measurement distribution of tails are caused by *outlying "true" measures* ("true" here means the estimation adjusted for measurement error).

(7) *Separation* indicates the number of statistically distinguishable measurement strata. This estimation is based on a hypothetical context that *outlying random noise* causes the tails of the measurement distribution.

(8) *Estimated discrimination* refers to how well the facet (e.g., items) can differentiate between better and less competent examinees. For example, in summary statistics for task facet (Table 17), the estimated discrimination indicates how a task distinguishes more and less proficient test-takers. As a rule-of-thumb, estimated discrimination of 1.0 follows Rasch model expectations. The over-generalization (values greater than 1.0) is thought to be beneficial, and usually corresponds to low mean-square values and vice versa.

(9) *Point measure r* generally refers to how "predictable" one's score is. It is the correlation between the value expected from Rasch model and the value observed

Table 12 displays an overview of the results of the 5-facet analysis: *test-taker*, *rating mode* (i.e., test delivery mode), *rater*, *task*, and *rating criterion*. For the MFRM results, I use "*rating mode*" and "*test delivery mode*" alternatively, because these indicate the same facet and

the conceptual focus is more on the severity (scores) of raters than the test-takers' performances. I conducted the MFRM analysis for five target facets to check the assumption of unidimensionality. Overall, the small RMSE values (< 0.5) across the five facets confirmed a good fit to the model with an overall precision of the measured elements in each facet (cf. Linacre, 2019).

Table 12.

Facets statistics summary for dataset 1

	Test-taker*	Rating mode	Rater**	Task	Rating criterion
M (measure)	0.59	0.00	0.00	0.00	0.00
SD (measure)	2.11	0.32	0.67	0.20	0.08
Model S.E.	0.35	0.05	0.09	0.04	0.05
Infit MnSq – Min, Max	0.82, 2.30	0.99, 1.17	0.96, 1.61	0.98, 1.01	0.98, 1.00
Misfitting case (over 1.5)	8 cases (2.30, 2.15, 2.11, 2.04, 1.95, 1.78, 1.80, 1.63)	None	1 case (1.61)	None	None
Outfit MnSq – Min, Max	0.83, 2.31	0.98, 1.17	0.96, 1.51	0.98, 1.00	0.98, 1.00
Misfitting case (over 1.5)	8 cases (2.31, 2.16, 2.07, 1.98, 1.95, 1.82, 1.75, 1.68)	None	1 case (1.51)	None	None
RMSE	0.43	0.05	0.09	0.04	0.05
Adj. (true) SD	2.06	0.31	0.66	0.20	0.06
Separation **	4.79	6.11	7.40	4.74	1.24
Strata	6.72	8.47	10.20	6.65	1.99
Reliability ***	.96	.97	.98	.96	0.61

*Test-taker is the only non-centered facet; includes extreme (i.e., perfect) scores.

** For practical use, a general standard is: person separation of 2 and reliability of 0.8

*** Inter-rater agreement opportunities: 900, exact agreements: 28.4%, Expected: 44.2%

I examined all facets for data-to-model fit, and most of the facets fell within the productive measurement range of 0.5 to 1.5 (Wright & Linacre, 1994), except for eight test-takers and one rater who misfitted the model. Nonetheless, because the "Reliability" is above 0.9 except for the rating criterion facet, the error distributions are narrower which could allow for more different measures to be squeezed into the "true" distribution. The more useful information

in this case "Separation." According to the "Separation" estimator, there were 4.79 statistically distinguishable levels in the test-taker facet, 6.11 in the rating mode facet, 7.40 in the rater facet, 4.74 in the task facet, and 1.24 in the rating criterion facet. Although the focus of this study is on the rating mode, I investigate summary statistics for each facet for an in-depth understanding of data-to-model fit.

Results of the 5-Facet Analysis

In this section, I report the summary statistics of each target facet (*test-taker*, *rater*, *task*, *mode*, *rating categories*) below. In the note below each table, I also report the results of measure summary chi-square statistics, a test of fixed effects hypothesis. The purpose of this analysis is to answer the question of "*are the measures of the elements in a facet all statistically the same, except for measurement error?*" (cf. Linacre, 2021), which particularly applies to rater facet, as the raters are expected to have the same leniency. The *p*-value lower than 0.05 indicate that the hypothesis is rejected.

Table 13.

Test-taker summary statistics

Test-taker ID	Observed (raw score) average	Fair average	Difficulty measure (in logits)	Model S.E.	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
084	6.78	6.64	1.68	0.23	2.30	2.31	-0.01	0.46
078	6.67	6.34	1.05	0.58	2.15	2.16	-0.16	0.79
006	5.83	5.92	0.14	0.24	2.11	2.07	-0.08	0.36
110	4.11	4.11	-3.85	0.27	2.04	1.98	-0.25	-0.10
096	5.94	5.82	-0.09	0.24	1.95	1.95	0.11	0.36
087	7.58	7.39	3.04	0.20	1.78	1.82	-0.09	0.51
082	7.03	6.87	2.15	0.22	1.80	1.75	0.46	0.37
083	7.53	7.33	2.95	0.20	1.63	1.68	-0.19	0.38

Note. Fixed (all same) chi-square = 4259.6; *df* = 89; significance = .00; Score range is from 1 (minimum) to 9 (maximum).

In test-taker summary statistics (Table 13), only the elements with infit mean-squares over 1.5 are reported because high mean-squares could distort or degrade the measurement system, whereas low mean-squares may be less productive for measurement but not degrading. As Table 13 indicates, eight test-takers with high mean-squares were detected. Test-taker 084 had the highest mean-square of 2.30, much larger than expected 1.0. The ratings of this student underfit (i.e., high mean-squares) the Rasch model, that they are too unpredictable from the Rasch measures. These misfitted data could bring a noise into the model fit. For summary statistics of all test-takers, see Appendix G.

Table 14.

Rater summary statistics

Rater	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*	Expected point measure *r**
E	6.40	6.21	-0.32	0.07	1.61	1.51	0.53	0.84	0.86
A	5.60	5.69	0.81	0.08	1.28	1.29	0.66	0.82	0.77
H	6.54	6.52	-0.99	0.10	1.23	1.26	0.75	0.88	0.85
F	6.35	6.17	-0.23	0.07	0.89	0.90	1.05	0.85	0.86
G	5.49	5.54	1.14	0.11	0.89	0.94	1.06	0.79	0.84
B	6.21	6.27	-0.43	0.07	0.76	0.77	1.22	0.73	0.78
D	6.81	6.24	-0.37	0.11	0.50	0.52	1.43	0.92	0.86
C	5.68	5.89	0.39	0.08	0.49	0.49	1.51	0.67	0.46

Note. Fixed (all same) chi-square = 407.1; *df* = 7; significance = .00

*It is the expected vale of point-measure correlation when the data fit the Rasch model. Negative point-biserial correlations indicate miskeyed or miscoded data.

Table 14 displays summary statistics for all eight raters. Rater E was one of the most lenient raters, and was the most misfitting rater (mean-squares > 1.5). The point measure *r* indicates high point-biserial correlation between observations and their corresponding average observations. The values show that all raters behaviors work in the same direction along the test delivery modes.

Table 15.

Rating mode summary statistics

Mode	Observed raw score average	Fair (M) average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
Audio	5.84	5.87	0.43	0.05	1.17	1.17	0.79	0.79
Video1	6.11	6.12	-0.11	0.05	0.88	0.87	1.12	0.85
Video2	6.31	6.21	-0.32	0.05	0.91	0.90	1.11	0.85

Note. Fixed (all same) chi-square = 114.4; *df* = 2; significance = .00.

The rating mode facet, of most relevance to this study, shows that the audio mode was substantially the most difficult mode compared to the other video modes (Table 15). The two video modes displayed very similar difficulty levels. The fair average score for the three modes were 5.87, 6.12, and 6.21 for the audio, video1, and video2 ratings respectively. The difference between audio and video2 rating modes was 0.47 of a band, while the difference between video1 and video2 was 0.20 of a band. The fixed (all same) chi-square value ($\chi^{2 =}$ 114.4) indicate that the mode of rating significantly affected test-takers' scores ($p < .001$).

The difference between the fair average scores was 0.25 (audio and video1), 0.34 (audio and video2), and 0.09 (video1 and video2) which are smaller than the smallest rating unit (i.e., 1). In real high-stakes context, however, the application of general rounding-down will lead to different score results. For example, on average, video1 and video2 delivery modes will have scores of 6, while the audio mode will have a score of 5.

The estimated discrimination value shows that the video1 mode had the highest value (1.12) followed by the video2 mode (1.11) and the audio mode (0.79). Generally, 1.0 is the

expected value[7] , and discrimination in the range 0.5 to 1.5 provide reasonable fit to the Rasch model (Linacre, 2021).

Table 16.

Rating category summary statistics

Rating Category	Observed raw score average	Fair (M) average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
Lexis	6.04	6.02	0.11	0.05	1.00	1.00	0.99	0.83
Fluency	6.11	6.09	-0.04	0.05	0.98	0.97	1.01	0.84
Grammar	6.13	6.10	-0.08	0.05	0.97	0.96	1.03	0.83

Note. Fixed (all same) chi-square = 7.6; *df* = 2; significance = 0.02

In Table 16, no misfitting cases were detected for the three rating categories. The difficult measure shows that the lowest difficulty measure is grammar, followed by fluency, and lexis. That is, test-takers received relatively higher scores in grammar while lower scores were given in lexis category.

Table 17.

Task summary statistics

Task	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
I-L	5.99	5.98	0.20	0.04	1.01	1.00	0.99	0.83
S-L	6.20	6.16	-0.20	0.04	0.96	0.96	1.03	0.84

Note. Fixed (all same) chi-square = 46.8; *df* = 1; significance = .00

Table 17 shows no misfitting cases for both task types. Interestingly, the difficulty measure shows that superior-level task (S-L) was less difficult than the intermediate-level task (I-L). S-L had higher estimiated discrimination (1.03) than I-L (0.99).

[7] https://www.winsteps.com/facetman/table7.htm

Empirical and Expected ICCs and CCCs

(a) (b)

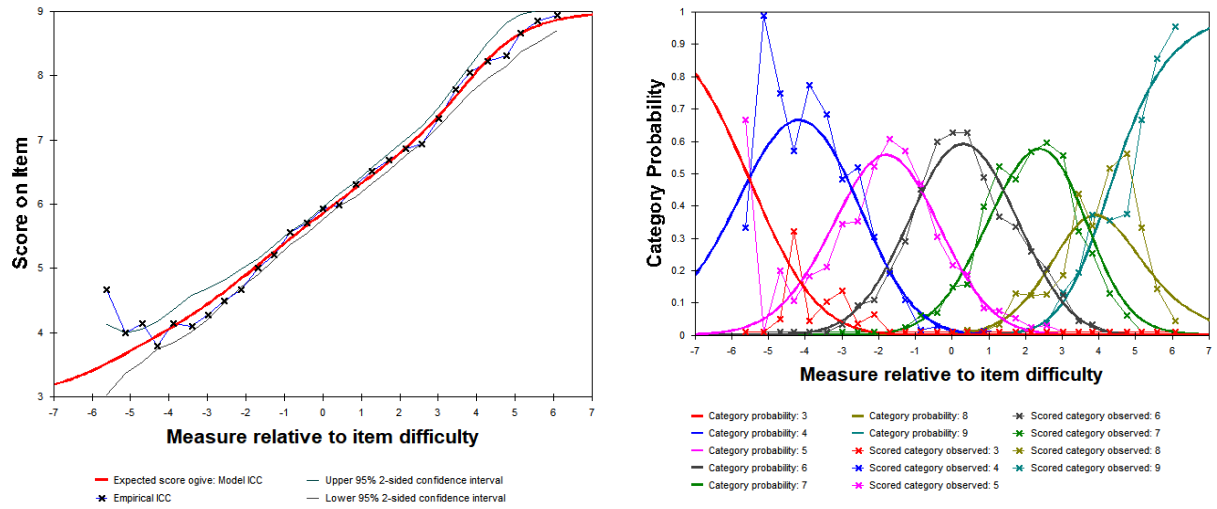


Figure 23. (a) Empirical and expected ICCs and (b) CCCs and observed scores for test indicators

As presented in Figure 23, the item characteristic curves (ICCs) and category probability curves (CCCs) provide further evidence for the valid functioning of the rating scale (Bond & Fox, 2015). In ICC, the empirical curve (thin blue line with x's) closely tracks the model (the continous red line) for most of the operational range of the scale. The curve is within the 95% confidence interval bands (light dark-green lines). From ICC, we can say that the inferences from rating-scale categories to measure are well-supported by the dataset 1.

In the CCC graph, the x-axis designates latent variable (i.e., speaking ability) regarding the difficulty of the tasks. The y-axis presents the probability of observing each category of the 9-point scale. The thinner lines with x's are the empirical category frequency lines, a summary of how the rating scale categories were used (Linacre, 2012). Each category presents a clear and separate peak with none of these categories over- or under-used (Eckes, 2015). Categories used in this data (3, 4, 5, 6, 7, 8, 9) have their thresholds rightly ordered from left to right, which implies that the rating scale categories were used and functioned in an intended order. Overall,

the ICC and CCC graphs indicate that the rating scales functioned as intended, which further supports the validity of the rating constructs.

Bias/Interaction

Following the overall analysis of facets, I performed a bias/interaction analysis between *rating mode* and *rater* facets to investigate whether and to what extent the three rating modes interact with the raters. I aimed to answer the question: "did the raters maintain their severity/leniency across the three rating modes?". To answer this question, I dummied the entire rating mode facet (audio, video1, video2) to prevent its contribution to measurement but available for interactions. The dummied facet anchors the model when running bias/interaction analysis. Although the dummy facet affected the data-model fit, the fit statistics were within an accepted range. The model I used for bias/interaction analysis is:

Interaction model: rating mode x rater → rating residual

Figure 24 displays two vertical bar charts that show size and significance of bias/interaction. For both bar charts, the horizontal line is the size value (how big) in logits, and the numbers above are the counts of interactions with that value (Linacre, 2021). Regarding the bar chart of the bias size, the number of interactions were found in different bias sizes. For example, one interaction has a bias/interaction size of 0 logit, and three interactions had bias/interaction sizes of -0.1 logit. Note that the scale is from left to right. That is, M is the mean value of the statistics for the facet, S indicates one sample standard deviation from each side of the mean, and Q indicates two sample standard deviations.

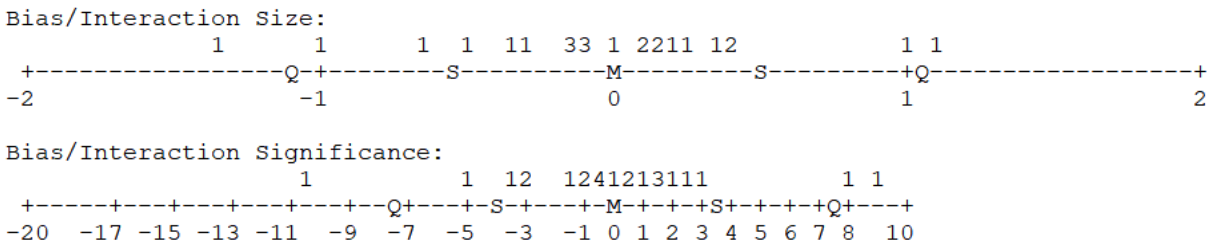


Figure 24. Vertical bar chart of bias/interaction size and significance from FACETS analysis

The bias/interaction significance bar chart displays a probability in a unit-normal deviate, *N* (0,1). For example, one interaction shows a bias/interaction significance of 0, which has a probability of less than .001 (double-sided). It is highly likely that this interaction did not happen by chance. The 8 interactions between the range of -1 to 1 logits indicate that the interactions may happened purposefully. Further inspection was conducted with summary statistics (Table 18) which contrasts raters' behaviors in three modes, and bias/interaction pairwise report (Table 19) which contrasts raters' behaviors in the rating mode facet.

Table 18 shows the contrast between local behavior with general behavior on the entire dataset. The eight raters are presented in an alphabetical order. Table 18 presents that all the raters but rater B, has an increasing Observed-Expected average from audio to video modes. These raters' severity decreased in video modes. They were giving increasingly higher than expected scores. Specifically, rater E was lenient in video2 mode and showed the largest Observed-Expected average. Rater D's severity in audio mode was the second largest Observed-Expected average. In addition, other five raters (A, C, E, F, H) also were severe when rating in audio mode. Some raters showed interaction with different test delivery modes: (a) raters A, C, D, and E showed negative interaction with the audio mode, (b) raters A, F, and H showed positive interaction with the video1 mode while rater C had negative interaction with the video1 mode, and (c) two raters, C and E, showed positive interaction with the video2 mode.

With the dummy facet anchored, I was able to run pairwise comparisons within *Facets*, by using the residuals from interactions with their own difficulty estimates (cf. Nakatsuhara et al., 2020). Table 19 shows the results of bias/pairwise report together with effect sizes (Cohen's *d*).

According to Cohen (1988), the guidelines for effect size estimation is: small (d = 0.2), medium (d = 0.5), and large (d = 0.8). The pairwise comparisons in Table 19 show that raters gave significantly lower scores in the audio rating mode than the video1 mode with small to near-medium effect sizes, which implies the extent of rater severity in rating modes. For example, if the effect size considered a 'medium' effect size (d = 0.5), then the difference between rating mode and rater severity is less than 0.5 standard deviation, which is not negligible. Furthermore, statistically significant difference indicates that rater severity in particular rating mode requires further inspection. Small effect size was found for rater A (d = 0.36), rater D (d = 0.29), rater F (d = 0.02), and rater H (d = 0.30). Rater E showed near-medium (d = 0.47) effect size. The raters gave lower scores in audio mode than video2 mode as well. Rater D showed near-medium effect size (d = 0.42), and Raters C (d = 1.25) and E (d = 0.80) showed large effect sizes. Lastly, the raters showed significantly different behaviors within the video modes. Two raters awarded significantly lower scores in video1 mode than video2 mode: Rater C (d = 1.04) with large effect size and Rater E (d = 0.33) with small effect size. The scores of video2 mode was significantly lower than video1 mode rated by Rater F (d = 0.08) and Rater H (d = 0.34) both in small effect sizes.

Table 18.

Summary statistics of bias/interaction (rater x rating mode)

Rater	Mode	Observed score	Expected score	Observed count	Obs-Exp average*	Bias size	Model SE	*t*	*df*	Prob.**	Infit MnSq	Outfit MnSq
A	Audio	644	672.39	120	-0.24	-0.48	0.13	-3.66	119	< .001	1.7	1.6
A	Video1	695	672.39	120	0.19	0.37	0.13	2.92	119	< .05	0.9	1
A	Video2	678	672.39	120	0.05	0.09	0.13	0.72	119	0.47	1	1
B	Audio	755	744.92	120	0.08	0.16	0.13	1.27	119	0.21	0.6	0.6
B	Video1	740	744.92	120	-0.04	-0.08	0.13	-0.62	119	0.54	0.7	0.7
B	Video2	740	744.92	120	-0.04	-0.08	0.13	-0.62	119	0.54	0.6	0.6
C	Audio	643	681.68	120	-0.32	-0.65	0.13	-5.01	119	< .001	0.4	0.4
C	Video1	660	681.68	120	-0.18	-0.36	0.13	-2.81	119	0.01	0.4	0.4
C	Video2	742	681.68	120	0.5	1.02	0.13	7.86	119	< .001	0.2	0.2
D	Audio	106	115.66	18	-0.54	-1	0.33	-3.01	17	0.01	0.7	0.7
D	Video1	120	115.66	18	0.24	0.42	0.31	1.35	17	0.19	1.1	1.1
D	Video2	729	723.6	108	0.05	0.08	0.12	0.68	107	0.5	0.3	0.4
E	Audio	684	767.62	120	-0.7	-1.37	0.13	-10.47	119	< .001	1.4	1.4
E	Video1	780	767.62	120	0.1	0.19	0.12	1.55	119	0.12	1.2	1.1
E	Video2	839	767.62	120	0.59	1.1	0.12	8.8	119	< .001	1.2	1.1
F	Audio	755	762.29	120	-0.06	-0.11	0.13	-0.91	119	0.36	0.8	0.8
F	Video1	778	762.29	120	0.13	0.24	0.12	1.97	119	0.05	0.7	0.8
F	Video2	754	762.29	120	-0.07	-0.13	0.13	-1.04	119	0.3	0.8	0.8
G	Audio	334	329.71	60	0.07	0.15	0.19	0.8	59	0.43	1	1.1
G	Video1	329	329.71	60	-0.01	-0.02	0.19	-0.13	59	0.89	0.6	0.6
G	Video2	326	329.71	60	-0.06	-0.13	0.19	-0.69	59	0.49	0.8	0.8
H	Audio	388	392.26	60	-0.07	-0.13	0.18	-0.75	59	0.46	1.3	1.3
H	Video1	406	392.26	60	0.23	0.42	0.18	2.4	59	0.02	1	1.1
H	Video2	383	392.26	60	-0.15	-0.29	0.18	-1.62	69	0.11	0.9	0.9

Table 18 (cont'd)

Note. Fixed (all = 0) chi-squared = 336.2; *df* = 24; $p < 0.001$. This is a chi-square test of the hypothesis that the biases presented in this Table xx are all the same apart from measurement error (Linacre, 2012). The probability of this hypothesis is below .001, which means that the hypothesis can certainly be rejected. Bias has played a significant role in the raters' behavior across test delivery modes.

*: the observed ratings subtracted by the expected ratings, on average. The formula for average is: (Observed – Expected) / Count

**: probability (*p*-value).

Table 19.

Bias/interaction pairwise report (rater x rating mode)

Rater	Target-Measure	SE	Obs-Exp Average	Rating Mode	Target Measure	SE	Obs-Exp Average	Rating Mode	Target Contrast	Joint SE	*t*	Welch *df*	*p*	Cohen's *d* *
A	1.27	0.13	-0.24	Audio	0.41	0.13	0.19	Video1	0.85	0.18	4.66	237	< .001	0.36
A	1.27	0.13	-0.24	Audio	0.69	0.13	0.05	Video2	0.57	0.18	3.12	237	< .05	0.24
A	0.41	0.13	0.19	Video1	0.69	0.13	0.05	Video2	-0.28	0.18	-1.54	237	0.12	0.13
B	-0.56	0.13	0.08	Audio	-0.32	0.13	-0.04	Video1	-0.24	0.18	-1.34	237	0.18	0.12
B	-0.56	0.13	0.08	Audio	-0.32	0.13	-0.04	Video2	-0.24	0.18	-1.34	237	0.18	0.14
B	-0.32	0.13	-0.04	Video1	-0.32	0.13	-0.04	Video2	0.00	0.18	0.00	238	1.00	0.26
C	1.03	0.13	-0.32	Audio	0.75	0.13	-0.18	Video1	-.28	0.18	1.55	237	0.12	0.34
C	1.03	0.13	-0.32	Audio	-0.63	0.13	0.50	Video2	1.67	0.18	9.11	237	< .001	1.25
C	0.75	0.13	-0.18	Video1	-0.63	0.13	0.50	Video2	1.39	0.18	7.56	237	< .001	1.04
D	0.55	0.33	-0.54	Audio	-0.86	0.31	0.24	Video1	1.41	0.45	3.13	33	< .001	0.29
D	0.55	0.33	-0.54	Audio	-0.53	0.13	0.05	Video2	1.08	0.35	3.06	22	.01	0.42
D	-0.86	0.31	0.24	Video1	-0.53	0.13	0.05	Video2	-0.33	0.33	-1.00	22	0.33	0.13
E	1.07	0.13	-0.70	Audio	-0.49	0.12	0.10	Video1	1.56	0.18	8.65	237	< .001	0.47
E	1.07	0.13	-0.70	Audio	-1.40	0.12	0.59	Video2	2.47	0.18	13.65	237	< .001	0.80
E	-0.49	0.12	0.10	Video1	-1.40	0.12	0.59	Video2	0.91	0.18	5.15	237	< .001	0.33
F	-0.10	0.13	-0.06	Audio	-0.46	0.12	0.13	Video1	0.36	0.18	2.03	237	0.04	0.02
F	-0.10	0.13	-0.06	Audio	-0.09	0.13	-0.07	Video2	-0.02	0.18	-0.09	237	0.93	0.11
F	-0.46	0.12	0.13	Video1	-0.09	0.13	-0.07	Video2	-0.38	0.18	-2.12	237	0.03	0.08
G	0.96	0.19	0.07	Audio	1.13	0.19	-0.01	Video1	-0.17	0.26	-0.66	117	0.51	0.05
G	0.96	0.19	0.07	Audio	1.23	0.19	-0.06	Video2	-0.28	0.26	-1.05	117	0.29	0.07
G	1.13	0.19	-0.01	Video1	1.23	0.19	-0.06	Video2	-0.10	0.26	-0.40	117	0.69	0.13
H	-0.79	0.18	-0.07	Audio	-1.35	0.18	0.23	Video1	0.55	0.25	2.23	117	0.03	0.30
H	-0.79	0.18	-0.07	Audio	-0.64	0.18	-0.15	Video2	-0.15	0.25	-0.62	117	0.54	0.03
H	-1.35	0.18	0.23	Video1	-0.64	0.18	-0.15	Video2	-0.71	0.25	-2.85	117	0.01	0.34

*****The effect size was calculated for Welch's *t*-tests. Cohen's *d*, the proportion of a standard deviation difference, was computed.

Note. The bias/interaction Table 19 displays only the most conspicuous interactions.

CHAPTER 14: RESULTS OF DATASET 2: ITEMS

Results

In this chapter, I first report the distribution of ITA Speaking test scores using means, standard deviations, skewness/kurtosis, and histograms. Table 20 shows that the test-takers received higher scores in the video-conferencing (synchronous, direct) delivery mode (*hereafter* VC) than the video-recorded (asynchronous, semi-direct) mode (*hereafter* VR), except for item 7. The score difference between the two delivery modes of all twelve items ranged between 0.29 and 1.46. Of all the items in both test delivery modes, item 5 in the live mode had the highest score (M = 49.58) while item 9 in the recorded mode had the lowest score (M = 46.53).

Table 21 shows the distribution of scores for 12 items of each test delivery mode. The scores were normally distributed, however, item 6 in the recorded mode had high kurtosis (8.64) that was higher than suggested criteria of below 7 (Byrne, 2010; Hair et al., 2010). A closer look at the frequency of scores in item 6 shows that more than half of the test-takers (n = 158) received score of 50, followed by score of 40 (n = 87), score of 60 (n = 36), and score of 30 (n = 4). I present the histograms (Figures 25 and 26) to provide readers a visual distribution of the scores.

Table 20.

ITA Speaking Test means for items

	VC Mode	VR Mode
	M (*SD*)	*M* (*SD*)
Item1	48.00 (6.60)	47.12 (6.47)
Item2	48.31 (6.83)	47.25 (6.45)
Item3	48.98 (7.13)	47.67 (7.04)
Item4	48.66 (7.01)	47.81 (7.13)
Item5	49.58 (6.71)	48.12 (6.77)
Item6	48.73 (6.81)	47.67 (7.12)
Item7	47.61 (6.61)	48.39 (6.81)
Item8	48.84 (6.21)	48.07 (6.81)

Table 20 (cont'd)

Item9	47.04 (6.43)	46.53 (6.65)
Item10	47.32 (6.82)	46.88 (6.84)
Item11	48.63 (6.72)	47.23 (6.65)
Item12	47.29 (6.52)	47.00 (7.20)

Note. For all items: minimum score = 20 and maximum score = 60.

Table 21.

Distribution of ITA Speaking Test scores

	VC mode		VR mode	
	Skewness	Kurtosis	Skewness	Kurtosis
Item1	-0.05	2.78	0.07	2.60
Item2	0.02	2.56	0.05	2.62
Item3	-1.02	9.52	-0.11	3.03
Item4	-0.24	3.30	0.06	2.41
Item5	-0.02	2.42	-0.01	2.55
Item6	-0.04	2.59	-0.96	8.64
Item7	0.16	2.46	-0.06	2.72
Item8	0.08	2.54	0.00	2.50
Item9	0.20	2.49	-0.04	2.74
Item10	0.19	2.43	0.03	2.63
Item11	0.03	2.51	-0.08	2.71
Item12	0.18	2.48	-0.84	8.35

Results of Paired *t*-test

To examine the mean differences between two delivery modes, I computed paired *t*-test for each item. Table 22 shows that all items had statistically and significantly higher means in the live delivery mode than the recorded delivery mode, except for items 9, 10, and 12. Five items had the significant difference at a *p*-value of < .05 level: items 1, 2, 4, 6, and 8. The other four items had significant differences at a *p*-value of <.001 level: item 3, 5, 7, and 11.

Table 22.

Paired t-test results comparing test delivery mode of each item

		t	*p*	Cohen's *d*
Item 1	VC VR	2.17	0.03	0.13
Item 2	VC VR	2.97	0.003	0.16

Table 22 (cont'd)

Item 3	VC VR	3.38	0.001	0.18
Item 4	VC VR	2.02	0.04	0.12
Item 5	VC VR	3.60	0.0004	0.22
Item 6	VC VR	2.44	0.02	0.15
Item 7	VC VR	3.08	0.002	0.12
Item 8	VC VR	2.11	0.04	0.12
Item 9	VC VR	1.37	0.17	0.08
Item 10	VC VR	1.15	0.25	0.06
Item 11	VC VR	3.66	0.00003	0.21
Item 12	VC VR	0.72	0.47	0.04

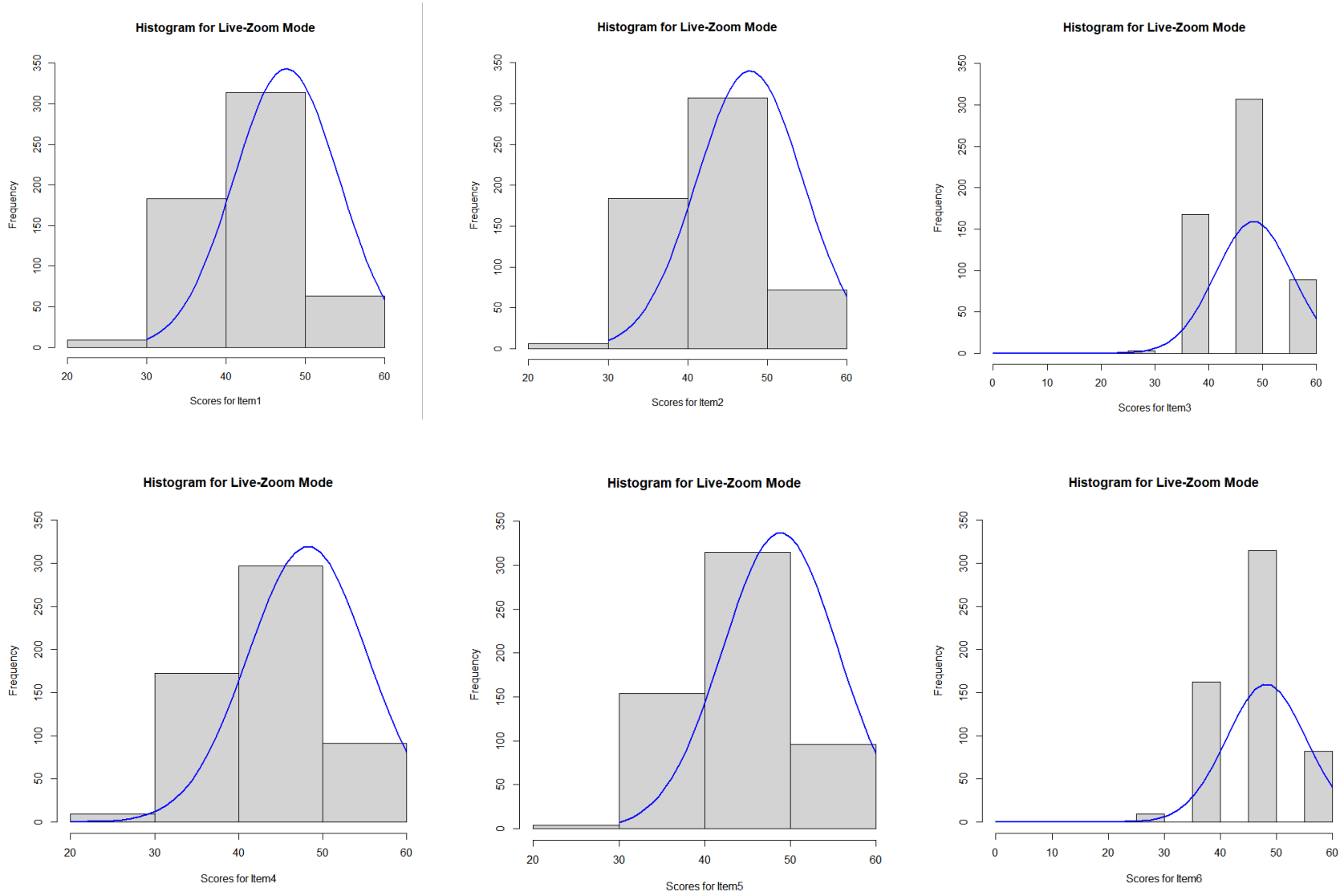
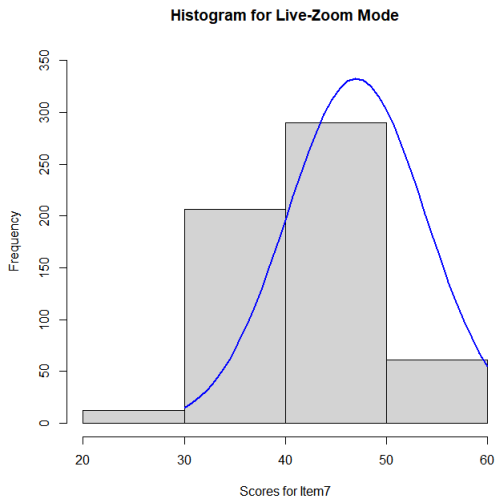
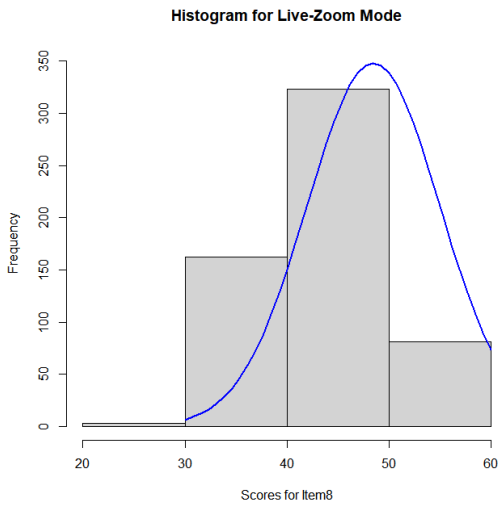
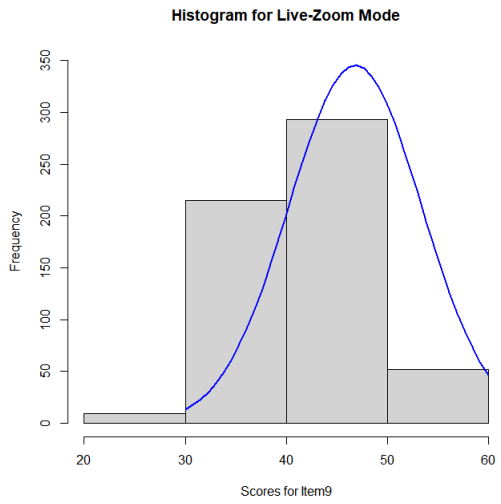


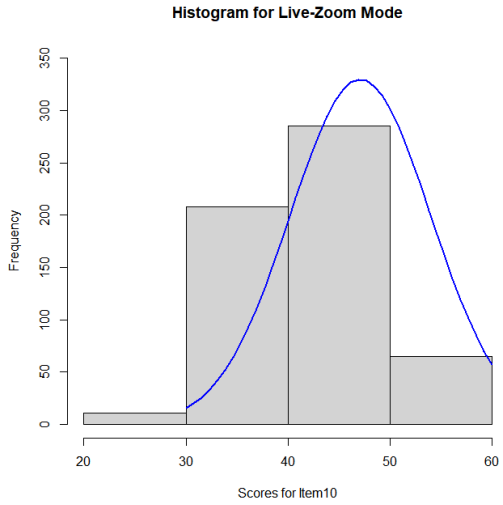
Figure 25. Histograms representing distribution of ITA Speaking test score categories from VC delivery mode

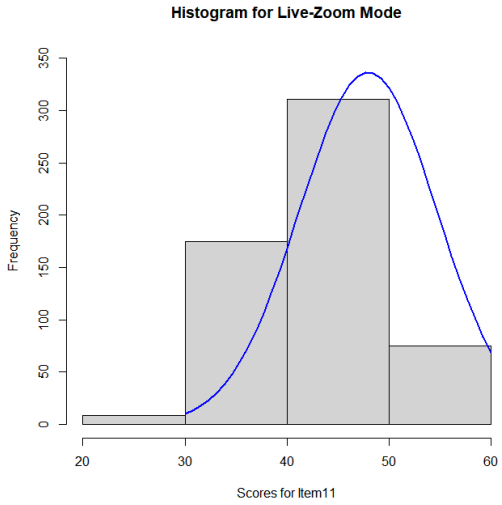
Figure 25. (cont'd)

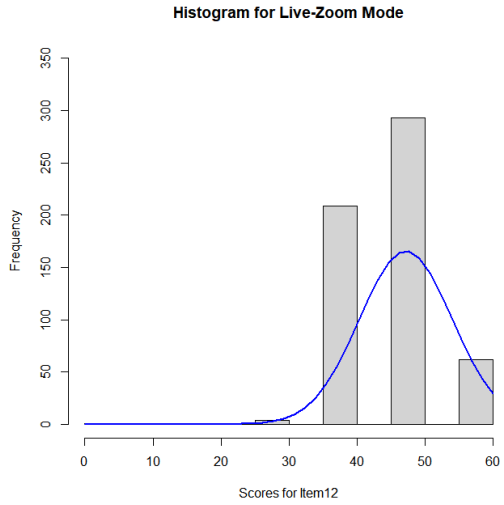












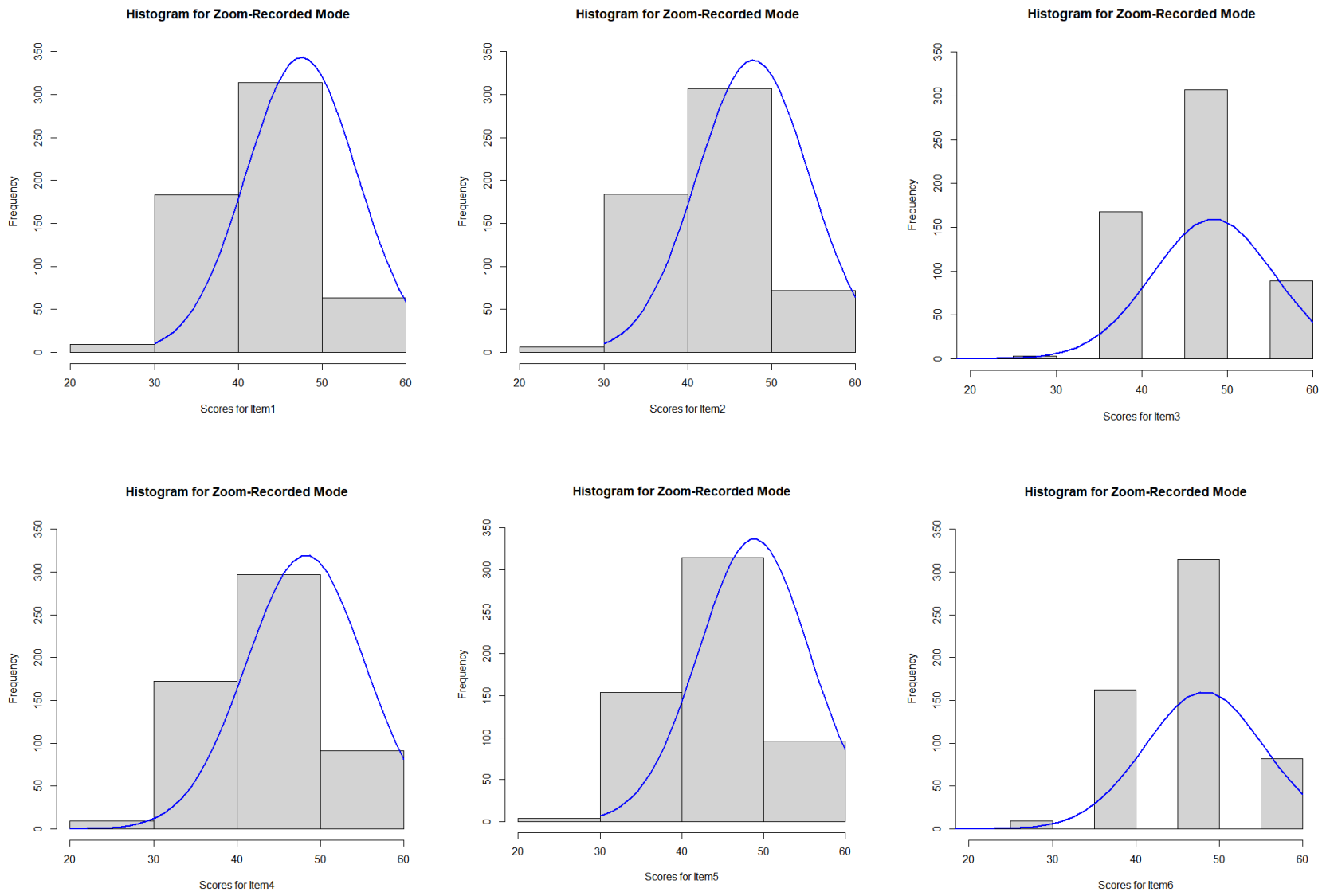
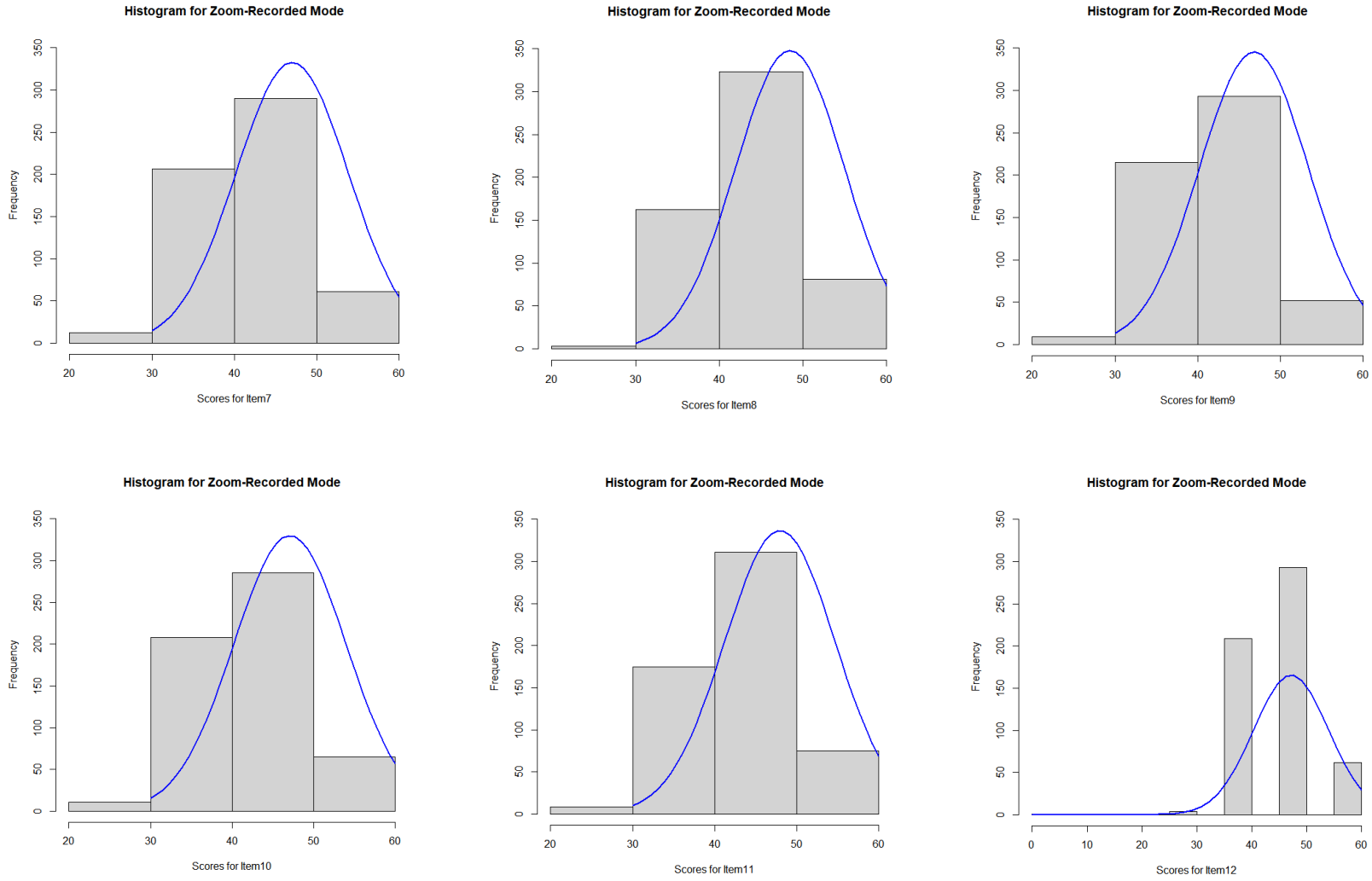


Figure 26. Histograms representing distribution of ITA Speaking test score categories from VR delivery mode

Figure 26. (cont'd)



Intra-Class Correlation Coefficient

ICC was not calculated in this dataset because of the missing data (scores). Each rater was assigned with different number of test-takers, ranging from minimum of 5 test-takers to maximum of 54 test-takers. Hence, SPSS ver. 24 showed the negative values.

Results of Confirmatory Factor Analysis (CFA)

CFA was computed for all test-takers to examine whether the scores for ITA Speaking test are comparable with speaking tests that are administered in two different test delivery modes (VC, VR).

Correlation Matrices

Table 23 presents the correlation matrices of the 12 items in each delivery mode. The correlation coefficients ranged from 0.37 to 0.69, which were higher within each mode.

Table 23.

Correlation matrices for indicator variables (N = 285)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1 VC-Item1	1																							
2 VC-Item2	0.56	1																						
3 VC -Item3	0.51	0.57	1																					
4 VC -Item4	0.58	0.59	0.57	1																				
5 VC -Item5	0.59	0.58	0.60	0.58	1																			
6 VC -Item6	0.57	0.69	0.61	0.61	0.60	1																		
7 VC -Item7	0.53	0.51	0.49	0.46	0.56	0.56	1																	
8 VC -Item8	0.55	0.60	0.49	0.55	0.55	0.62	0.51	1																
9 VC -1tem9	0.51	0.55	0.50	0.48	0.51	0.49	0.51	0.53	1															
10 VC -Item10	0.52	0.60	0.55	0.51	0.56	0.58	0.54	0.47	0.56	1														
11 VC-Item11	0.57	0.58	0.54	0.60	0.61	0.59	0.57	0.60	0.51	0.54	1													
12 VC-Item12	0.52	0.56	0.55	0.51	0.51	0.56	0.56	0.54	0.60	0.61	0.55	1												
13 VR-Item1	0.49	0.53	0.47	0.45	0.44	0.46	0.42	0.43	0.51	0.48	0.45	0.46	1											
14 VR-Item2	0.43	0.58	0.45	0.49	0.45	0.45	0.37	0.43	0.47	0.49	0.46	0.46	0.65	1										
15 VR-Item3	0.46	0.46	0.59	0.50	0.42	0.46	0.39	0.44	0.43	0.43	0.46	0.47	0.59	0.58	1									
16 VR-Item4	0.47	0.52	0.43	0.53	0.45	0.49	0.45	0.46	0.42	0.42	0.50	0.48	0.62	0.62	0.59	1								
17 VR-Item5	0.39	0.49	0.48	0.47	0.48	0.46	0.47	0.45	0.49	0.50	0.46	0.50	0.62	0.67	0.60	0.63	1							
18 VR-Item6	0.45	0.45	0.44	0.44	0.44	0.44	0.41	0.42	0.43	0.42	0.47	0.50	0.53	0.52	0.54	0.56	0.56	1						
19 VR-Item7	0.43	0.47	0.42	0.43	0.41	0.40	0.50	0.39	0.39	0.44	0.44	0.43	0.56	0.58	0.59	0.58	0.59	0.51	1					

Table 23 (cont'd)

20 VR-Item8	0.44	0.54	0.50	0.50	0.47	0.51	0.45	0.55	0.44	0.50	0.50	0.52	0.65	0.64	0.64	0.62	0.63	0.55	0.59	1			
21 VR-Item9	0.43	0.52	0.50	0.51	0.46	0.51	0.47	0.50	0.50	0.42	0.50	0.45	0.48	0.57	0.55	0.58	0.59	0.58	0.51	0.62	1		
22 VR-Item10	0.47	0.50	0.46	0.44	0.46	0.49	0.45	0.45	0.43	0.52	0.48	0.45	0.63	0.55	0.56	0.62	0.60	0.55	0.54	0.59	0.62	1	
23 VR-Item11	0.47	0.51	0.51	0.49	0.45	0.52	0.48	0.47	0.53	0.47	0.55	0.50	0.63	0.59	0.57	0.65	0.59	0.58	0.58	0.65	0.62	0.62	1
24 VR-Item12	0.40	0.45	0.48	0.40	0.43	0.43	0.46	0.40	0.41	0.47	0.48	0.48	0.55	0.54	0.53	0.56	0.54	0.47	0.54	0.60	0.57	0.55	0.57	1

Note. The first initials represent the test delivery mode; VC (synchronous video-conferencing mode) and VR (asynchronous video-recorded mode).

Model Fit Statistics

Model fit statistics were examined for two CFA models, single-factor (Figure 27) and correlated two-factor models (Figure 28). Table 24 shows the model fit indices, with correlated two-factor model having the excellent fit. CFI is higher than 0.95, and both SRMR and RMSEA have values less than 0.07. While single-factor model shows the good fit for SRMR and RMSEA, the low value of CFI indicates that the two latent factors best represent the construct of academic English language oral communication ability, specifically as a teaching assistant. Thus, I selected the correlated two-factor model as the best fitting model.

Table 24.

Fit statistics for the two models

Model	SBχ^2	df	CFI	SRMR	RMSEA	AIC	BIC
Single-factor	4777.864	276	0.889	0.058	0.084	41503.358	41766.084
Correlated two-factor (VC and VR)	4777.864	276	0.978	0.030	0.037	41100.501	41366.876

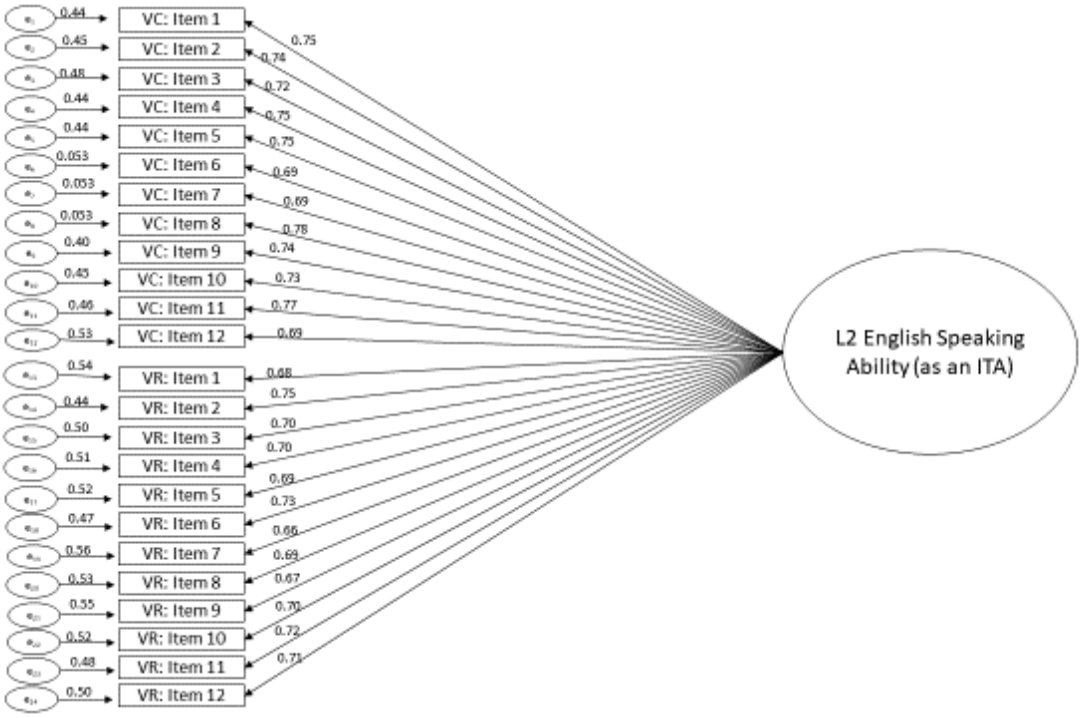


Figure 27. Single-factor model

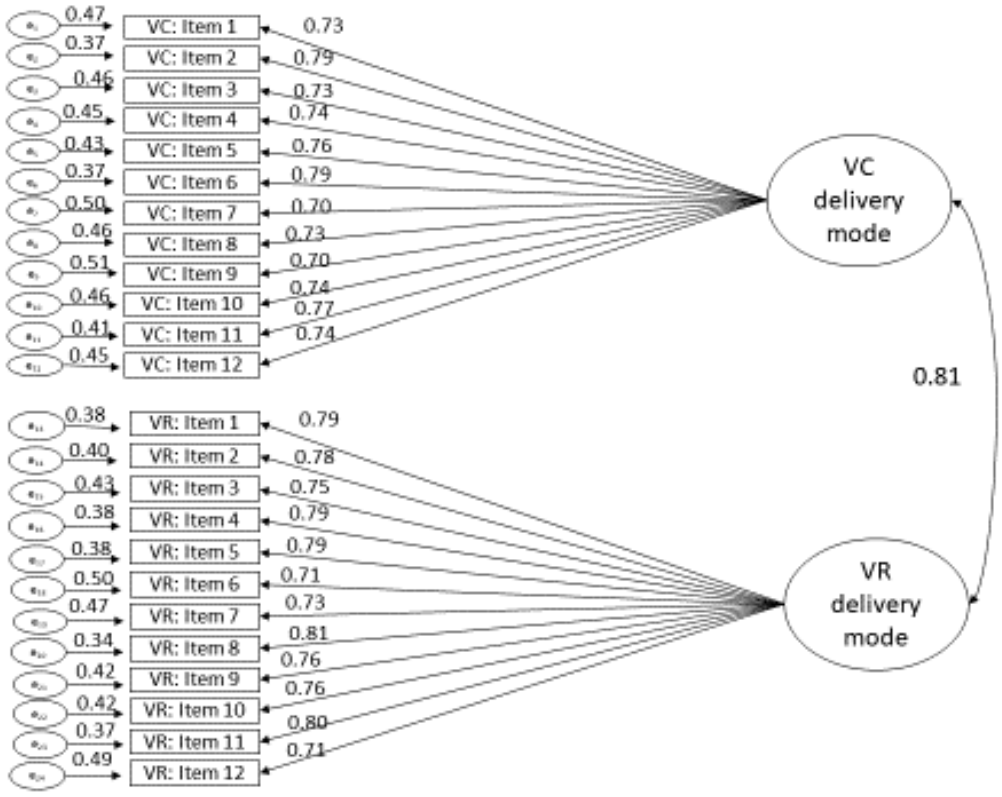


Figure 28. Correlated two-factor model (VC vs. VR)

Description of the Final Model

Figure 28 above shows the correlated two latent factors (VC, VR). The prospective ITA's academic oral communication ability as a teaching assistant within each test delivery mode was measured using 12 items with a scale of 20 to 60 that load on each factor. The factor inter-correlations are above 0.80 (VC and VR: $r = 0.81$), which indicates poor discriminant validity (cf. Kline, 2016). This model specifies that the delivery modes of this speaking test are not distinct from one another. The item loadings (weights) on each latent factor are indicated on the straight arrow which range from .70 to .81. The item intercepts (means) are displayed within each indicator variable (square), which range from 6.5 to 7.9. Item error variances and covariances are presented by parameters ε1-ε24 in a small circle.

Results of Multifaceted Rasch model (MFRM) Analysis

MFRM analysis was conducted using 6,825 measurable responses. The data summary indicated the successful estimation, as shown by the mean standard residual (Resd) and the mean standardized residual (StRes) of 0.00, and the standard deviation (S.D.) of 1.00. Rasch analysis explained this dataset's variance by 54.95%.

Description of the FACET Variable Map

Based on the terminologies described in the 'Description of the final model' subsection in Chapter 13, I report the findings of the dataset 2 below.

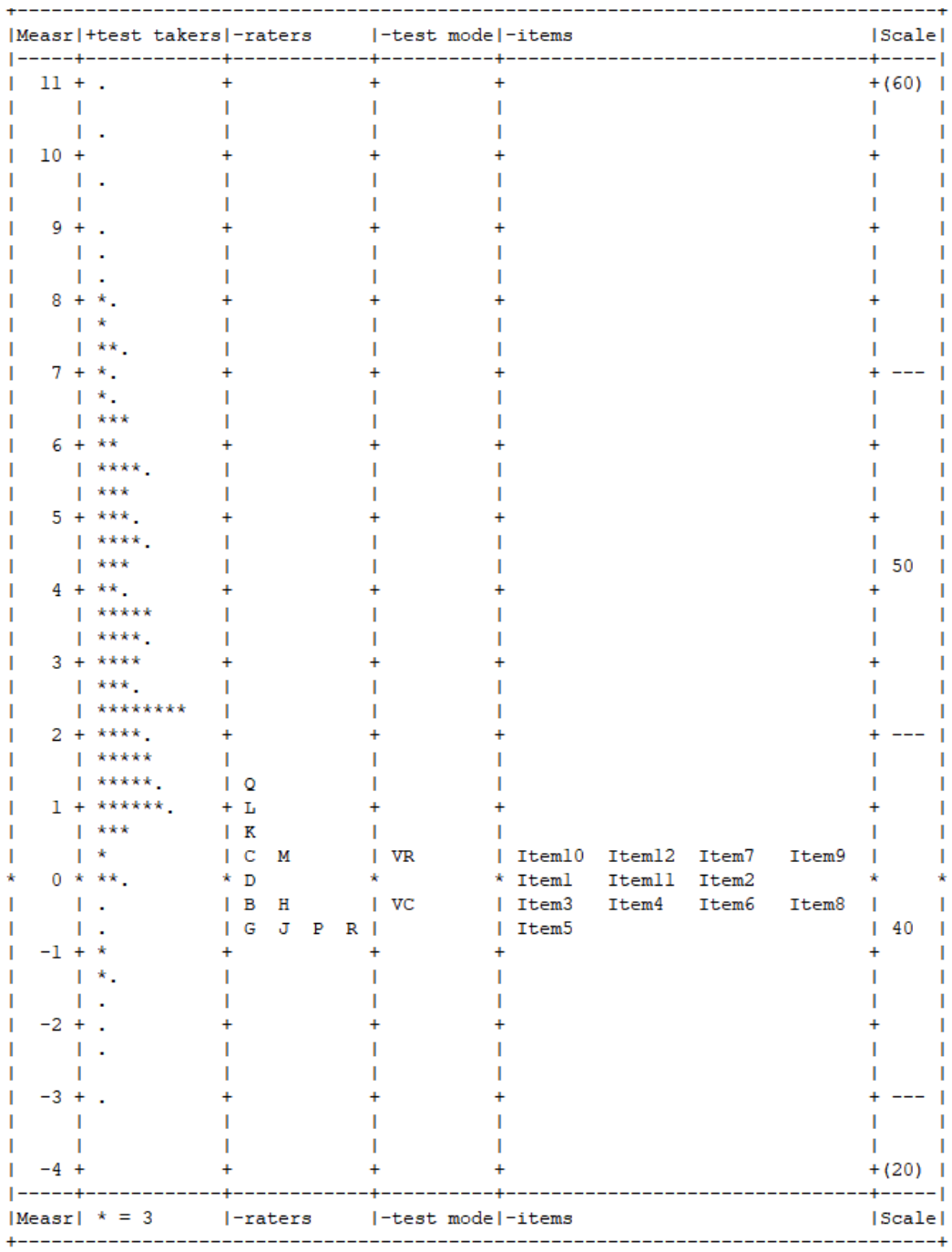


Figure 29. Variable map from the FACETS analysis of the dataset 2

The variable map (Figure 29) shows the overview of the results with key variables (*test-takers*, *raters*, *test delivery mode*, *items*) along with logit scale (the first left column) and the scale (the first column on the right) used by the raters to assess prospective ITAs' oral

performance. The plot shows the estimates of the four facets. Since the definition and function of each column were explained in the dataset 1, I interpret the map unique for this dataset.

The "Measr" column (logit scale) shows that the scale ranged from 11 logits to -4 logits. According to the "test-takers" column which displays 285 ITAs' estimated oral communication ability, most of them are above the top category of the easiest item (items 10, 12, 7, and 9) with a greater probability of 50% exceeding. Test-takers with the highest proficiency are located on the top of the column and the lowest proficiencies are located on the bottom of the column. A dot (*) in the test-taker column indicates the number of people, 1 to 3[8] (Linacre, 2021).

The raters are located differently ("raters" column), with the most lenient raters displayed on the lower end of the column (raters G, J, P, and R) and most severe rater is displayed on the higher end of the column (rater Q). The narrower distribution of rater severity measures (range = 2.04) than the test-taker proficiency measures (range = 13.93) indicates that the raters' individual differences had little impact on the test-takers' scores.

Of most importance to this study, the "test mode" column shows that the test-takers received lower scores in recorded delivery mode (VR) than their same responses in the live delivery mode (VC). The different location of two delivery modes indicates that a test-taker's oral response was rated harsher when presented in a VR mode while raters were more lenient when assessing in VC delivery mode.

[8] https://www.winsteps.com/winman/table1.htm

The “items” column displays the difficulty of the 12 items. The test-takers had more difficulty receiving higher scores for items 10, 12, 7, and 9. Item 5 was easier for them to achieve higher scores.

Lastly, the 5-point rating scale used by the raters to assess test-takers’ speech samples are presented in the ‘scale’ column. Prospective ITAs with proficiency measures approximately between +2 logits and -3 logits were more likely to receive a rating of 40; those between about +7 logits to +2 logits more likely to receive a rating of 50; and those between +11 to +7 logits likely received a score of 60 (perfect score).

Facets Statistics Summary

Following the description of variable map, in this section, I report the data-to-model fit for dataset 2 at item-level. The summary statistics for all facets (Table 25) and each facet (Tables 26, 27, 28, and 29) are presented below. The summary statistics for each facet are reported starting from *test-taker* facet, followed by *rater*, *test delivery mode* (i.e., rating mode), and *item* facets.

The overall facets statistics summary (Table 25) was conducted to examine the unidimensionality assumption. The RMSE values were below 0.5 in all four facets, which indicate a good fit to the Rasch model with an overall precision of the measured elements in each facet.

Table 25.

Facets statistics summary

	Test-taker*	Rater**	Test delivery mode	Item
M (measure)	3.35	0.00	0.00	0.00
SD (measure)	2.52	0.70	0.30	0.32
Model S.E.	0.47	0.13	0.04	0.09

Table 25 (cont'd)

Infit MnSq – Min, Max	0.01, 2.96	0.72, 1.33	0.98, 1.02	0.89, 1.12
Misfitting case (over 1.5)	26 cases (2.96, 2.57, 2.33, 2.37, 2.22, 2.20, 2.07, 1.96, 1.99, 1.91, 1.84, 1.82, 1.83, 1.70, 1.76, 1.82, 1.80, 1.64, 1.60, 1.56, 1.63, 1.64, 1.64, 1.57, 1.60, 1.53)	None	None	None
Outfit MnSq – Min, Max	0.01, 3.01	0.67, 1.40	0.97, 1.03	0.85, 1.16
Misfitting case (over 1.5)	28 cases (3.01, 2.56, 2.47, 2.36, 2.21, 2.20, 2.11, 2.10, 2.05, 1.91, 1.89, 1.85, 1.82, 1.82, 1.82, 1.74, 1.80, 1.80, 1.75, 1.75, 1.73, 1.68, 1.65, 1.64, 1.63, 1.63, 1.58, 1.56)	None	None	None
RMSE	0.48	0.15	0.04	0.09
Adj. (true) SD	2.47	0.68	0.29	0.31
Separation***	5.19	4.42	7.73	3.32
Strata	7.25	6.23	10.64	4.76
Separation reliability***	0.96	0.95	0.98	0.92

*Test-taker is the only non-centered facet; includes extreme (i.e., perfect) scores.

** Inter-rater agreement opportunities were 0 in this dataset.

***General standard for practical use: person separation of 2 and reliability of 0.8.

All four facets were examined for model-data fit. The infit and outfit mean square values between range of 0.72 to 1.33 indicate that that three facets (*rater, test delivery mode, item*) were within the productive range. The *test-taker* facet had 26 prospective ITAs who showed misfitting infit meansquare values (1.53 to 2.96) and 28 of them who had misfitting outfit meansquare values (1.56 to 3.01). The "Separation" value indicates that *test-taker* facet had 5.19 statistically distinguishable levels, *rater* facet with 4.42, *test delivery mode* facet with 7.73, and *item* facet with 3.32. Different from the summary statistics in dataset 1 which had *rater* facet as having the highest distinguishable level of 7.40, *test delivery mode* was the highest in the current dataset.

Results of the 4-Facet Analysis

Here, I report the summary statistics of the four target facets (test-taker, rater, mode, item) sorted by infit mean-square values. I first report test-taker facet (Table 26), rater facet (Table 27), mode facet (Table 28), and item facet (Table 29).

Table 26.

Test-taker summary statistics

Test-taker ID	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
210	49.17	49.00	3.80	0.50	2.96	3.01	0.17	0.08
070	40.00	39.40	-1.02	0.55	2.57	2.56	0.39	-0.31
125	49.58	49.21	3.92	0.49	2.33	2.47	0.12	0.37
114	46.67	46.74	2.68	0.43	2.37	2.36	-1.12	0.31
165	51.25	51.38	5.32	0.49	2.22	2.21	0.15	0.50
109	50.42	50.10	4.51	0.52	2.20	2.20	0.41	0.60
143	50.00	49.58	4.16	0.49	2.07	2.11	0.32	0.44
268	47.92	47.94	3.22	0.45	1.96	2.10	0.04	0.25
177	41.25	41.71	0.57	0.51	1.99	2.05	0.36	0.51
221	41.67	42.17	0.81	0.48	1.91	1.91	0.26	0.59
038	49.58	50.05	4.48	0.51	1.84	1.89	0.55	0.07
279	40.00	39.19	-1.18	0.55	1.82	1.85	0.62	0.09
154	45.83	44.61	1.85	0.41	1.83	1.82	-1.18	-0.44
046	48.33	48.35	3.43	0.47	1.70	1.82	0.45	0.58
151	52.50	52.74	5.99	0.44	1.76	1.82	0.02	-0.07
092	51.25	50.23	4.60	0.48	1.82	1.74	0.43	0.40
161	49.58	48.84	3.70	0.52	1.80	1.80	0.59	0.23
162	47.50	48.99	3.79	0.44	1.64	1.80	0.27	0.62
062	40.00	40.22	-0.41	0.53	1.60	1.75	0.82	0.23
254	36.67	36.53	-2.49	0.41	1.38	1.75	0.22	0.89
219	59.17	59.64	10.18	0.76	1.13	1.73	0.84	-0.03
230	51.67	51.47	5.37	0.47	1.56	1.68	0.54	0.24
084	50.00	50.72	4.92	0.52	1.63	1.65	0.68	0.08
222	48.75	48.41	3.46	0.49	1.64	1.64	0.56	-0.14
248	51.67	51.54	5.40	0.47	1.64	1.63	0.45	0.00
099	40.42	39.83	-0.71	0.55	1.57	1.63	0.70	-0.24
231	53.75	53.60	6.36	0.41	1.60	1.58	-0.75	0.17
144	47.50	46.54	2.60	0.44	1.49	1.56	0.38	0.44
239	55.83	55.34	7.04	0.41	1.53	1.49	-1.81	0.45

Note. Fixed (all same) chi-squared = 7847.2; *df* = 284; significance = .00

Table 26 above shows 26 test-takers' summary statistics who have infit mean-squares higher than 1.5. These test-takers' scores, whether their scores are higher or lower than expected scores (fair average), indicate the unpredictability from the Rasch measures which bring noise to the model fit.

It should be noted that the seven test-takers (IDs 210, 070, 125, 114, 165, 109, 143) have infit mean-squares higher than 2.0, which flags the unexpected patterns in on-target observations. While mean-square between 1.5 and 2.0 indicate that the element (test-taker) is unproductive for construction of measurement, it is not degrading. However, mean-squares are over 2.0 imply that elements distort or degrade the measurement system. The highest infit mean-square (2.96) was detected in test-taker 210 who has an observed score of 49.17 and fair average of 49.00.

Table 27.

Rater summary statistics

Rater	Total score	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
P	20,290	47.08	49.55	-0.83	0.10	1.33	1.40	0.58	0.69
B	54,850	49.19	48.82	-0.37	0.07	1.23	1.28	0.71	0.76
R	7,000	48.95	49.18	-0.59	0.19	1.22	1.19	0.79	0.48
L	2,210	46.04	45.59	1.10	0.33	1.09	1.13	0.94	0.66
C	43,400	46.97	47.45	0.34	0.07	1.10	1.12	0.88	0.81
J	9,560	46.86	49.33	-0.68	0.15	1.02	1.04	0.94	0.57
M	33,270	47.80	47.61	0.26	0.08	1.03	1.02	0.98	0.60
D	57,960	48.30	48.23	-0.04	0.06	0.93	0.93	1.09	0.75
Q	7,530	48.27	45.29	1.21	0.18	0.84	0.83	1.16	0.83
G	5,500	45.83	49.08	-0.53	0.21	0.85	0.81	1.16	0.83
H	44,020	48.27	49.04	-0.50	0.07	0.81	0.77	1.21	0.69
K	40,790	46.56	46.81	0.61	0.08	0.72	0.67	1.29	0.75

Note. Fixed (all same) chi-squared = 324.0; *df* = 11; significance = .00

The twelve raters' summary statistics in Table 27 shows that all raters' infit mean squares are in a range for productive for measurement (0.5 to 1.5). No raters were found to force other raters to be reported as misfitting.

The severity of the raters has to be understood in two terms, because only eight raters (raters B, C, D, H, J, K, M, and P) assessed speech samples in both delivery modes. Among the eight raters, rater J was most severe while rater D was most lenient. Among the other four raters who assessed only the recorded delivery mode, rater L was most sever while rater P was most lenient. All raters have high point measure correlation coefficients.

Table 28.

Rating mode summary statistics

Mode	Total score	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
Live	164,430	48.26	48.54	-0.21	0.04	1.02	1.03	0.98	0.73
Reco-rded	161,950	47.38	47.72	0.21	0.04	0.98	0.97	1.02	0.75

Note. Fixed (all same) chi-square = 60.7; *df* = 1; significance = .00

Total score in Table 28 shows that the raters are more severe in the VR delivery mode than the VC delivery mode. This implies that test-takers received higher scores in the VC delivery mode, while their same oral performance received lower scores in the VR delivery mode. Both rating modes are in a productive infit mean-square range.

Table 29.

Item summary statistics

Item	Total score	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrim.	Point measure *r*
Item 4	27,450	48.24	48.53	-0.21	0.09	1.12	1.16	0.89	0.74
Item 7	26,760	47.03	47.33	0.39	0.09	1.15	1.13	0.85	0.71
Item 10	26,800	47.10	47.41	0.35	0.09	1.09	1.09	0.91	0.73
Item 3	27,520	48.45	48.71	-0.30	0.09	1.02	1.05	0.98	0.74
Item 5	27,790	48.84	49.04	-0.50	0.09	0.99	0.98	1.01	0.75
Item 6	27,420	48.27	48.56	-0.22	0.09	0.99	0.98	1.01	0.75
Item 9	26,640	46.82	47.10	0.49	0.09	0.99	0.98	1.00	0.73
Item 1	27,070	47.57	47.90	0.12	0.09	0.97	1.0	1.03	0.74
Item 12	26,850	47.27	47.60	0.27	0.09	0.94	0.94	1.06	0.74
Item 11	27,290	47.96	48.28	-0.07	0.09	0.94	0.94	1.07	0.76

Table 29 (cont'd)

Item 8	27,580	48.47	48.73	-0.32	0.09	0.94	0.93	1.07	0.74
Item 2	27,210	47.82	48.14	.00	0.09	0.89	0.85	1.12	0.77

Note. Fixed (all same) chi-square = 132.2; *df* = 11; significance = .00

Table 29 shows the twelve items are in a productive infit mean-square range. Item 2 has the highest estimated discrimination value (1.12) while item 4 has the lowest (0.89). Test-takers received the lowest score for item 9 (total score = 26,640), which indicates that this item was the most difficult for the test-takers to achieve high scores as in item 5 (total score = 27,580). Items 7, 10, and 12 also have the low total scores which imply that these were also difficult for the test-takers.

Empirical and Expected ICCs and CCCs for ITA Speaking Test Indicators

(a) (b)

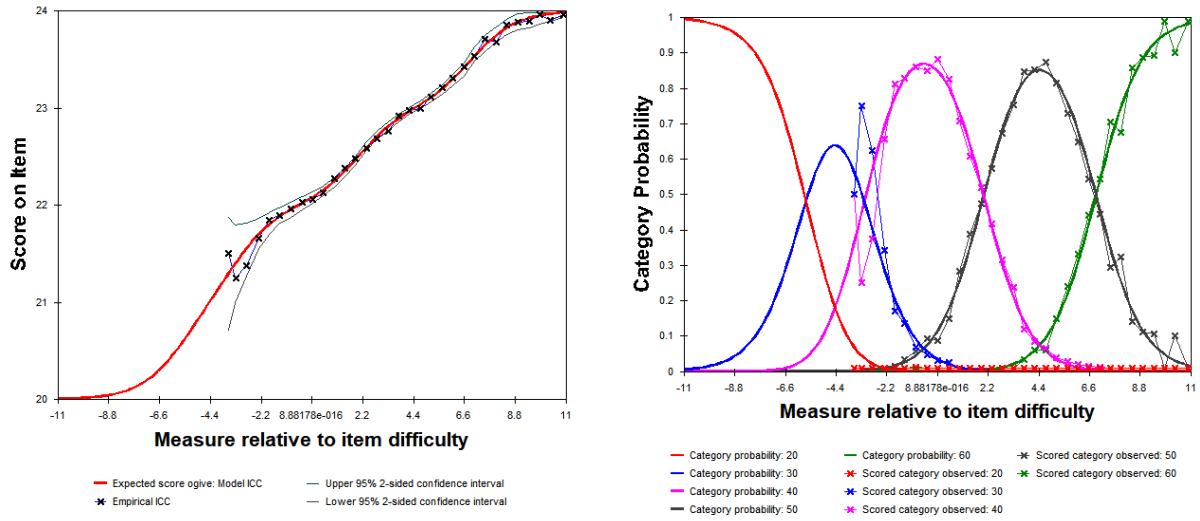


Figure 30. (a) Empirical and expected ICCs, (b) category probability curves and observed scores

The validity of rating scale was further examined by generating ICCs and CCCs (Figure 30). Similar to the ICCs in dataset 1, the current empirical curve is close to the model for most of the operational range of the scale. The empirical curve is within the 95% CI bands. The shape of

ICCs indicate that the inferences made from 5-scale rating categories to measure are well supported by the dataset 2.

The CCCs graph shows that each rating category (20, 30, 40, 50, 60) presents a clear peak with their usage in a similar frequency. The rightly ordered thresholds of the five categories (20 on the left, the lowest score and 50 on the right, the highest score) indicate that the rating scale categories functioned in an intended order. The graphs of ICCs and CCCs show that the rating scales were used by raters as intended. These further support the validity of rating constructs of ITAs' oral communication ability.

Bias/Interaction

The results of variable map, summary statitics, and graphs showed how each element for each facet fitted to the model and the valid function of rating categories. To understand the interaction between the *rating mode* and *rater* facets, current focus of this study, I conducted the bias/interaction analysis. The same question as in dataset 1 was asked, that is, whether the raters maintained their severity/leniency across two rating modes (VC, VR). The entire rating mode facet was dummied for the interaction analysis. Below is the model used for bias/interaction analysis:

Interaction model: rating mode (i.e., test delivery mode) x rater → rating residual

Prior to analyzing the full interactions, I first examined the bias/interaction size and significance which are useful to visually understand the distributional pattern, also a summary of the bias/interaction statistics in the following. Each bias estimation is produced in Figure 31 as vertical bars which shows the distribution of reported statistics.

The upper vertical bar chart, "*bias/interaction size*", shows the number of interactions found in different bias sizes and significance. The biggest number of bias/interaction size is 4, in

the 0 logits in size and near -1 SD in size. While there are a few outliers that require special attention, the majority of interactions are close to 0 (within ± 1SD).

The lower vertical bar chart, "*bias/interaction significance*", shows that 4 interactions have a bias/interaction significance (*t*-value) of about 0. The significance of 0, probability of $p < .001$ (double-sided), indicates that the 4 interactions did not happen by chance.

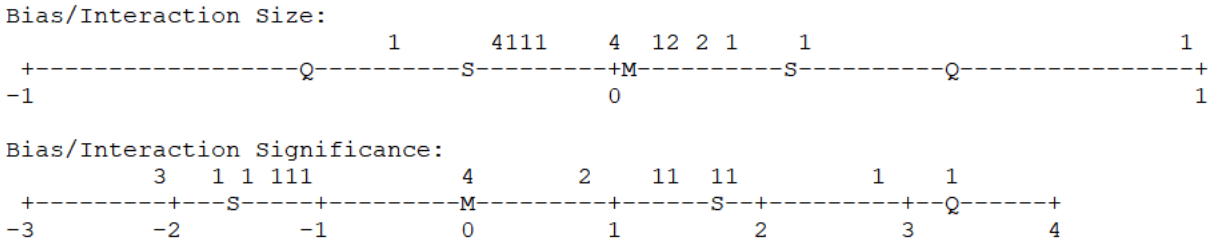


Figure 31. Vertical bar chart of bias/interaction size and significance from FACETS analysis

Next, I report the results of bias/interaction statistics. Table 30 displays the values of the bias/interaction terms, that is, it provides each rater's behavior across two test delivery modes on the entire dataset. The twelve raters who have most conspicuous interaction with the delivery mode are presented in an alphabetical order.

The "Observed-Expected average" decreased from live delivery mode to recorded delivery mode for all reported raters, except for raters G, L, Q, and R who rated only the recorded delivery mode. These raters' severity increased when rating the recorded delivery modes. Rater J had biggest change, who became most severe in the recorded mode (the gap between the live and the recorded delivery mode = -0.29), followed by rater C (-0.11). Raters B and M showed the smallest change (-0.05), however, they also became severe when assessing the speech samples in the recorded delivery mode.

Three raters showed noticeable interaction with the two delivery modes. Raters C and J had interaction with both delivery modes ($p < .05$), rater D had interaction with the recorded delivery mode ($p < .05$).

In terms of rater C's change in severity, I report a closer inspection from the Table 30. Rater C had an observed count of 336, which means this rater assessed speech samples in the live delivery mode for 336 times. Based on rater C's overall severity and the difficulty of the live delivery mode, the expected summed score was near 16,177.06. However, the observed ratings (16,200) were 0.07 higher than the expected average. Rater C has 0.07 rating-points leniency in the live delivery mode, which has 0.33 logits (bias size) with precision of 0.12 logits (model S.E.). A *t*-test with the hypothesis that this bias is due to measurement error with a null hypothesis, that there is no statistically discernable bias in rater C's ratings in the live delivery mode has a $t = 2.76$, rejected at $p < .05$ (two-sided). Overall, the results indicate that rater C does have noticeable bias in the live delivery mode.

In the recorded delivery mode, rater C assessed speech samples in the recorded delivery mode for 588 times, who had expected summed score around 27,222.85. However, the observed ratings (27,200) were 0.04 lower than the expected average. Different from the live delivery mode, rater C has 0.04 rating-points severity in the recorded delivery mode, which has -0.19 logits (bias size) with precision of 0.09 logits (model S.E.). The hypothesis for *t*-test is rejected ($t = -2.10$, $p < .05$). Rater C showed discernable bias in the recorded delivery mode as well, by becoming more severe than in the live delivery mode. Table 31 further contrasts each rater's behavior across delivery modes.

Lastly, I report the pairwise comparisons with effect sizes (Cohen's *d*) in Table 31. The eight raters assessed speech samples in both rating modes, and all of them gave higher scores in

the live mode than in the recorded mode. Among those, six raters (B, C, D, H, J, and M) showed statistically significant *p*-values for paired *t*-tests. I interpret the results of these six raters below.

Rater B: In the context of the live delivery mode, rater B is 0.64 lenient (i.e. awarding higher ratings; Obs-Exp Average = 0.02) but in the context of the recorded delivery mode, rater B is 0.38 less lenient (i.e., awarding lower ratings; Obs-Exp Average = -0.03). Overall, rater B is 0.26 more lenient in the live delivery mode than with the recorded delivery mode. The paired *t*-test shows that rater B's change in leniency is highly significant ($p = 0.05$) with a large effect size ($d = 0.08$).

Rater C: Rater C was 0.02 lenient in the live delivery mode (Obs-Exp Average = 0.07), while became 0.51 less lenient (Obs-Exp Average = -0.04) in the recorded delivery mode. "Target Contrast" shows that rater C was, overall, 0.52 more lenient in the live mode than with the recorded mode. This different leniency was statistically significant ($p = 0.01$) with a large effect size ($d = 0.25$).

Rater D: In the live delivery mode, rater D was 0.40 lenient (Obs-Exp Average = 0.03) and 0.04 less lenient in the recorded delivery mode (Obs-Exp Average = 0.04). The paired *t*-test shows that rater D's 0.36 leniency in the live delivery mode than in the recorded delivery mode was statistically significant ($p = 0.01$) with large effect size ($d = 0.50$).

Rater H: Rater H also was more lenient in the live delivery mode, with 0.73 leniency (Obs-Exp Average = 0.03) than in the recorded delivery mode in which the rater became 0.45 less lenient (Obs-Exp Average = -0.03). Overall, rater H had statistically significant 0.29 leniency in the live mode than in the recorded mode ($p = 0.05$) with small effect size ($d = 0.16$).

Rater J: Rater J was 1.67 lenient in the live delivery mode (Obs-Exp Average = 0.21) and 0.31 less lenient in the recorded delivery mode (Obs-Exp Average = 0.08). The paired t-test

shows rater J was 1.35 more lenient in the live mode with a *p*-value of 0.002 and small effect size ($d = 0.09$).

Rater M: In the live delivery mode, rater M had 0.08 leniency (Obs-Exp Average = 0.04) and 0.31 less leniency in the recorded mode (Obs-Exp Average = 0.04). Overall, rater M was 0.39 more lenient in the live mode than in the recorded mode, that the change was statistically significant ($p = 0.02$) with small effect size ($d = 0.16$).

Table 30.

Summary statistics of bias/interaction (rating mode x rater)

Rater	Mode	Observed score	Expected score	Observed count	Obs-Exp average*	Bias size	Model SE	*t*	*df*	Prob.**	Infit MnSq	Outfit MnSq
B	Live	32,120	32,105.29	659	0.02	0.11	0.09	1.25	658	0.21	1.2	1.3
B	Recorded	22,010	22,024.25	444	-0.03	-0.15	0.10	-1.47	443	0.14	1.2	1.2
C	Live	16,200	16,177.06	336	0.07	0.33	0.12	2.76	335	0.01	1.2	1.2
C	Recorded	27,200	27,222.85	588	-0.04	-0.19	0.09	-2.10	587	0.04	1.1	1.1
D	Live	35,120	35,098.36	708	0.03	0.15	0.08	1.82	707	0.07	0.9	1.0
D	Recorded	22,840	22,861.53	492	-0.04	-0.20	0.10	-2.08	491	0.04	0.9	0.9
G	Recorded	5,500	5,499.9	120	0.00	0.00	0.21	0.02	119	0.98	0.8	0.8
H	Live	19,310	19,296.91	396	0.03	0.16	0.11	1.45	395	0.15	0.8	0.8
H	Recorded	23,990	24,002.82	504	-0.03	-0.12	0.10	-1.26	503	0.21	0.8	0.7
J	Live	2,790	2,777.7	60	0.21	0.98	0.30	3.30	59	< 0.05	1.1	1.2
J	Recorded	6,770	6,782.22	144	-0.08	-0.37	0.18	-2.13	143	0.03	1.0	1.0
K	Live	28,040	28,030.99	600	0.02	0.08	0.09	0.83	599	0.41	0.7	0.6
K	Recorded	12,750	12,759.2	276	-0.03	-0.17	0.13	-1.24	275	0.21	0.8	0.7
L	Recorded	2,210	2,210.05	48	0.00	-0.01	0.33	-0.02	47	0.99	1.1	1.1
M	Live	16,480	16,465.71	348	0.04	0.19	0.12	1.66	347	0.10	1.0	1.0
M	Recorded	16,790	16,804.11	348	-0.04	-0.20	0.12	-1.68	347	0.09	1.0	1.0
P	Live	13,650	13,643.4	288	0.02	0.11	0.13	0.84	287	0.40	1.4	1.4
P	Recorded	6,640	6,646.25	143	-0.04	-0.20	0.18	-1.12	142	0.26	1.3	1.3
Q	Recorded	7,530	7,530.13	156	0.00	0.00	0.18	-0.03	155	0.98	0.8	0.8
R	Recorded	7,000	6,999.92	143	0.00	0.00	0.19	0.02	142	0.99	1.2	1.2

Note. Fixed (all = 0) chi-squared = 52.3; *df* = 20; $p < .001$. A chi-square test of the hypothesis that the biases presented in this Table xx are all the same except for measurement error (Linacre, 2021). The *p*-value under 0.001 indicates that the hypothesis is rejected, that the bias had significant impact in the raters' assessment across two test delivery modes.

Table 30 (cont'd)

*: the observed ratings subtracted by the expected ratings, on average. Example of rater C's Obs-Exp average calculation: (16,200-16,177.06)/336 = 0.07.

**: probability (*p*-value) *Note*. Only the most conspicuous interactions are displayed

Table 31.

Bias/interaction pairwise report (rater x rating mode)

Rater	Target Measr	SE	Obs-Exp Average	Test Mode	Target Measr	SE	Obs-Exp Average	Test Mode	Target Contrast	Joint SE	*t*	Welch *df*	*p*	Cohen's *d*
B	-0.64	0.09	0.02	Live	-0.38	0.10	-0.03	Recorded	-0.26	0.13	-1.93	956	0.05	0.08
C	-0.02	0.12	0.07	Live	0.51	0.09	-0.04	Recorded	-0.52	0.15	-3.47	707	0.01	0.25
D	-0.40	0.08	0.03	Live	-0.04	0.10	-0.04	Recorded	-0.36	0.13	-2.77	1081	0.01	0.50
H	-0.73	0.11	0.03	Live	-0.45	0.10	-0.03	Recorded	-0.29	0.15	-1.92	845	0.05	0.16
J	-1.67	0.30	0.21	Live	-0.31	0.18	-0.08	Recorded	-1.35	0.34	-3.93	102	0.002	0.09
K	0.29	0.09	0.02	Live	0.53	0.13	-0.03	Recorded	-0.24	0.16	-1.49	534	0.14	0.09
M	-0.08	0.12	0.04	Live	0.31	0.12	-0.04	Recorded	-0.39	0.17	-2.36	693	0.02	0.16
P	-1.16	0.13	0.02	Live	-0.85	0.18	-0.04	Recorded	-0.31	0.22	-1.40	284	0.16	0.13

Note. only the most conspicuous interactions are displayed.

CHAPTER 15: RESULTS OF DATASET 2: ITEM TYPES

Results

In this last section, I report on the findings of eight item types that were used in ITA Speaking test (dataset 2). As described in the Methodology section, these eight item types were differently distributed across five test forms in terms of frequency. I first report the descriptive statistics and paired samples *t*-tests, followed by CFA, measurement invariance, and MFRM analyses.

Table 32.

ITA Speaking Test means for item types

	VC mode		VR mode	
	Mean	*SD*	Mean	*SD*
Item type 1 (Comparison)	49.06	6.17	48.07	6.13
Item type 2 (Supported Opinion)	48.19	5.73	47.05	5.74
Item type 3 (Hypothetical)	50.41	6.30	48.38	6.17
Item type 4 (Comparison)	48.57	6.62	44.86	9.49
Item type 5 (Role Play)	47.61	6.61	46.58	7.01
Item type 6 (Description/Explanation)	48.61	6.21	47.74	6.78
Item type 7 (Graph presentation)	46.51	6.36	46.18	6.45
Item type 8 (Classroom announcement)	47.29	6.52	46.86	7.28

Note. Description of item types are in the parentheses.

Table 32 shows the mean and standard deviations of scores of each item type for each test delivery mode. Overall, item type 3 in the live mode had the highest mean score (M = 50.41), and item type 4 in the recorded mode had the lowest mean score (M = 44.86). For the mean scores within each delivery mode, item type 3 was the highest (M = 50.41) and item type 7 was the lowest (M = 46.51) in the live mode, while item type 3 was the highest (M = 48.38) and item type 4 (M = 44.86) was the lowest in the recorded mode. For all item types, scores were higher in the live delivery mode than in the recorded delivery mode.

Table 33.

Distribution of ITA Speaking Test scores

Table 33

	VC mode		VR mode	
	Skewness	Kurtosis	Skewness	Kurtosis
Item type 1	.137	-.531	.139	-.290
Item type 2	.037	-.218	.026	-.072
Item type 3	-.031	-.452	-.146	-.201
Item type 4	.067	-.435	.062	-.478
Item type 5	.157	-.525	-.012	-.314
Item type 6	-.049	.088	.019	-.498
Item type 7	.204	-.432	.177	-.418
Item type 8	.185	-.507	-.749	5.062

The distribution of the current dataset was examined using skewness and kurtosis. Table 33 shows that the scores for both delivery modes were normally distributed. Although item type 8 was within the range for parametric distribution, kurtosis in the recorded mode was high (5.062).

Results of Paired *t*-test

The paired *t*-test was conducted with ordinal variables, widely reported score type in standardized speaking tests such as IELTS (scores are generally reported by applying the rounding-down convention, cf. Nakatsuhara et al., 2020). To generate the ordinal variable, I rounded-down the decimals of the averaged scores. Table 34 shows that item types 6, 7, and 8 had non-significant differences between the scores of the VC mode and the VR mode. That is, only item types 1, 2, 3, 4, and 5 showed significantly higher scores in the live mode than in the recorded mode.

Table 34.

Paired t-test results for comparing each item type

		N	Mean	*SD*	*t*	*p*
Item Type 1	VC	284	48.99	6.16	3.17	0.001

Table 34 (cont'd)

	VR	284	47.84	5.96		
Item Type 2	VC	284	48.00	5.74	3.67	0.000
	VR	284	46.93	5.85		
Item Type 3	VC	169	50.41	6.30	2.89	0.004
	VR	169	48.76	6.47		
Item Type 4	VC	284	48.57	6.62	2.30	0.02
	VR	284	47.68	6.65		
Item Type 5	VC	284	47.61	6.61	2.66	0.008
	VR	284	46.48	7.00		
Item Type 6	VC	284	48.61	6.21	1.83	0.68
	VR	284	47.92	6.90		
Item Type 7	VC	284	46.51	6.36	0.53	0.60
	VR	284	46.30	6.52		
Item Type 8	VC	284	47.28	6.52	0.00	1.00
	VR	284	47.28	6.63		

Note. VC = video-conferencing (live, synchronous) mode; VR = video-recorded (recorded, asynchronous) mode.

Results of Confirmatory Factor Analysis (CFA)

Confirmatory factor analysis was conducted on dataset 2 at item type level, for all test-takers and the two test delivery modes (VC, VR) to examine the psychometric dimensionality of the ITA speaking test. Items were parceled into eight item types following the original design of the test. The item-parceling approach (i.e., a process of combining raw scores into subscales prior to analysis by averaging item responses into parcel scores, Meade & Kroustalis, 2006) provides advantages compared to an item-level CFA, such as the improvement if model estimation and fit/indicator reliability and the reduction in the number of parameters specifying a given latent factor (e.g., Dorans & Lawrence, 1999; Little, Cunningham, & Shahar, 2002; Sawaki & Sinharay, 2017).

I conducted the item type level approach because this study satisfied two criteria suggested by Meade and Kroustalis (2006) required for parcel-level analysis. First, the main focus of this study is to investigate the latent factor structure (i.e., prospective ITAs' oral communication ability) rather than the relationships of individual items to latent factors. Second, the unidimensionality of the items on which the parcels (item types) were theoretically designed within the ITA speaking test designers from pilot and main test scores.

In the following, I report the correlation matrices, fit statistics for the CFA models, and the results of measurement invariance. Since CFA is based on continuous variables, I used only the mean scores for the analysis.

Correlation Matrices

The correlation matrices in Table 35 indicate that the correlation coefficients were relatively stronger for each mode. The correlations coefficients show a wide range, from the smallest correlation coefficient of 0.22 to the strongest correlation coefficient of 0.73.

Table 35.

Correlation matrices for indicator variables (N = 285)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1 VC-IT1	1															
2 VC -IT2	0.73	1														
3 VC -IT3	0.64	0.67	1													
4 VC -IT4	0.64	0.76	0.65	1												
5 VC -IT5	0.56	0.63	0.52	0.53	1											
6 VC -IT6	0.61	0.68	0.54	0.62	0.48	1										
7 VC -IT7	0.50	0.58	0.58	0.47	0.52	0.47	1									
8 VC -IT8	0.53	0.64	0.59	0.55	0.57	0.53	0.56	1								
9 VR-IT1	0.63	0.59	0.50	0.43	0.41	0.43	0.44	0.46	1							
10 VR-IT2	0.56	0.71	0.55	0.57	0.48	0.55	0.57	0.52	0.75	1						
11 VR-IT3	0.43	0.51	0.49	0.41	0.38	0.40	0.40	0.45	0.61	0.66	1					
12 VR-IT4	0.26	0.33	0.40	0.30	0.37	0.22	0.26	0.32	0.27	0.40	0.39	1				
13 VR-IT5	0.52	0.54	0.57	0.40	0.43	0.44	0.43	0.48	0.64	0.66	0.58	0.26	1			
14 VR-IT6	0.47	0.60	0.50	0.54	0.44	0.55	0.46	0.52	0.66	0.74	0.63	0.28	0.65	1		
15 VR-IT7	0.43	0.49	0.46	0.37	0.53	0.41	0.42	0.39	0.61	0.65	0.52	0.39	0.49	0.60	1	
16 VR-IT8	0.43	0.55	0.57	0.42	0.51	0.43	0.53	0.57	0.57	0.69	0.60	0.33	0.59	0.65	0.68	1

Note. VC = video-conferencing mode; VR= video-recorded mode. IT is an acronym for item type. For example, "VC-IT1" is a label for item type 1 in the video-conferencing (live) delivery mode. Followed 'Type' denotes the item types (1 = Comparison, 2 = Supported Opinion, 3 = Recommendation, 4 = Hypothetical, 5 = Role Play, 6 = Description/Explanation, 7 = Graph Presentation, 8 = Classroom Announcement).

Model Fit Statistics

An overall model fit was examined for the two hypothesized CFA models, single-factor model (Figure 32) and correlated two-factor model (Figure 33). Table 36 indicates that the single-factor model has a poor fit with low CFI and high RMSEA. The correlated two-factor model shows an excellent fit considering the CFI, SRMR, and RMSEA values.

Table 36.

Fit statistic for the two models

Model	SBχ^2	df	CFI	SRMR	RMSEA	AIC	BIC
Single-factor	1973.377	120	0.837	0.967	0.131	16284.724	16434.959
Correlated two-factor (live and recorded)	1937.377	120	0.932	0.046	0.085	16108.249	16261.614

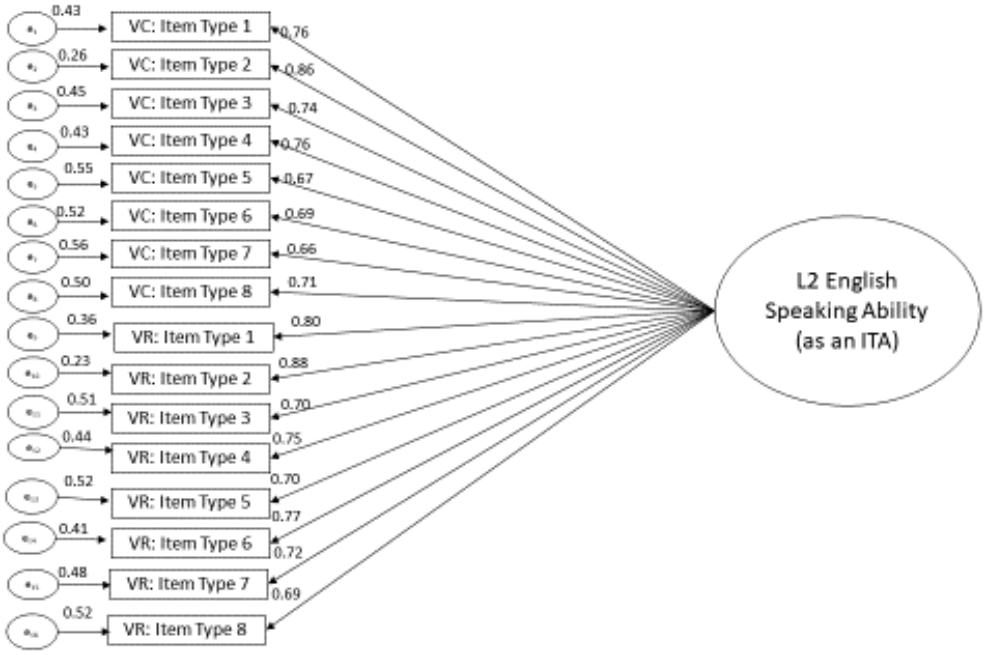


Figure 32. Single-factor model

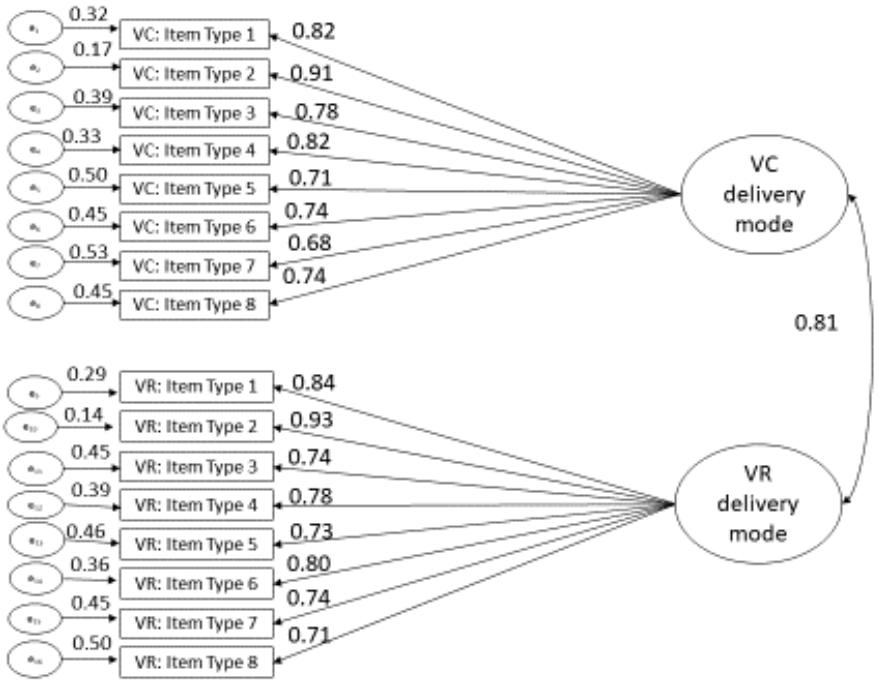


Figure 33. Correlated two-factor model (VC vs. VR)

Description of the Final Model

Figure 33 shows the diagram of the correlated two-factor model. The two latent factors are the live and the recorded delivery modes. The factor inter-correlations are above 0.80 (live and recorded: $r = 0.81$), which indicates poor discriminant validity (cf. Kline, 2016). This indicates that the live and recorded delivery modes are hard to distinguish from one another. The item loadings (weights) on each latent factor are presented on the straight arrow, which ranges between 0.42 to 0.90. The live delivery mode has relatively higher factor loadings which implies that the item types are more strongly loaded to the live delivery mode. The intercepts (means) of item types are around 7.5. The error variances and covariances of item types are displayed in small circles.

Results of Multifaceted Rasch model analysis (MFRM)

In this last section, I report the MFRM findings for item types with ordinal variables (mean scores that were applied with rounding-down convention). The analysis was conducted for a total of 4,321 measurable responses. The mean standard residual and mean standardized residuals are 0.00, and standard deviation is 1.00. The results indicate the successful estimation. The variance of the current dataset was explained by 40.75%, by Rasch measurement.

Description of the FACET Variable Map

The variable map (Figure 34) shows the information of the four target facets (*test-taker*, *rater*, *test mode*, *item*) in relation to the logits on a scale with 5 breaks or areas on the logit scale that are justified as separate bins or general levels. For the current MFRM analysis, all averaged scores were transformed into the 5 breaks (range from 20 to 60), a general approach to reporting scores. In the following, I interpret the variable map.

The current variable map shows "Measr" (for measurement) column that runs from +7 to -2 logits. The "test-takers" column (range: 6.81 to -1.70 logits) displays 285 prospective ITAs' distribution, with test-takers having higher probability of receiving higher scores on the upper column.

The "raters" column (range: +0.74 to -0.47 logits) shows somewhat different locations of the raters regarding their severity. This map represents that the raters L and Q were most severe while the raters H, J, P, and R were most lenient.

Of the most relevant and focus in this study, the "*test mode*" column (range: +0.11 to -0.11 logits) shows that the raters were harsher in the recorded delivery mode than in the live delivery mode. This result is the same with the MFRM analysis with items.

The “items” column (range: +0.42 to -0.35 logits) shows that item type 7 was most difficult, while item type 3 was the easiest. Item type 5 showed the approximately similar difficulty with item type 8. Item types 2 and 4 show similar difficulty, and item types 1 and 6 show similar difficulty.

Lastly, the scale shows that the scores are considered as ordinal with 5 breaks between 20 and 60. Test-takers with proficiency measures between about +7 to +5 logits were more likely to receive higher scores, around 60 (perfect score) to around 57.

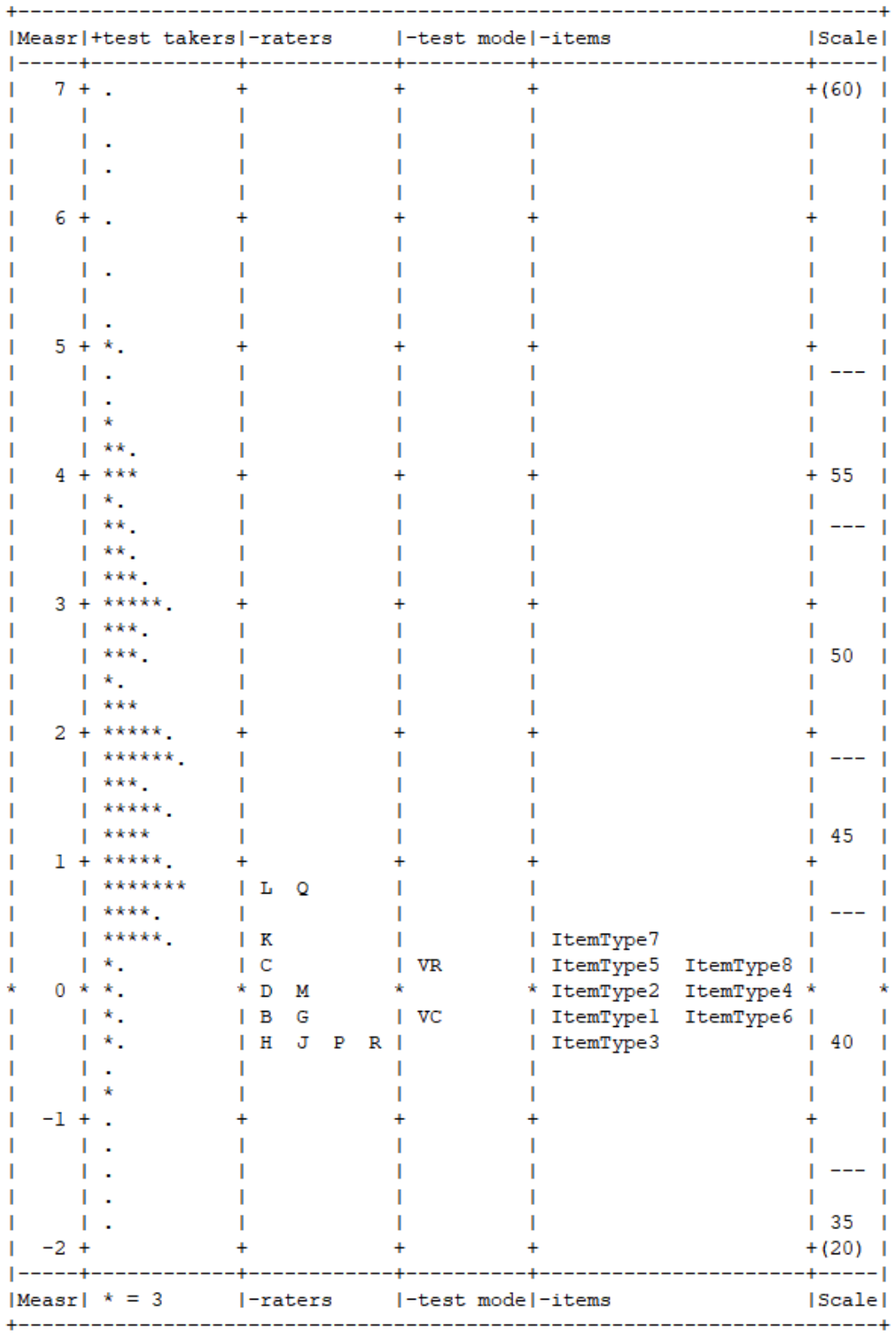


*Figure 34.*Variable map from FACETS analysis for Dataset 2

Note. VC = video-conferencing (live) mode; VR = video-recorded (recorded) mode.

Facets Summary Statistics

In this section, I report the summary statistics for the target facets. Table 37 presents the results.

Table 37.

Facets statistics summary

	Test-taker*	Test mode	Rater**	Item Type
M (measure)	1.91	0.00	0.00	0.00
SD (measure)	1.53	0.16	0.42	0.25
Model S.E.	0.33	0.03	0.09	0.05
Infit MnSq – Min, Max	0.02, 2.81	0.99, 1.01	0.64, 1.45	0.55, 1.31
Misfitting case (over 1.5)	39 cases (2.81, 2.76, 2.44, 2.73, 2.64, 2.37, 2.36, 2.40, 2.41, 2.28, 2.21, 2.27, 1.68, 1.98, 1.88, 2.01, 1.70, 1.87, 1.85, 1.97, 1.86, 1.95, 1.90, 1.90, 1.90, 1.75, 1.60, 1.83, 1.78, 1.79, 1.67, 1.78, 1.66, 1.68, 1.62, 1.63, 1.58, 1.52, 1.53)	None	None	None
Outfit MnSq – Min, Max	0.02, 2.99	0.98, 1.04	0.58, 1.53	0.58, 1.28
Misfitting case (over 1.5)	43 cases (2.99, 2.86, 2.82, 2.79, 2.75, 2.56, 2.53, 2.46, 2.44, 2.33, 2.33, 2.31, 2.27, 2.24, 2.13, 2.11, 2.03, 2.03, 2.01, 2.00, 1.87, 1.95, 1.80, 1.88, 1.79, 1.89, 1.86, 1.83, 1.74, 1.80, 1.79, 1.79, 1.79, 1.78, 1.74, 1.73, 1.73, 1.58, 1.61, 1.60, 1.60, 1.57, 1.54)	None	1 case (1.53)	None
RMSE	0.34	0.03	0.11	0.05
Adj. (true) SD	1.49	0.15	0.41	0.24
Separation	4.37	5.85	3.80	4.59
Strata	6.17	8.13	5.40	6.45
Separation reliability	0.95	0.97	0.94	0.95

*Test-taker is the only non-centered facet; includes extreme (i.e., perfect) scores.
**Inter-rater agreement opportunities = 0

All facets were in productive measurement range except for the *test-taker* facet, which shows 39 misfitting cases with infit mean-square higher than 1.5. Overall, the "Separation" value indicates that *test-taker* facet had 4.37 distinguishable levels, test mode facet had 5.85 distinguishable levels, *rater* facet had 3.80 distinguishable levels, and *item type* had 4.59

distinguishable levels. The detailed summary statistics of each facet are presented in the following. All summary statistics for each facet are sorted by infit mean-square values.

Results of the 4-Facet Analysis

Identical to the structure of the results reported above, I report the findings of the four target facets: test-taker (Table 38), rater (Table 39), rating mode (Table 40), and item type (Table 41).

Table 38.

Test-taker summary statistics

Test-taker ID	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
177	41.07	41.50	0.20	0.38	2.81	2.99	0.47	0.29
219	59.64	59.84	6.68	0.84	0.76	2.86	0.78	-0.38
161	49.06	48.26	1.96	0.34	2.76	2.82	0.48	0.22
143	48.44	48.05	1.89	0.32	2.44	2.79	0.47	0.42
114	46.25	46.22	1.41	0.28	2.73	2.75	-0.15	-0.13
138	44.69	43.50	0.78	0.27	2.64	2.56	-0.55	0.33
099	40.71	40.23	-0.36	0.40	2.37	2.53	0.50	-0.41
151	52.50	52.57	3.52	0.29	2.36	2.46	0.04	-0.05
210	48.93	48.88	2.17	0.36	2.40	2.44	0.32	0.07
107	50.00	49.44	2.39	0.35	2.41	2.33	0.79	0.66
109	50.94	50.49	0.34	0.34	2.28	2.33	0.45	0.58
038	50.00	50.32	2.77	0.35	2.21	2.31	0.63	0.17
250	39.69	40.21	-0.37	0.37	2.27	2.27	0.70	0.23
062	38.57	39.03	-0.91	0.34	1.68	2.24	0.72	-0.14
156	51.25	51.30	3.14	0.32	1.98	2.13	0.58	0.40
167	52.19	52.12	3.40	0.30	1.88	2.11	0.61	0.70
154	45.63	44.32	0.98	0.27	2.01	2.03	-0.59	-0.37
214	53.21	55.61	4.22	0.30	1.70	2.03	0.16	0.04
096	50.71	50.42	2.87	0.37	1.87	2.01	0.68	0.08
162	46.25	48.10	1.90	0.29	1.85	2.00	0.63	0.66
222	48.21	47.90	1.84	0.34	1.97	1.87	0.65	-0.46
144	47.50	46.30	1.43	0.30	1.86	1.95	0.45	0.19
221	41.43	41.90	0.34	0.36	1.95	1.80	0.56	0.54
211	47.14	45.87	1.33	0.31	1.90	1.88	0.32	-0.38
165	51.88	51.81	3.30	0.31	1.90	1.79	0.49	0.51
231	53.44	53.08	3.65	0.28	1.90	1.89	0.27	0.12
209	49.29	48.45	2.02	0.37	1.75	1.86	0.94	-0.20
074	41.07	40.66	-0.16	0.38	1.60	1.83	0.63	-0.17
102	48.57	47.93	1.85	0.35	1.83	1.74	0.64	0.51
254	36.56	36.53	-1.59	0.26	1.41	1.80	0.58	0.83

Table 38 (cont'd)

248	51.56	51.28	3.13	0.32	1.78	1.79	0.51	0.01
173	47.14	45.62	1.27	0.31	1.79	1.79	0.42	-0.14
124	48.13	48.91	2.18	0.31	1.67	1.79	0.94	0.23
084	50.00	50.70	2.92	0.38	1.78	1.78	0.51	0.03
183	49.29	49.58	2.45	0.37	1.66	1.74	0.97	-0.03
149	52.50	52.57	3.52	0.29	1.68	1.73	0.52	-0.09
272	40.31	39.74	-0.60	0.37	1.62	1.73	0.88	0.29
098	46.43	45.50	1.24	0.30	1.63	1.58	-0.08	-0.21
230	51.25	50.88	2.99	0.33	1.50	1.61	1.03	0.30
092	51.07	50.23	2.73	0.35	1.50	1.60	0.66	0.59
125	50.63	50.09	2.66	0.33	1.58	1.60	0.82	0.54
123	49.69	49.79	2.54	0.34	1.52	1.57	0.90	0.09
139	43.13	43.60	0.81	0.29	1.44	1.54	0.98	0.64
181	45.36	46.63	1.50	0.29	1.53	1.50	0.05	0.18
025	52.81	51.70	3.27	0.28	1.35	1.50	0.52	0.26
126	49.69	48.91	2.18	0.36	1.50	1.49	0.96	0.44

Note. Fixed (all same) chi-squared: 5931.4; $df = 284$; $p < .001$

Overall, Table 38 shows that there are 39 test-takers who showed infit mean-squares higher than 1.5. Of those, 13 test-takers (IDs 177, 161, 143, 114, 138, 099, 151, 210, 107, 109, 038, 250, 154) had infit mean-squares higher than 2.0. Test-taker 177 had the highest infit mean-square (2.81) and had an observed score of 41.07 and fair average of 41.50.

Table 39.

Rater summary statistics

Rater	Total score	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrim.	Point measure *r*
P	12,635	46.80	49.37	-0.47	0.07	1.45	1.53	0.70	0.70
R	4,350	48.88	49.08	-0.35	0.13	1.44	1.43	0.85	0.42
L	1,365	45.50	45.11	0.74	0.23	1.23	1.36	1.08	0.63
B	34,880	49.20	48.80	-0.24	0.05	1.23	1.30	0.78	0.79
C	27,355	46.92	47.44	0.18	0.05	1.08	1.09	0.96	0.83
J	6,110	46.64	49.06	-0.34	0.10	1.06	1.09	0.87	0.60
M	20,960	47.85	47.74	0.10	0.06	1.01	1.01	0.96	0.63
D	36,815	48.25	48.30	-0.08	0.04	0.92	0.96	1.02	0.78
Q	4,705	48.01	45.09	0.74	0.13	0.82	0.84	1.13	0.87
H	28,145	48.36	49.19	-0.39	0.05	0.83	0.78	1.19	0.72
G	3,195	45.64	48.64	-0.19	0.15	0.65	0.62	1.01	0.88
K	25,975	46.55	46.98	0.30	0.05	0.64	0.58	1.23	0.79

Note. Fixed (all same) chi-squared = 232.7; $df = 11$; $p < .001$

Table 39 shows that the twelve raters had productive infit mean-square ranges. In terms of rater severity, rater J showed the most severity among the eight raters who judged both delivery modes. Rater D showed the most leniency. Within the four raters (raters G, L, Q, R) rater L had more severity, while rater Q had more leniency.

Table 40.

Rating mode summary statistics

Mode	Total score	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrim.	Point measure *r*
Live	103,960	48.20	48.41	-0.11	0.03	1.01	1.04	1.00	0.76
Recorded	102,530	47.38	47.70	0.11	0.03	0.99	0.98	1.00	0.78

Note. Fixed (all same) chi-squared = 35.2; $df = 1$; $p < .001$

The two delivery modes in Table 40 have productive infit mean-squares. The raters were more lenient in the live delivery mode (VC) than in the recorded mode (VR), as indicated by the total score.

Table 41.

Item type summary statistics

Item type	Total score	Observed raw score average	Fair average	Difficulty measure (in logits)	SE	Infit mean square	Outfit mean square	Estimated discrimin.	Point measure *r*
Item type 5	26,760	47.03	47.26	0.23	0.05	1.31	1.28	1.13	0.73
Item type 7	26,400	46.40	46.52	0.42	0.05	1.13	1.13	1.22	0.73
Item type 3	16,800	49.56	49.08	-0.35	0.07	1.12	1.11	1.25	0.73
Item type 6	27,455	48.25	48.47	-0.13	0.05	1.08	1.11	1.10	0.75
Item type 8	26,850	47.27	47.53	0.16	0.05	1.08	1.04	1.27	0.75
Item type 4	27,375	48.11	49.35	-0.09	0.05	1.03	1.03	0.84	0.77
Item type 1	27,565	48.44	48.64	-0.19	0.05	0.77	0.84	0.73	0.80
Item type 2	27,285	47.95	48.20	-0.04	0.05	0.55	0.58	0.55	0.85

Note. Fixed (all same) chi-squared = 146.8; $df = 7$; $p < .001$

Table 41 shows the eight item types with productive infit mean-square values (sorted by infit mean-square values). Test-takers had more difficulty receiving higher scores in item type 3, while they were more likely to receive higher scores in item type 5.

Empirical and Expected ICCs and CCCs for ITA Speaking Test Indicators

(a) (b)

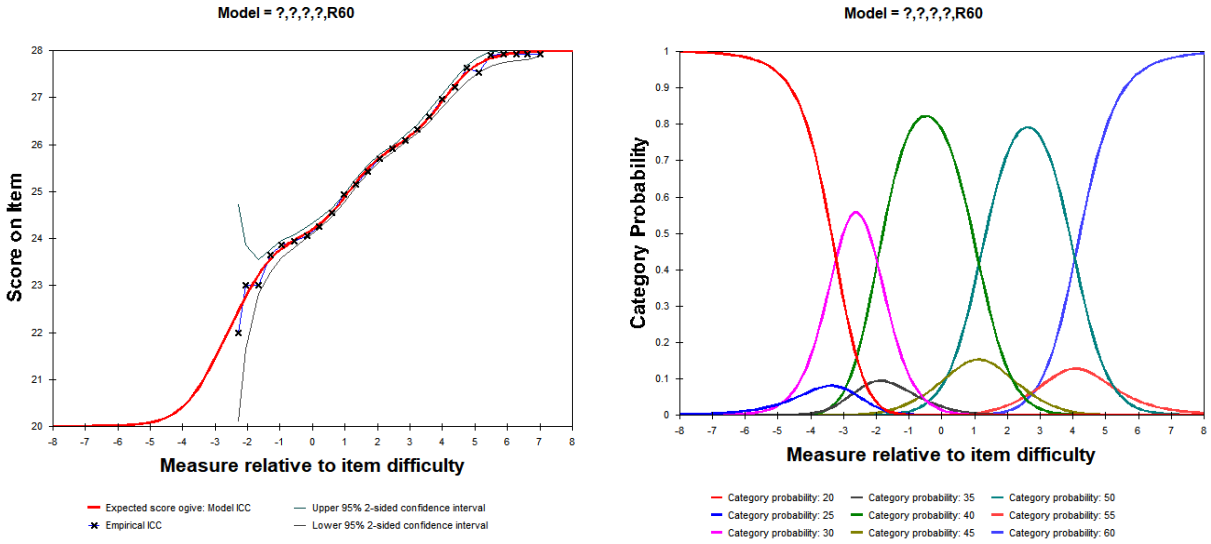


Figure 35. (a) Empirical and expected ICC, (b) category probability curves and observed scores

Figure 35 displays the ICCs and CCCs for further inspection of the validity of rating scale using ordinal variables. The ICCs graph shows that empirical ICC is within 95% CI bands, and closely follows the model ICC. The shape of ICCs indicate that the inferences made from the ordinal variable (i.e., rounded-down mean scores, cf. Nakatsuhara et al., 2020, p.13) are well supported by the current dataset.

The CCCs graph shows clear peaks (Figure 35, b). However, not all "categories" are clearly peaked, for example, less frequently used show low probability (i.e., low peak). The categories are rightly ordered, however, they overlap with each other. Hence, it should be noted that the intended order of the rating scale may not function as clearly as the rating scales in

dataset 1 or dataset 2 for item-level. Overall, the CCCs with ordinal variable show a relatively clear function of the categories. Nevertheless, the different height of peaks indicate that the categories are not used in a similar frequency, which implies that the validity of the rating construct using ordinal variables is not strongly supported.

Bias/Interaction

The results of the variable map, summary statistics, and ICCs support the model-to-data fit for the target facets and elements. The shape of CCCs, however, indicate that the rating construct is not fully supported. Regardless of the weak support from CCCs, I report the bias and interaction findings.

The same question, whether the raters are consistent with their severity/leniency across two rating modes (live, recorded), was asked. The same model is used:

Interaction model: rating mode x rater → rating residual

In the vertical bar chart (Figure 36), the upper chart ("bias/interaction size") shows the biggest number of bias/interaction size is 9 in the 0 logits in size, near the mean value (M).

The "*bias/interaction significance*" shows there are about 5 interactions that have a bias/interaction significance (*t*-value). Four interactions have a bias interaction with significance of 0 and one interaction has a bias interaction with significance of approximately $p = 0.02$ (double-sided). These five interactions imply that these did not happen by chance.

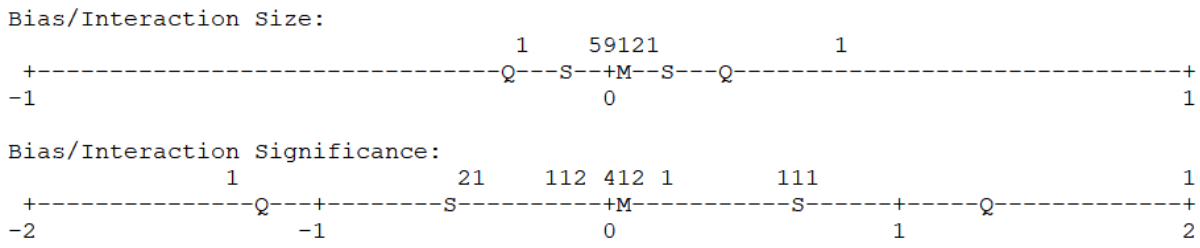


Figure 36. Vertical bar chart of bias/interaction size and significance from FACETS analysis

Lastly, I report the bias/interaction statistics. Table 42 indicates different patterns of "Observed-Expected Average" across the eight raters who assessed both delivery modes. Three raters showed a decrease of "Observed-Expected Average" from the live to the recorded mode (rater C: 0.03 to -0.02, rater J: 0.28 to -0.12, rater P: 0.01 to -0.01) and three other raters showed an increase from the live (VC) to the recorded (VR) mode (rater B: -0.02 to 0.03, rater K: -0.02 to 0.04, rater M: -0.01 to 0.01). Two raters had the same "Observed-Expected Average" in the two delivery modes (rater D: 0.00, rater H: 0.01).

Only one rater, J, showed statistically significant interaction with the live delivery mode ($p = 0.05$). Rater J's expected summed score was around 1,769.18, but the observed score was 0.28 higher in average. The *t*-test result shows that rater J had bias when rating the live delivery mode: J was more lenient in the live delivery mode.

Table 43 shows the bias/interaction pairwise with effect size. As the raters' "Observed-Expected Average" indicated, their leniency/severity differed across delivery modes. Raters C, J, and P, who had decreased "Observed-Expected Average" showed higher scores in the live delivery mode (VC). Raters B, K, and M, who had increased "Observed-Expected Averaged" showed higher scores in the recorded delivery mode (VR). Rater D gave approximately the same scores across two modes, while rater H gave slightly higher scores in the recorded delivery mode.

The *t*-test results indicate that only rater J showed significantly higher scores in the live delivery mode (VC) than in the recorded delivery mode (VR) ($p = 0.02$). Rater J in this dataset was 0.54 more lenient in the live delivery mode (VC), and J's change was significant with small effect size ($d = 0.23$).

Table 42.

Summary statistics of bias/interaction (rater x rating mode)

Rater	Mode	Observed score	Expected score	Observed count	Obs-Exp average	Bias size	Model SE	*t*	*df*	Prob.	Infit MnSq	Outfit MnSq
B	VC	20485	20493.78	420	-0.02	-0.03	0.06	-0.52	419	0.60	1.2	1.3
B	VR	13915	13906.01	281	0.03	0.04	0.07	0.64	280	0.53	1.2	1.2
C	VC	10220	10212.67	213	0.03	0.05	0.08	0.60	212	0.55	1.1	1.1
C	VR	17135	17142.22	370	-0.02	-0.03	0.06	-0.46	369	0.64	1.1	1.1
D	VC	22225	22223.77	449	0.00	0.00	0.06	0.07	448	0.94	0.9	1.0
D	VR	14590	14591.13	314	0.00	-0.01	0.07	-0.08	313	0.94	0.9	0.9
G	VR	3195	3194.96	70	0.00	0.00	0.15	0.01	69	0.99	0.7	0.6
H	VC	12350	12352.78	253	0.01	-0.02	0.08	-0.21	252	0.83	0.9	0.9
H	VR	15315	15312.07	321	0.01	0.01	0.07	0.20	320	0.84	0.8	0.7
J	VC	1780	1769.18	39	0.28	0.39	0.20	1.98	38	0.05	1.2	1.4
J	VR	4330	4340.79	92	-0.12	-0.16	0.12	-1.29	91	0.20	1.0	1.1
K	VC	17650	17657.47	378	-0.02	-0.03	0.06	-0.48	377	0.63	0.6	0.6
K	VR	8325	8317.54	180	0.04	0.06	0.09	0.69	179	0.49	0.7	0.6
L	VR	1365	1365.01	30	0.00	0.00	0.23	0.00	29	0.99	1.2	1.4
M	VC	10305	10306.32	218	-0.01	-0.01	0.08	-0.11	217	0.92	1.0	1.1
M	VR	10655	10653.58	220	0.01	0.01	0.08	0.12	219	0.91	1.0	1.0
P	VC	8465	8463.59	179	0.01	0.01	0.09	0.12	178	0.90	1.3	1.5
P	VR	4170	4171.28	91	-0.01	-0.02	0.12	-0.16	90	0.88	1.7	1.7
Q	VR	4705	4705.03	98	0.00	0.00	0.13	0.00	97	0.99	0.8	0.8
R	VR	4350	4349.96	89	0.00	0.00	0.13	0.01	88	0.99	1.4	1.4

Note. Fixed (all = 0) chi-squared = 7.7; *df* = 20; *p* = 0.99. Different from the results of chi-square tests in this study, the *p*-value indicates that the chi-square is not rejected. That is, the bias did not have significant impact in the raters' assessment of item types across two delivery modes.

Higher score = higher bias measure

Table 43.

Bias/interaction pairwise report (rater x rating mode)

Rater	Target Measure	SE	Obs-Exp Average	Test Mode	Target Measure	SE	Obs-Exp Average	Test Mode	Target Contrast	Joint SE	*t*	Welch *df*	*p*	Cohen's *d*
B	-0.21	0.06	-0.02	VC	-0.29	0.07	0.03	VR	0.08	0.09	0.82	608	0.41	0.34
C	0.13	0.08	0.03	VC	0.21	0.06	-0.02	VR	-0.08	0.10	-0.76	453	0.45	0.01
D	-0.08	0.06	0.00	VC	-0.07	0.07	0.00	VR	-0.01	0.09	-0.10	686	0.92	0.26
H	-0.38	0.08	-0.01	VC	-0.41	0.07	0.01	VR	0.03	0.10	0.29	540	0.77	0.35
J	-0.73	0.20	0.28	VC	-0.19	0.12	-0.12	VR	-0.54	0.23	-2.37	68	0.02	0.23
K	0.33	0.06	-0.02	VC	0.24	0.09	0.04	VR	0.09	0.11	0.84	353	0.40	0.34
M	0.11	0.08	-0.01	VC	0.09	0.08	0.01	VR	0.02	0.11	0.16	435	0.87	0.27
P	-0.48	0.09	0.01	VC	-0.45	0.12	-0.01	VR	-0.03	0.15	-0.20	183	0.84	0.24

CHAPTER 16: RATERS' VERBAL REPORT

Results

I now move on to reporting the results of eight raters' verbal report data in dataset 1. Based on the qualitative findings, I compared the raters' perceptions across the audio and the two video modes (video1, video2). I present the final coding scheme which includes four main themes. I present each theme and the exemplary excerpts below.

Theme 1

Most frequently mentioned by the raters was that the video modes (video1, video2) provided more information than the audio mode, which helped them to understand test-takers' performance beyond what they can hear. Facial expressions, eye-gaze, and head orientation were commonly mentioned nonverbal cues that affected raters' perception towards test-takers, signaled test-takers' struggle during their responses, and showed test-takers' focus during their responses.

For example, when background noise was too loud (e.g., test-takers had to take the test in the lab, some test-takers who had children took test at home or at a café), these nonverbal cues helped raters to overcome the noise and guess the content. The test-takers' visual information supported raters to assign scores and provided fuller context of what test-takers were doing or focusing on when responding. Several remarks were made about this, such as:

> "Eye gaze was the number one thing that I paid attention to … [for example,] people would look down more if they were having language-related problems as opposed to memory-wise where they would just shift back and forth." (Rater 7, video modes)

> "I preferred video recordings because I had a chance to not only listen to what the test-taker was saying but also see their verbal expressions. It especially helped me a lot when assigning the fluency category." (Rater 11, video modes)

> "There was a test-taker who was in a café and there was jazz music on the background. She had good English skill so it didn't affect her speaking performance,

> but if I had her in the audio mode, I may have had great difficulty understanding her. When I saw her that she was in a café and she signaled some cues when the music got louder, it was okay for me to understand what she was saying." (Rater 1, video modes)

While positive influence of visual information was mentioned by several raters, three raters mentioned there was no difference between the video1 and video2 modes. However, they all agreed that rating was easier when nonverbal cues were displayed. For example, rater 5 mentioned that "different from the audio mode, it felt easier to rate when I was able to see test-takers' faces".

Theme 2

Between the video modes, four raters mentioned that video2 mode was more helpful and straightforward for them to assess test-takers' performance, while only one rater mentioned it was video1. The video2 was preferred mainly because of the absence of examiner' visual information, which allowed raters to focus on test-takers' performance with less distraction:

> "Because there were a couple of times where I would pick up on examiner's facial expressions rather than test-takers', I tried to ignore what the examiner was doing [in the video1 mode]. So, in the video2 mode, there was enough information and less distractions … more straightforward and maybe less interference [when rating]." (Rater 2, Video2)

Interestingly, while several raters mentioned that examiner's visual information was distracting, all raters frequently mentioned that the video1 mode provided fuller picture representing an authentic and real-life oral communication. For example:

> "[Video1 mode] allowed me to get a wider picture of everything […] I felt it was a bit more real in the sense that [a test] taker was trying to engage with the examiner." (Rater 1, video1)

> "I didn't think there was a difference in rating [between the video1 and video2 modes] but I thought test-takers' engagement with an examiner seemed to better show test-takers' speaking ability because they are actually talking to someone." (Rater 9, video1)

Theme 3

Although the video modes were more helpful than the audio mode when rating, this wasn't always the case. When visual information was distracting, or because visual features were not clearly described in the rubric, raters sometimes preferred the audio mode to the video modes. This happened when the raters were giving scores for each linguistic category:

> "Being able to only listen to the audio was better, because I didn't have to switch back and forth between screen and rubric. I constantly looked at the rubric while listening." (Rater 10, audio mode)

> "The audio mode helped me better to rate the test-takers. I sometimes found the visual cues kind of distracting, like their hand gestures" (Rater12, audio mode)

> "The audio mode allowed me to concentrate more on [test-takers'] language, really focus on what they are saying without any sort of distractions" (Rater 7, audio mode)

Another case of the audio mode being sufficient was related to the raters' familiarity and their experience with assessing audio-recorded speech samples:

> "Most of the raters including myself, have experience as raters to give scores just by listening to the speech samples. Video modes are somewhat new, and I am not used to rate test-takers with their visual information included." (Rater 5, audio mode)

Theme 4

This theme emerged from remarks about the raters' personal preference of visual information that, in turn, affected their preferences for the test-takers. Several nonverbal features were mentioned, for example, test-takers' eye gaze, head orientation, which raters mentioned were related to test-takers' willingness to engage in the task and confidence in what they are saying. For example:

> "I liked the speaker whom I felt were confident in what they were saying. For example, when a test-taker looked straight to the camera, it made me feel like he knew what he was saying and that he was giving out those answers" (Rater 1, video modes)

> "There were test-takers whom I really enjoyed listening to …they looked straight into the camera and seemed focused … they were engaging to listen to" (Rater 2, video modes)

There were nonverbal features raters did not prefer, such as test takers who looked away from the camera (e.g., looking down or avoiding eye contact with camera) or test-takers who were wearing masks on which made it difficult to read their lip movements. Another visual information was test-takers' virtual backgrounds that were funny or messy, which made them look childish or not prepared. For some raters, these backgrounds distracted focusing on test-takers' performances. Below are several excerpts about non-preferred visual information:

> "There were others not many but there was a test-taker who felt weird because I couldn't see her face because she kept her head down. I couldn't really see what she was trying to say" (Rater 1, video modes)

> "Another test-taker had a funny background, and it was distracting because it made her look a bit like childish. So, I think in those senses if they have a messy background, [it] can be distracting and gives you an impression of the personality in a negative way." (Rater 1, video modes)

> "In some cases, there were test-takers who looked the other way, [they] didn't look straight into the camera. In that case, I didn't feel comfortable because I had to put more energy [to understand what they are saying]. So, I personally preferred the clear visual display of a test-taker." (Rater 5, video modes)

CHAPTER 17: DISCUSSION OVERVIEW

The purpose of the current study was three-fold: to (a) investigate the latent structure of the video-conferenced speaking proficiency tests using the test-takers' oral performance data, (b) compare raters' scores when assessing test-takers' oral performances under different test delivery modes (dataset 1: audio-only, video1, video2, dataset 2: video-conferencing, video-recorded), and (c) explore raters' perspectives on the three different test delivery modes when assessing oral performances (dataset 1). In Chapters 13 through 16, I reported the findings of the three research questions. For the first two research questions, I interpreted the findings of the dataset 1 in Chapter 13, and discussed the results of the dataset 2 in Chapters 14 and 15. Regarding the third research question, I interpreted the findings of the raters' verbal data in Chapter 16.

In this chapter, I discuss the findings of datasets 1 and 2 considering the findings of previous studies and theoretical backgrounds, specifically regarding the influence of the participants' visual information in the video modes on raters' behavior and how different characteristics of different delivery modes (e.g., synchronous/asynchronous, cognitive load) affected rater behavior. I also discuss the findings in relation to the importance of nonverbal features in the speaking construct, and the construct validity of the delivery modes in speaking tests.

For the sake of clarity, I follow the same order of the Data Findings sections (Chapters 18, 19, and 20). In Chapter 18, I first discuss the results for RQs 1 and 2 of the dataset 1. Followed by is Chapter 19, where I discuss the results for RQs 1 and 2 of the dataset 2. Lastly, in Chapter 20, I discuss the findings of raters' verbal reports in dataset 1, which answers RQ 3.

CHAPTER 18: DISCUSSION ON DATASET 1

Delivery Modes and the Speaking Construct

Considering the underlying structure of the speaking construct focused on this study (i.e., both verbal and nonverbal features contribute to the speaking ability being measured), I conducted confirmatory factor analysis (CFA) to investigate the construct validity of speaking tests with different delivery modes. So far, using factor analysis for speaking test delivery modes has not been the major focus of previous research, with an exceptional study by Zhou (2015) who conducted exploratory factor analysis (EFA) to compare the face-to-face and audio modes.

Prior to discussing the CFA findings, it is worth noting that test-tasks are generally derived from the definition of the construct to be measured. The rationale of representing the defined construct is to ensure that all tasks capture the construct aimed at and only for this construct (Ziegler & Hagemann, 2015). Consequently, tasks belonging together in a scale are expected to capture differences in the same underlying construct. In this study, the speaking tasks were adopted from the Oral Proficiency Interview-computer (OPIc), a test developed based on an approach where each task measures English language oral communication ability by modeling the process of language use (Spolsky, 1985) in the indirect (asynchronous) mode (conventionally used is the virtual format in which an avatar gave questions, and test-takers took the test in a designated space with computers. However, after the outbreak of COVID-19, test-takers take the OPIc in their home, with their personal computer or tablet; for detail, see this link: https://www.languagetesting.com/test-delivery-logistics).

Among the three plausible latent models in this current study, the CFA findings indicated that a correlated three-factor model that consisted of audio, video1, and video2 modes as latent factors was the best fitting model with an excellent fit. This also indicates that other two latent

models, single-factor model and correlated two-factor model, have failed to support their hypotheses.

Regarding the single-factor model, the poor fit failed to support the unidimensionality of the speaking construct. While there has been no research conducted to investigate the comparability of speaking test delivery modes using CFA, the current finding provides counterevidence of what very few previous studies have suggested about the delivery modes, regarding the unidimensionality of speaking assessment. For example, Nakatsuhara et al. (2017) stated that from their MFRM findings, a lack of misfit across rating scales provides indirect evidence of unidimensionality, in that the modes (audio-recorded, video-recorded, video-conferencing) are measuring the same construct. In another study, Zhou (2015) stated that EFA evidenced the unidimensional factor structure across the two delivery modes (face-to-face, audio-recorded).

However, the CFA findings of the unidimensionality here indicate a different story, that L2 English speaking ability measured across the tasks of different delivery modes are not unidimensional. Simply put, speaking construct measured by tasks with different difficulties (intermediate level, superior level) across three delivery modes (audio, video1, video2) are not considered unidimensional, because the systematic differences within the task variance are caused by more than one variance source (i.e., one latent variable, that is, one type of test delivery mode) (cf. Rubio, Berg-Weger, & Tebb, 2001; Ziegler & Hagemann, 2015). This CFA finding revealed that the tasks of different modes are not measuring one attribute of the defined speaking construct. This is not surprising, considering the previous research that demonstrated nonverbal behaviors are a component of the construct of speaking, distinct from linguistic features (e.g., Choi & Lantolf, 2008; Gullberg, 2006; Jenkins &Parra, 2004; Kendon, 2004).

In particular, video delivery modes naturally expand the construct of speaking being measured; demonstrating test-takers' competence in the use of linguistic features (e.g., vocabulary, fluency, grammar) is not sufficient, they must also communicate efficiently with the interlocutor by coping with various social situations (cf. Kendon, 1981). In the video modes, test-takers transmit information using their nonverbal behavior which is different from linguistic components. The nonverbal behaviors (e.g., facial expression, eye gaze, manner of speaking, clothing) contribute to the interaction between participants, whether the message is intended or not (Goffman, 1963). For example, during speech production the test-taker looks at the examiner's nonverbal behavior – head nodding, responding at the boundaries of speech units – which sends the test-taker a message that the similar unit of the produced speech is received by the examiner (e.g., Allen & Guy, 1977; Beattie, 1978; Kendon, 1967). Thus, from this point of view that nonverbal behaviors (e.g., gesture) are a symbolic language which has an arbitrary form-meaning pairs (Kendon, 2000; Saussure, 1959), gestures are understood to create additional meaning to the spoken utterance which, in turn, may overcome the limitations imposed by the speech to some extent (Kendon, 2000).

Next, the correlated two-factor model failed to support that the two latent factors (i.e., L2 speaking ability in the audio and the video modes) measure the same speaking construct. While the correlated two-factor model was not supported, it is worth mentioning that the model fit was not excellent but showed its potential for good fit. Taken altogether, the failed support of the unidimensionality and the two latent factors indicate that there may be distinguished differences across the three delivery modes. The support from the three correlated-factor indicates that these three delivery modes may measure the same underlying speaking construct (i.e., test-takers' English speaking ability which consists of verbal and nonverbal aspects).

Within the three correlated-factor model, the high intercorrelations across the three latent variables further supported that L2 speaking abilities across three delivery modes (audio, video1, video2) potentially measure the same speaking construct. Thus, it is important to note that the three delivery modes may not be statistically distinct test-taking contexts. That is, the findings of CFA demonstrate that whether the speaking test is held in conceptually distinguishable delivery modes, they are statistically inseparable when actual speaking is held. Overall, the three delivery modes do not discriminate the underlying speaking construct being measured.

I suggest two potential reasons for the high inter-correlations across the latent variables within the three correlated-factor model. First, the use of same rubric and tasks across the three different modes may have caused the high inter-correlations. It is likely that using the same rubric and tasks overcame the differences (e.g., absence/presence of visual information) imposed by different rating modes, leading to high correlations. For all three modes, the rubric included only the analytic scales for linguistic features (fluency, lexis, grammar), which is suited for the audio-only mode that measures the narrowest range of the speaking construct (e.g., Fulcher, 2003; Nakatsuhara et al., 2020). Hence, the constrained focus on verbal features may have affected raters' final scores. For example, even in the video1 mode that provides the most extensive information, because raters were only required to focus on the narrowest range of test-takers' speaking ability, the scores across audio or video modes were likely to be similar. In sum, because the instruments (rubrics, tasks) used were the same for all delivery modes, it could have led to the high inter-correlations among latent factors.

Second, the high inter-correlations between the two video modes imply that the absence or presence of the examiner's visual information may not have significantly influenced the final scores test-takers received. In other words, whether the examiner's visual information was

present (video1) or not (video2), rater could have mainly focused on test-takers' visual display and assessed their L2 speaking ability by watching their oral performance. If this was the case, the examiner's visual information was likely to be disregarded by raters when awarding scores (e.g., Nakatsuhara et al., 2020). Furthermore, because the tasks were monologic, raters didn't have to consider interacting with test-takers (cf. Roever & Ikeda, 2022). Overall, within the video-recorded modes, raters' focus may have been on test-takers' performance, regardless of the presence of examiner.

It is important to note that the framework and rating scale of the rubric in this test focused on demonstrating competence in linguistic features – fluency, grammar, vocabulary – while nonverbal components were not articulated in the rubric. No specific guidance on nonverbal behavior regarding what to rate and how to rate may have influenced raters' judgment, such as not reflecting the impact of nonverbal interaction between the test-taker and the examiner (video1), or neglect the nonverbal behavior of the test-taker (video2). That is, such flattened construct of speaking described in the rubric affects raters' judgments on nonverbal behaviors; regardless of different speech performance between the video1 and video2 modes, the lack of description and scale of nonverbal components could have led to no statistically significant difference in scores.

Considering the characteristics of CFA, it should be acknowledged that the scores used for CFA is given by raters, not the values given by test-takers themselves. This indicates that the observed variables may not represent the "true" L2 speaking ability, and because mean scores were used, additional information (e.g., rater differences reflected in the raw scores) is limited. These characteristics of scores may have affected high inter-correlation across the latent variables (mode).

Furthermore, from a statistical point of view, the findings that the three delivery modes are measuring the same speaking construct warrants further investigation. In structural equation modeling (SEM), the latent factors are allowed to correlate when they all measure a higher construct (Anderson & Gerbing, 1988; Rubio et al., 2001). To confirm whether the different delivery modes are contributing to the same speaking construct, further research testing the assumption of higher-order CFA analysis is necessary to appropriately assume that the correlations between the latent factors indicate that the factors are measuring the same construct. Also, while there is no definite minimum sample size required for statistical precision, the sample size in this dataset 1 may not be large enough (typical sample size in SEM: $N < 200$, see Kline, 2016). Thus, while gaining large sample size may be difficult, it is highly recommended for further SEM analysis to achieve adequate statistical power and statistical precision (Kline, 2016).

Nonetheless, this CFA finding provides an overall conceptualization of how the underlying L2 speaking ability is measured across the different modes, which distinguishes from more standard statistical techniques (e.g., ANOVA, multiple regression) that analyze observed variables only (Kline, 2016). That is, the three latent variables (delivery modes) are distinct, but considering the high inter-correlations, they may be independent contexts when it comes to assessing test-takers' underlying L2 speaking ability.

Rater Behavior across Audio, Video1, and Video2 Modes

The MFRM results of the dataset 1 comparing the audio and two video-recorded (video1 mode displayed both test-taker and examiner, video2 mode displayed only test-taker) delivery modes in a speaking test suggest that while the video modes generated comparable range of test scores, audio mode had distinguishably lower scores than the video modes. In rater behavior and

the difficulty of tasks (intermediate, superior levels), some differences were observed while similar scores were found across linguistic categories (fluency, lexis, grammar). Overall, the findings confirm the statistically significant score difference between the audio and video-recorded modes reported by Nakatsuhara et al. (2020). These results further support the findings of previous studies (e.g., Conlan et al., 1994; Larson, 1984; Nakatsuhara et al., 2020; O'Loughlin, 2001; Styles, 1993) that the audio-only rating condition could limit the assessment of test-taker performance.

The bias/interaction pairwise analyses between rater and rating modes revealed mixed findings considering effect sizes. First, audio and video1 modes showed that except for the three raters (B, C, G) who showed no significant differences, the other five raters had small effect sizes. Second, for the audio and video2 modes, mixed effect sizes were discovered. While raters B, F, and G showed no significant differences between the modes, other raters had small to large effect sizes. Particularly, raters B and G showed no significant differences for all bias/interaction pairwise analyses.

The mixed findings of rater behavior across the rating modes indicate that while visual information in the video modes may affect rater severity, this wasn't the case for all raters. That is, raters may have different ways of using nonverbal cues during assessments, or they are simply not aware of how to use the nonverbal cues in their assessment ratings. I discuss further about rater behavior across the audio and video modes considering two aspects: (a) raters who showed interaction with the rating modes, and (b) raters who did not show interaction with the rating modes. While qualitative aspects of raters' different behaviors are discussed in RQ3, in this section, I elaborate on the possible reasons for the current findings in relation to previous studies that investigated the impact of visual cues on listeners' comprehension.

First and foremost, regarding the raters' interaction with the delivery modes, the tendency of rater awarding higher scores in the video modes than audio mode is in line with previous studies' findings that visual information positively affected raters' behavior (e.g., Nakatsuhara et al., 2020; cf. Conlan et al., 1994; Kenyon & Malabonga, 2001; O'Loughlin, 2001; Qian, 2009; Styles, 1993). For example, in Nakatsuhara et al.'s study, raters reported that test-takers' visual cues supported their comprehension of test-takers' intended message, and helped to understand fluency related features such as pauses, hesitation, repetition, and awkwardness. Considering that test-takers' use of gestures and their speech is known to form an integrated system of message, their use of gestures may have supported raters' comprehension (cf. Kelly et al., 2010), for example, assisting raters' memory for interpretation of speech (e.g., Beattie & Shovelton, 1999; Overoye, 2019). Thus, in the video modes, the gestured speech was likely to deliver test-takers' intended message clearly, while in the audio mode where visual cues are diminished, the speech may be occasionally incomplete for raters to fully comprehend (cf. Knight & Sweeney, 2007). Overall, such absence/presence of visual information may have influenced rater severity.

Another reason for raters' different behavior across the modes may be due to participants' gaze, an important visual feature to consider, specifically in video-conferenced platforms which mainly display participants' upper torso focused on their faces. In human communication, gaze is one of the salient nonverbal features and a part of a social skill that has been extensively found to profoundly affect a listener's perception towards a speaker (e.g., Kleinke, 1986). It is also the strongest visual cues in face-to-face interaction, and is linked to variety of functions, such as managing social interaction and speakers' intention, also grabbing listeners' attention (e.g., Frischen, Bayliss, & Tipper, 2007; Ijuin et al., 2018; Senju & Hasegawa, 2005). Furthermore, gaze is a central facet of social interaction that even

linguistically adept adults rely on to interpret social behavior (Frischen et al., 2007). Thus, within the face-focused delivery modes in this study, raters may have been affected by a test-taker's use of gaze (e.g., directly looking into the camera, looking away from the camera) because their direct or averted gaze may have sent additional information (e.g., preferences of tasks, focus during responding, linguistic difficulties during response, emotions) also reflecting their social skills.

In sum, raters assessing in the video modes were likely to consider test-takers' use of nonverbal features, which positively affected their task performance and rating behavior. Based on these findings, I echo Nakatsuhara et al.'s (2017, 2020, 2021) statement that the speaking construct measured under the video condition is much closer to the face-to-face condition (most authentic oral interaction context) than the audio-only context. In this regard, the semi-direct audio-only delivery mode that is grounded on the individualistic psycholinguistic-oriented construct can be considered as measuring less wide range of the speaking construct, which led to lower scores than video modes.

Second, regardless of the positive impact of nonverbal features on rater behavior, it is important to note that in this study, several raters' interactions were non-biased, indicating that visual information does not always enhance comprehension or support raters' understanding of test-takers' disfluency. This result is in accordance with previous studies that showed no significant differences between the rating modes with or without visual information (e.g., Beltrán, 2016; Kenyon & Tschirner, 2000; Shohamy, 1994). When rating video-recorded modes (which was the case for video1 and video 2 in this study), raters may have treated the video modes as another type of audio-recorded speech sample. For example, because raters do not have to interact with test-takers in the video-recorded (asynchronous) modes, it may be up to the raters

to decide to what extent they will consider nonverbal information in their rating. Thus, these raters were likely to look away from the screen and focus on the rubric to assess test-takers' linguistic performance, which is no different from the audio-only rating mode. Even though they may have watched the video-recordings occasionally, visual cues may have not greatly affected their assessment towards test-takers' performance, because they may not have actually seen them all.

Another potential reason may be raters' familiarity or experience with video-recordings. Because raters in this study had their rating experience mainly with the face-to-face mode or audio-only mode, video-recorded modes were new to them. Different from the raters who showed biased performance, it is likely that when rating, these raters decided to ignore information they were not familiar with, that is, additional information imposed by nonverbal cues.

In summary, the mixed findings of rater behavior across the delivery modes indicate that nonverbal cues are used differently among raters, and this may increase rater variance, affecting the final scores test-takers receive. To increase reliability of raters' performance, then, further research concerning the standardization of nonverbal features in speaking construct is required. Also, raters should be trained in terms of how to use nonverbal cues such as reflecting or not reflecting the visual information into their rating. Although such attempts are timely needed, it is important to point out that the growing number of theoretical discourse defining and measuring nonverbal features as speaking ability has been in recent few decades (cf. Burgoon & Bacue, 2008). Thus, different from linguistic features that are relatively "objective" and clearly defined as the speaking construct from the theoretical frameworks and empirical support, difficulty remains when it comes to measuring nonverbal cues as the construct (cf. Roever & Ikeda, 2022).

To complicate things further, nonverbal skills are mediated by construct-irrelevant features (e.g., culture, personality and psychology types, working memory, social setting, degree of formality, shared knowledge) which makes it difficult to distinguish between test-takers' ability of using nonverbal cues. Nonetheless, if speaking tests underrepresent the construct of communicative competence or interactional competence by not including nonverbal behavior, something now considered as a major component of the speaking construct, then the tests may produce scores that do not infer test-takers' actual ability to orally communicate in real life. Since score users tend to interpret speaking test scores as indicators of test-takers' actual speaking ability in oral communication (Roever & Ikeda, 2022), redefining the speaking construct is in great need. Thus, continued attempt to standardize the impact of nonverbal cues and train raters/examiners to appropriately use the nonverbal information is timely necessary, as video modes are increasingly used in everyday to professional contexts.

CHAPTER 19: DISCUSSION ON DATASET 2

In this dataset, I aimed to answer two research questions. Same as dataset 1, for RQ1, I carried out CFA to investigate which test delivery mode (i.e., the video-conferencing (live) mode, the video-recorded mode; *hereafter* VC and VR) best represents test-takers' speaking ability. For RQ2, I compared examiner/rater behaviors across two test delivery modes for the ITA Speaking Test. I discuss the findings of RQ2 in two categories: (a) findings at item level, and (b) findings at item type level.

Delivery Modes and the Speaking Construct

With the goal of investigating how the rating modes (VC, VR) are contributing to the underlying construct being measured (i.e., prospective ITAs' speaking ability), I conducted CFA. Two CFAs were conducted, one for items and another for item types. Since the same models (i.e., single-factor model, correlated two-factor model) were hypothesized and examined for both items and item types, I discuss the findings together in this section.

The ITA Speaking Test used in this study is developed based on an approach where each speaking test item purportedly measures test-takers' English speaking ability at a graduate level. This test is used to meet the requirements for teaching assistants, specifically their English language ability such as clarity and comprehensibility of speech (cf. Common European Framework of Reference, or CEFR, 2020[9]). This speaking test conventionally used the semi-direct mode (audio-only mode), and during pandemic, have been adopting double-rating, by using the synchronous (VC; video-conferencing mode using Zoom application) and

[9] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

asynchronous (VR; video-recorded mode) formats. The findings of CFA used in this study indicated that the two modes (VC, VR) may potentially measure the same speaking construct.

Between the two possible models (single-factor model and correlated two-factor model), the single-factor model showed poor fit and failed to support the unidimensionality. This is a similar finding to the dataset 1, which indicates that the items/item types of two different delivery modes are not considered unidimensional, because the systematic differences within the item/item type variance are caused by more than one variance source (i.e., one type of delivery mode; Rubio et al., 2001; Ziegler & Hagemann, 2015). Furthermore, this finding also provides counter-evidence of the few previous studies (e.g., Nakatsuhara et al., 2017; Zhou, 2015) which stated that MFRM and EFA findings provide supportive evidence for the unidimensionality. In other words, the current findings' failure of supporting unidimensional factor shows that the items/item types of the VC and VR modes are not measuring one attribute of the defined speaking construct. Overall, this finding provides evidence that could raise a question regarding the previous studies' doubts on the equivalence of the construct measured in different modes (e.g., Nakatsuhara et al., 2021).

In this dataset, the correlated two-factor model – VC and VR modes – was selected with excellent fit for both items and item types. Such findings indicate that the two distinct delivery modes may measure the same underlying speaking construct, and this interpretation is further supported by the high inter-correlation between the two latent factors (i.e., delivery modes). Overall, it can be understood that regardless of the context, the two modes are independent and statistically inseparable, which do not discriminate the underlying speaking construct.

I suggest a potential reason for the high inter-correlations between the two latent variables. That is, the use of the same rubric and items/item types for both delivery modes could

have led to non-discriminant final test scores. Specifically, the holistic scale focused on ITAs' linguistic skills, also in line with the language-focused rubric used in dataset 1. Such constrained description of the narrowest aspect of the speaking construct and the use of monologic tasks could have led to similar final scores (e.g., Roever & Ikeda, 2022). Even more, although holistic scale may be simple and efficient for rating, the reduced approach flattens the construct of speaking being measured, even narrower than using analytic scales.

As I have mentioned above, lack of description considering the nonverbal behavior as a component for the speaking construct could have led to the high inter-correlations. Between the VC and VR modes, because there is no specific guidance on how to assess the nonverbal behavior presented on the screen, raters may have neglected the visual information that might have distracted them either positively or negatively (Nakatsuhara et al., 2020). Or, raters might have had no particular preferences when assessing in either mode (cf. Lavolette, 2013). In essence, absence of rating scales and frameworks for nonverbal behavior in the rubric may have led to various rater behavior. That is, raters' different ways of disregarding or treating the visual information could have caused high inter-correlations among latent variables.

With regards to the characteristics of CFA, another reason may reside in the observed variables, that the scores were dependent on raters, not on the test-takers themselves. Thus, the observed variables may not accurately represent the test-takers' true speaking ability. Further, because the mean scores were used for CFA, additional information regarding variation (rater differences) were diminished.

The current finding has its own implications considering the dimensionality of the underlying construct being assessed. In particular, I would like to point out that the findings of the items/item types failing to support the unidimensionality warrants further investigation. If a

local language testing center is unaware that a measure may be multidimensional (i.e., measuring more than on attribute of the speaking construct), it can cause problems (e.g., inaccurate measures that could lead to erroneous conclusions about the measure; Rubio, Berg-Weger, & Tebb, 2001) when evaluating the speaking properties of prospective ITAs.

Complicating things further, I would like to note that while unidimensionality is an essential property for measurement, it is not sufficient evidence for construct validity (Anderson & Gerbing, 1988). Therefore, local testing centers and test designers should take into account that (a) the unidimensionality of the test score does not necessarily indicate that the items have to measure only one attribute of the speaking construct (cf. Bejar, 1983), and (b) depending on the dimensionality of items/item types, the test score could be multidimensional (Ziegler & Hagemann, 2015). Thus, if the differences within an underlying speaking construct are due to a set of different delivery modes, the test score should adequately reflect those attributes through, for example, an analytic rubric. As suggested in dataset 1, one approach could be investigating the higher-order model, with the underlying speaking construct as the highest latent construct and see if items/item types of the different modes load highly load on to this highest factor.

Rater Behavior between the VC and VR Modes

Item-Level Analysis

The purpose of the MRFM analysis with items was to understand rater behavior between the two different modes of the video-conferenced ITA Speaking Test: (a) synchronous (VC) mode and (b) asynchronous (VR) mode, which has not been investigated in previous research. The MFRM results of comparing the VC and VR delivery modes showed that raters were harsher in the VR mode than in the VC mode. Some differences were also found in rater severity and item difficulty.

The bias/interaction pairwise analyses between rater and rating modes revealed six raters' small to medium effects sizes, out of eight raters who assessed test-takers' performances in both modes. Within the six raters, only one (rater D) showed significant difference in medium effect size. Other five raters (B, C, H, J, K, M) showed small effect sizes, which implies that the difference is negligible regardless of its significance. Overall, these results are in line with previous studies in which raters were more lenient when assessing the direct mode than semi-direct mode (note, however, that the previous studies' direct and semi-direct were mainly the face-to-face mode and audio mode; e.g., Conlan et al., 1994; Larson, 1984; Nambiar & Goon, 1993; Styles, 1993).

The current findings can be understood in relation to the previous studies (e.g., Nakatsuhara et al., 2020, 2021) that used MFRM to investigate rater behavior in the VC or VR modes by comparing them with the face-to-face mode. These studies suggested that both the VR and VC modes could be a complement to the face-to-face mode (Nakatsuhara et al., 2020, 2021). In light of what Nakatsuhara et al. highlighted, I suggest that the video-conferenced speaking tests will bring the positive impact especially where double-ratings for the face-to-face mode are required. In particular situations where speaking tests cannot be held in face-to-face context (e.g., test-takers who are in different country or region; cf. Ockey, Timpe-Laughlin, Davis, & Gu, 2019), the VC mode could be the best alternative way to understand test-takers' speaking proficiency in real time. Similar to this dataset, double-ratings in the VR mode can be held if needed.

Nonetheless, the MFRM findings of this current study revealed that raters behaved differently between the VC and VR modes. In the synchronous VC mode, the raters tended to

give higher scores, while in the asynchronous VR mode, lower scores were given. Some items showed statistically significant differences in scores between the VC and VR modes.

I suggest three possible reasons for the raters' different behavior between the VC and VR modes found in this study, considering (a) different cognitive demands imposed by the VC and VR modes, and (b) the presence/absence of examiners'/raters' nonverbal engagement with test-takers in the two modes.

First of all, different cognitive demands imposed by different test delivery modes (VC, VR) may have affected rater behavior. When raters participate as examiners (test givers), they play a dual role, by assessing test-takers' live-performance using a holistic scale while simultaneously participating as a test giver. Although interactive skills are not assessed in this test type, examiners may have used nonverbal tokens (e.g., back-channeling, nodding, gazing to camera) to encourage candidates and signal comprehension (Nakatsuhara et al., 2017). Other factors such as time pressure or paying attention to the internet stability (Nakatsuhara et al., 2020) could have imposed extra cognitive demands and may have prevented raters from solely focusing on assessing test-takers' speaking proficiency.

Such cognitive demands given to the examiners in the synchronous VC mode likely required them to balance between the different features of the test, which may have distracted their attention from solely focus on rating. Raters in the VR mode, on the contrary, could have less cognitive demand with no time pressure and no need to multi-task like examiners. This is associated with the point of what previous studies have stated, that raters assessing in the VR modes are more likely to notice negative aspects of test-takers (e.g., dysfluency features, head orientation, awkwardness, nosy background), which examiners in the VC mode might have

missed (cf. Nakatsuhara et al., 2017, 2020). Raters' negative noticing could have led them to give lower scores in the VR mode than in the VC mode.

Secondly, the absence and presence of raters' engagement with test-takers across the modes may have led to different rater behavior. When assessing test-takers' performances, examiners in the VC mode generally interact with test-takers while raters in the VR mode "watch" the interaction between examiner and test-taker. Although the test in this study was one-way, in which an examiner gives monologic tasks to a prospective ITA to respond, both participants are engaged to a certain degree within interaction during task performance, potentially nonverbally. Specifically, when the mode is live (synchronous VC mode), participants use nonverbal cues to mutually communicate their messages. Although monologic tasks were used, it is likely that examiner and test-taker continuously assess, integrate, and consider what one another can see (Schober, 1993; Schober & Clark, 1989). In turn, raters' and test-takers' nonverbal behavior contributes to test-takers' and raters' nonverbal behavior, which makes it difficult to identify test-takers' original, individual contribution during the test (cf. Roever & Ikeda, 2022). Thus, in the VC mode, a social engagement between examiners and test-takers occurs to some extent, possibly using nonverbal cues during test-takers' responses, which may consequently affect raters' severity in awarding final scores.

In the VR modes, however, raters are not required to engage with test-takers, but to watch other examiner interact with test-takers/ or just the test-takers' response. Within this context, raters are not nonverbally co-constructing mutual (dis)agreements with test-takers. Rather, raters watch and assess already recorded test-takers' responses. Thus, in the VR mode, some nonverbal cues between examiner and test-taker (e.g., gaze, gesture, nodding, facial expressions) may not be decoded by the rater, which could affect, either positively or negatively, rater behavior.

Considering the lower scores in the VR mode, the absence of engagement with test-takers could have negatively influenced the way raters gave final scores to test-takers. Overall, being able to interact with the test-taker may produce different contexts for examiners from raters who do not simultaneously co-constructmeaning either verbally or nonverbally.

Overall, the nonverbal behaviors could have influenced raters' judgment on the test-takers' oral performance. If such nonverbal behaviors contributed to the interaction between the test-taker and the rater, then, it can be said that nonverbal behaviors should be considered as construct-relevant features of the speaking construct. This is an interesting finding, since nonverbal behavior was considered as construct-irrelevant within speaking assessment.

Looking back at the history of the speaking assessment, the original English proficiency test could be the Cambridge Certificate of Proficiency in English (CPE), launched in 1913, to assess the knowledge of phonetics. The two World Wars paved a way for a dramatic shift in L2 speaking assessment; in the US, Foreign Language Institute (FSI) and Oral Proficiency Interview (OPI) was introduced in 1952. These oral assessments focused on analytic components (e.g., pronunciation, fluency, grammar, vocabulary, accentedness). Subsequent development was made to oral assessments, including the Interagency Language Round table (ILR) and American Council for the Teaching of Foreign Language (ACTFL) OPI. While significant steps have been made toward the construct of speaking with the growth of speaking research in SLA – multi-faceted speaking construct – only the verbal features were considered as a component of the measured speaking proficiency.

Although few, scholars such as Kendon and McNeill consistently emphasized the crucial role of nonverbal behaviors within speech. One of the nonverbal cues frequently discussed is "coordination of action in interaction" (Kendon, Sebeok, & Umiker-Sebeok, 2016, p.21). That is,

within two-way interaction, nonverbal behavior serves as cues to another participant about when to continue or stop the talk, pass the opportunity to talk for an interactant (Kendon, 1967). Another well-known example is eye-gaze, specifically the direction pattern (e.g., Argyle & Dean, 1965; Exline, 1963; Kendon, 1967; Nielson, 1964). For example, during speech, gaze functions as an index of participants' attention, specific information regarding what is going on around the participants, and reflecting the process of planning and producing the speech (Kendon et al., 2016).

With the advance in technology (i.e., video-conferencing platforms are used in speaking tests) and visual information of participants are found to play a significant role that could affect fairness and reliability of rater behavior, the importance of nonverbal behavior as a component of the speaking construct cannot be disregarded.

In addition to the suggested potential reasons, it is worth noting that the data was collected right after the outbreak of the pandemic, when the unexpected transition occurred, from conventional semi-direct ITA Speaking Test (audio-only mode) to video-conferenced speaking tests. Most prospective ITAs and raters were new to the video-conferenced speaking test format, having less experience and familiarity than the conventionally used test modes (e.g., audio-only, face-to-face). While cognitive demands and engagement issues may be important reasons, the potential impact of rater experience and familiarity with the VC and VR modes should not be disregarded.

Lastly, besides the mode differences, the mixed findings across items (i.e., score differences between the modes for each item as indicated by the results of paired *t*-tests, different item difficulties as displayed in the variable map) indicate that item difficulty may be another reason for raters' different behavior between the modes. Depending on item difficulty, test-takers

may show different responses, or raters may show different severity. In sum, when administering the ITA Speaking Test and inferring the scores, test designers should be aware of different rater behavior between the VC and VR modes, and should not neglect the potential impact of item difficulties on test-taker performances and rater behaviors.

Item Type-Level Analysis

The goal of the MFRM analyses with the scores of item types was to investigate rater behavior between the two delivery modes (VC, VR). The findings were similar to the findings with the items; raters were harsher in the VR mode than in the VC mode. Some differences were found in item types and rater severity.

The bias/interaction pairwise analyses between rater and rating modes showed one rater's small effect size out of eight raters. Overall, the similar findings of items and item types indicate that rater behavior tend to be the same when rating the VC and VR modes. Thus, potential reasons for different rater behavior between the two modes in items may also be the case for item types.

CHAPTER 20: DISCUSSION ON RATERS' VERBAL REPORT

Role of Visual Information in Rater Behavior

To answer the third research question, of qualitatively exploring the raters' perceptions towards three different rating conditions, I analyzed raters' verbal data. The results of the raters' verbal reports clearly supported their higher scores in the video modes, that having visual information (facial expressions, gaze, head orientation) helped them to (a) understand test-takers' intended message when there was background noise, (b) understand why test-takers had disfluency features such as sudden pause or repetition, (c) engagingly listen and watch test-takers' performances, and (d) think fondly of test-takers when nonverbal cues are employed appropriately. This finding suggests that nonverbal features affect raters' comprehension and their ratings, particularly the fluency category (cf. Nakatsuhara et al., 2020), with raters giving higher scores in the video modes than in the audio mode. The raters tended to prefer the additional information given in the video modes, and this may have positively affected their scores. In the audio modes where visual information is restricted, raters were likely to give lower scores in particular when they had difficulty overcoming the background noise (e.g., music at a café, children playing or crying) or when test-takers suddenly paused while speaking.

Although the raters reported their preference for the visual information, in some cases, the audio mode was sufficient when assigning scores for analytic categories (fluency, lexis, grammar). This mostly happened when raters (a) found several visual features distracting, (e.g., test-takers' excessive use of hand gestures, frequent averted gaze, messy background), (b) were more familiar and had experience with rating audio-recordings, and (c) were confused about how to use nonverbal cues in their rating because the rubric didn't describe those. This finding reveals that raters' familiarity to the rating modes is important, because it affects how they use test-

takers' linguistic information, descriptions in the rubric, and give scores based on their prior rating experiences. The raters' new experience with the video modes may have brought confusion regarding how they should use nonverbal cues, and lack of description of nonverbal features in the rubric may have added the fuzziness of utilizing visual information to their ratings.

It appears that visual information could potentially work either positively or negatively towards the final decisions raters make, that is, the scores test-takers will receive. This finding follows the results of previous studies that showed visual information could be either positive (e.g., Jenkins & Parra, 2003; Lam, 2018; May, 2011; Nakatsuhara et al., 2020; Nambiar & Goon, 1993; Neu, 1990; Roever & Kasper, 2018) or negative (e.g., Bejar et al., 2000). It is worth noting that, regardless of negative impacts, the scores indicate that nonverbal information mostly positively affected raters' final scores.

Another important finding was the impact of nonverbal features on raters' personal preferences for test-takers. While empirical studies in social science or neuropsychology have evidenced that nonverbal cues (e.g., direct/averted gaze, facial expressions) significantly affect listeners' perception towards addressers' attractiveness or preferences (e.g., Ijuin et al., 2018; Palanica & Itier, 2012; Wieser et al., 2009), it has been less discussed in SLA and LT fields. Having a bias towards test-takers because of their nonverbal cues or visual information should be prevented for the sake of fairness issues; however, raters in this study tended to have several preferred nonverbal features that motivated them to listen more carefully. Most preferred and mentioned nonverbal cues were direct eye gaze, that is, looking straight into the camera. Although the effect of direct versus averted gaze on final scores warrants further investigation, raters were fond of test-takers who did not look away from the camera, as it showed their

willingness to engage in the task and their confidence in what they are saying. On the contrary, raters were “confused,” “fatigued,” or “felt weird” when test-takers’ background was messy, or when test-takers avoided looking into the camera. Taken together, while nonverbal cues are associated with construct-irrelevant features (e.g., culture, individual differences, L1 influence), it is hard to deny that the nonverbal information from test-takers affect rater behavior to some degree. How significantly the nonverbal information in video modes affects raters’ final scores warrants further investigation.

CHAPTER 21: CONCLUSION

In this Chapter, I conclude the current study by suggesting important topics for further research. Then, I provide implications specifically for local language testing centers and test developers who plan to continue and develop using video-conferencing applications in their L2 speaking proficiency assessments.

Room for Further Research

In this current study, I discovered that the different delivery modes are distinct factors which measure the same underlying speaking construct (i.e., L2 speaking ability). However, these modes were found to be independent, that is, they are not statistically distinct test-taking contexts and do not discriminate the measured underlying L2 speaking ability. It is worth mentioning that this finding may stem from the use of same rubric across different delivery modes in each dataset, that raters' judgment was based on the same analytic scaling (dataset 1: fluency, grammar, vocabulary, dataset 2: holistic). To understand more about the different delivery modes, I compared them, and found that (a) visual information (i.e., nonverbal behaviors such as gestures and gaze, test-takers' displayed background), (b) different range of visual information displayed on the screen (visual display of only test-taker or both test-taker and examiner), and (c) asynchronous/synchronous video modes (VC, VR) do matter when considering the final scores test-takers receive. Grounded on these current findings, I discuss avenues for future research that need further investigation to develop reliable inter- and intra-ratings, and to validate the constructs of the different delivery modes used in video-conferenced speaking tests.

Above all, the rubrics used in this current study lacked clear a description of nonverbal cues, and the same rubric was used across different delivery modes for each dataset.

Nevertheless, considering that the same rubric was used and delivery modes were the only difference when rating, the higher scores in the video modes than audio mode indicate that visual information may have affected rater behavior, possibly in a positive way. Furthermore, the mixed findings of the interaction between rater and rating mode revealed that the raters had different ways of using nonverbal cues when assessing speech samples. For example, as raters reported, nonverbal cues affected their rating behavior but they were confused on how to use these cues when they were giving scores. This is not too surprising, given that the use of nonverbal information was not part of rater training, nor was it mentioned on the rubric. To minimize the confusion in the use of nonverbal cues when rating video modes, rubrics should develop further from the psycho-linguistically oriented scales. That is, the measured construct of speaking needs to be expanded; from the assessment of test-taker's competence in verbal features to their ability to cope with social situations in which verbal and nonverbal features are used for interaction. The rapid advance of technology enables testing and education institutions to administer video-conferencing applications that will display participants' visual information. Thus, speaking ability will be defined in way that a test-taker achieves successful and efficient oral communication using both verbal and nonverbal features.

In this regard, further research should focus on developing explicit description and standardization of nonverbal cues, and how these can minimize rater variance and lead to accurate assessment of test-takers' L2 speaking proficiency. Importantly, the use of verbal information must become part of rater-training programs, otherwise, raters will be left to their own thoughts and interpretations of nonverbal behavior and will have to decide on their own whether it should, either consciously or unconsciously, factor into scores. Advice for test-takers should also be given, such as how to prepare themselves in light of the nonverbal information

that may be conveyed during the test. For example, test-takers can practice looking straight to the camera rather than looking down or sideways to give a message that their attention is to the test giver/rater, also showing confidence in their own performances. Another advice for test-takers is to have a clean background, that could minimize any disturbing impact on raters' perception towards test-takers which could be construct-irrelevant. Further research is thus required, to understand what color or type of backgrounds best work for testing contexts, and testing institutions should work on standardizing the background of test-takers and the visual display of test-takers (e.g., whether to display only the torso of a test-taker or their faces, have test-takers' head image of a certain size).

If the nonverbal information is included as the speaking construct and reflected on the rated scores, then it is essential to instruct L2 learners with critical nonverbal cues that could affect their oral performances. As mentioned above, learners can simply be informed with the direction of eye-gaze, that it shows the attention of a speaker, therefore it is better to look straight to the camera rather than looking sideways. If gestures are to be considered as the construct of speaking, then important nonverbal features such as gaze, head nodding, using particular hand gestures should be instructed. However, nonverbal behaviors are complex in its nature because of the influence of learners' own background including their culture, community, personality, age, and gender (e.g., (e.g., Itier & Batty, 2009; Palanica & Itier, 2012; Palermo & Rhodes, 2007). Thus, instruction and standardization criteria for nonverbal features need careful approach with thorough empirical evidence. Nonetheless, I would like to note that clear-cut description of nonverbal features as speaking ability is a complex issue, because the theoretical frameworks for the speaking construct (i.e., communicative competence, interactional competence) lack explicit and detailed discussion of nonverbal features in oral communication. For example, how are gaze

and gestures going to be assessed? Nonverbal features are considered as part of an interactional competence, in that nonverbal cues tend to enhance or detriment efficient communication depending on numerous construct-irrelevant features (e.g., context, L1, culture, personality). However, how to measure and assess them becomes difficult as nonverbal behavior is mediated by construct-irrelevant features, such as context and culture (Gullberg et al., 2008), and personality and psychological traits (e.g., introvert vs. extrovert: Feyereisen & de Lannoy, 1991). Thus, further research should investigate how to narrow down the critical nonverbal features that need to be described in the rubric, and thus are to be considered as part of L2 users' speaking ability. This would, most certainly, impact language teaching, and possibly could open the classroom up more fully to discussions and instruction on how nonverbal behavior can best compliment meaning making in the targeted language and culture.

Next, rater training and norming on how to use nonverbal cues is essential. In validating the speaking assessments, these are fundamental processes that allow raters to develop effective scoring standards as they gain deeper understanding of test rubrics (Beltrán, 2016). Regardless, rater bias can exist even after training and norming sessions, but it will provide guidance for the raters who are unfamiliar with the visual information in video-conferenced delivery modes.

Another further research should focus on how the different range of visual information within the video-conferenced delivery mode affect rater behavior. Depending on the choice of technology (e.g., wide-angle view, head-mounted camera, a scene-oriented camera), the quality, utility, and range of visual information exchanged among participants will differ (cf. Gergle, Kraut, & Fussell, 2004, 2013). As discussed above, the face-to-face and synchronous video-conferencing modes showed different results when compared to the video-recorded mode, largely due to the different range of visual information provided to participants. While previous

studies (e.g., Nakatsuhara et al., 2020, 2021) reported that the video-conferencing mode can be an alternative to the face-to-face mode based on the scores; I found that in this study, the measured underlying construct of speaking across different modes showed mixed findings regarding unidimensionality (i.e., the findings of CFA supported the multi-dimensionality while MFRM supported the unidimensionlaity). These different findings of unidimensionality, which are dependent on the type of statistical analysis, may indicate that the use of different test delivery modes of speaking assessment is not robust in its nature. Hence, further investigation is necessary to understand the potential impact of how different ranges of visual information displayed (i.e., face-to-face mode: full access to visual information, synchronous video-conferencing mode: partial display of participants' body image) impact raters' perception and their behaviors.

Lastly, not only raters but test-takers' perceptions towards different test delivery modes should be explored in more detail. When speaking tests are held in synchronous format, both examiners and test-takers are exchanging their visual information to one another. As raters in this study have mentioned that they had preferred visual cues and visual information from a test-taker, it is likely that test-takers will have raters they are more comfortable interacting with (e.g., test-takers' anxiety surveyed in Kim & Craig, 2012). Test-takers' emotional status or confidence in responding to tasks may be affected by not only the difficulty of tasks, but also how examiners are sending off their signals that could either have positive or negative impact on test-takers' performances.

Overall, the delivery modes in recently administered video-conferenced speaking assessments are still their early stage of utilization. To provide constructively valid and reliable scores to stakeholders, future research could investigate further the different types of video

modes and how these should be selectively used to provide an accurate measurement of test-takers' L2 speaking ability.

Implications for Local Testing Centers and Test Developers

Taken together, I suggest two broad implications of this study. These implications would be particularly useful for local testing centers and test developers who are currently using video-conferenced speaking tests, and plan to continue using it.

The first implication is that different types of video modes should be used cautiously. Depending on the type of video mode administered, test-takers could receive different test scores. While the test scores in the video modes were higher than the audio mode (dataset 1), differences in scores were also observed within the video modes (video1, video2 in dataset 1, and VC, VR in dataset 2). The highest scores were observed when test-takers participated in direct (synchronous) format than their performances assessed in video-recorded format. Within the recorded format, scores were higher when only test-takers' visual information was provided. This finding implies that besides the addition of visual information in the video modes, other things also influence rater behavior such as visual display, as well as asynchronous and synchronous visual information. Thus, local test centers may want to administer the video-conferenced speaking tests with caution, or use double-rating to minimize the score variations.

Another way to minimize the impact of visual information would be rater training and norming sessions. When raters are informed on how to use nonverbal features during the assessment, test-takers' speaking ability could be assessed more accurately. This implication again leads to the importance of explicitly describing the role of nonverbal features in the speaking construct, both theoretically and empirically. If language centers decided to ignore

nonverbal features as a measured category, then this should be explicitly mentioned in the rubric to prevent raters' mixed use of nonverbal cues.

The second implication is also in line with the use of nonverbal information. While most of the focus in applied linguistics and language testing research has been on raters' perspective and their use of nonverbal information, test-takers' awareness of the impact of their nonverbal information or cues could influence their test scores, their behavior, or their demand for instruction on nonverbal behavior. Hence, testing centers could inform test-takers in advance on how to prepare for their speaking tests in light of the nonverbal information that may be conveyed during the test. Most importantly, nonverbal features that are general and can be applied to all test-takers should be stressed in instruction and in test preparation. Because test-takers' nonverbal behaviors are closely related to their individual differences, such as their cultural background (prohibiting use of some particular gestures may not be recommended in some cultures) and personal or psychological characteristics (giving higher scores to test-takers who gesture more is not recommended), nonverbal behaviors should be carefully considered by rater training programs (note, in some cultures, direct gaze to an interlocutor is considered as rude).

While video-conferencing applications are becoming widely used in educational and speaking test contexts, it is important to consider test-takers who are visually impaired. Language testing centers and test developers should develop rubrics and tasks that will prevent visually impaired test-takers from fairness issues. Detailed research should be conducted about how visual performance testing (video-based testing) will have an impact on visually impaired test-takers and the scores they will receive.

To conclude, as education contexts are increasingly adopting video-conferenced platforms, the use of video-conferenced speaking assessments is an important issue for many testing centers. The addition of visual information tends to have positive impacts on rater behavior; however, within the video modes, there are variations that could bring different impacts on test scores. Therefore, testing centers are encouraged to learn more about how different video delivery modes affect participants, and which delivery mode best suits the purpose of a particular speaking test (e.g., ITA Speaking Test, test for assessing undergraduate international students' English speaking proficiency, test for assessing adult L2 learners' L2 speaking proficiency). When testing centers use the video modes with caution, these different formats of video modes could facilitate the assessment of L2 speaking proficiency in a more flexible and efficient way.

APPENDICES

APPENDIX A: Background Questionnaire for Test-Takers

Background questionnaire for test-takers (dataset 1)

The questionnaires are partially adapted from the studies by Gary et al (2019) and Nakatsuhara et al. (2016).

Q1. What is your name?

Q2. What is your gender?

Q3. Please indicate your age range
a. Older than 46 years old
b. 31 – 45 years old
c. 22 – 30 years old
d. Younger than 21 years old

Q4. What country are you from?

Q5. What is your first language?

Q6. What other languages do you communicate in? (Besides English and your native language)

Q7. How long have you been living in the United States?
a. 3 years or more
b. 1-2 years
c. 6 months to 1 year
d. Less than 6 months
e. I have not lived in the United States

Q8. What is your student status at your university/college?
- Graduate
- Undergraduate
- Pre-university
- Other

Q8-1. If you chose 'other', please indicate your current status.

Q9. What is your (desired) area of interest or major at your university/college?
a. Humanities (history, language, culture)
b. Business (economics, computer science, finance)
c. Social Sciences (psychology, education, linguistics)
d. Natural Sciences (math, physics, biology)
e. Other _____

Q10. How long did you study English in your home country before coming to the United States?
a. 6 years or more
b. 3-5 years

c. 1-2 years
d. Less than 1 year
e. I have not yet been able to travel to the United States for my university studies.

Q11. How long have you been studying English in the United States?
a. 6 years or more
b. 3-5 years
c. 1-2 years
d. Less than 1 year
e. I am enrolled in a U.S. university, but I have not yet been able to travel due to COVID-19 or travel restrictions.

Q12. Have you ever taken a standardized English language speaking proficiency test? (Please check all that apply)
a. TOEFL iBT Speaking
b. TOEIC Speaking
c. IELTS
d. OPIc
e. Other

Q13. How many years have you used English in an English-speaking environment (both in your own country and in the United States)?
a. 3 years or more
b. 1-2 years
c. 6 months to 1 year
d. Less than 6 months

Q14. How would you describe your English speaking ability?
a. I can easily communicate complex ideas in English
b. I can communicate complex ideas in English, but I have to work hard to do it
c. I can communicate simple ideas pretty easily, but I cannot express complex ideas
d. I have to work very hard to communicate even basic ideas in English

Q15. Approximately how many hours do you use a computer in a week?
a. 30 hours or more per week
b. 20-29 hours per week
c. 10-19 hours per week
d. 5-9 hours per week
e. 0-4 hours per week

Q16. How often do you use the computer to video chat (using Skype, Zoom, Google Hangouts, FaceTime, WeChat, WhatsApp, etc.)?
a. Every day
b. More than once a week
c. Once a week
d. Once a month
e. Never

Q17. If you use video chat apps, how many people do you usually video chat with at one time during a conversation? (check all that apply)

a. More than 3 people
b. 3 people
c. 2 people
d. 1 person

Q18. What language do you use when you video chat?

a. English
b. First language
c. Mostly English
d. Mostly first language

Q19. If you use video chat apps, what is the purpose of using it? (check all that apply)

a. For work
b. For school work
c. To video chat with friends
d. To video chat with family
e. Other

Q19-1. If you chose 'other', please write your purpose of using video chats.

Q20. Please describe your comfort level with learning new computer programs and technologies.

a. Very comfortable
b. Comfortable
c. Somewhat comfortable
d. Not at all comfortable

Q21. Which of the following best describes your interest in learning new computer programs and technologies for communicating with others?

a. Highly interested
b. Somewhat interested
c. Not at all interested

Q22. For each of the four columns, please select the option that best describes you. "When I speak English with a group of people that I don't know, I":

feel very shy	feel fairly shy	feel a little shy	don't feel shy at all
I am never a leader hate to talk	I am not usually a leader don't like to talk	I am often a leader like to talk	I am almost always a leader love to talk

Q23. On a scale from 1 to 5, please indicate what you think of your English language proficiency skills.

	1	2	3 Average	4	5 Very good

	Not good at all				
Listening					
Speaking					
Reading					
Writing					
Pronunciation					
Grammar					
Vocabulary					

Q24. Answer the question so it is true for you: "When I am using chatting apps on my phone, and I am speaking to others in my **native language**, I feel more comfortable speaking with them

a. without video (audio-only)
b. with both video and audio
c. Both audio-only and video are equally comfortable

Q25. Answer the question so it is true for you: "When I am using chatting apps on my phone, and I am speaking to others in **English**, I feel more comfortable speaking with them

a. without video (audio-only)
b. with both video and audio
c. Both audio-only and video are equally comfortable

----------Thank you ----------

APPENDIX B: Background Questionnaire for Raters

Background questionnaire for raters (dataset 1)

The questionnaires are partially adapted from the studies by Gary et al (2019) and Nakatsuhara et al. (2016).

Q1. What is your name?

Q2. What is your gender?

Q3. Please indicate your age range
a. Older than 46 years old
b. 31-45 years old
c. 22-30 years old
d. Younger than 21 years old

Q4. What country are you from?

Q5. What is your first language?

Q6. What other languages do you communicate in? (Besides English and your native language)

Q7. (For L2 English raters) How long have you been in the United States?
a. 3 years or more
b. 1-2 years
c. 6 months to 1 year
d. Less than 6 months
e. I have not lived in the United States

Q8. What is your student status at your university?
a. MA
b. PhD

Q9. What is your major at your university?

Q10. Are you a native speaker in English?
a. Yes
b. No

Q11. (For L2 English raters) How long did you study English in your own country before coming to the United States?
a. 6 years or ore
b. 3-5 years
c. 1-2 years
d. Less than 1 year

Q12. (For L2 English raters) Have you ever taken a standardized English language speaking proficiency test? (Please check all that apply)

a. TOEFL iBT Speaking
b. TOEIC Speaking
c. IETLS
d. OPIc
e. Other
f. I haven't took any English speaking tests

Q13. (For L2 English raters) How would you describe your English speaking ability?

a. I can easily communicate complex ideas in English
b. I can communicate complex ideas in English, but I have to work hard to do it
c. I can communicate simple ideas pretty easily, but I cannot express complex ideas
d. I have to work very hard to communicate even basic ideas in English

Q14. (For L1 English raters) What is your second language?

Q15. (For L1 English raters) How long did you study your second language?

a. 6 years or more
b. 3-5 years
c. 1-2 years
d. Less than 1 year

Q16. (For L1 English raters) How long did you study/live in the country that used second language?

a. 6 years or more
b. 3-5 years
c. 1-2 years
d. Less than 1 year
e. Never lived

Q17. (For L1 English rater) How would you describe your second language speaking ability?

a. I can easily communicate complex ideas in second language
b. I can communicate complex ideas in second language, but I have to work hard to do it
c. I can communicate simple ideas pretty easily, but I cannot express complex ideas
d. I have to work very hard to communicate even basic ideas in second language

Q18. (For L1 English rater) From scale of 1 to 5, please indicate how you think of your own second language proficiency.

	1 (not good at all)	2	3 (average)	4	5 (very good)
Listening					
Speaking					
Reading					
Writing					
Pronunciation					
Grammar					

Vocabulary					

Q19. How did you **teach** English as a second/foreign language?

Q20. Do you have experience as a **rater** of English speaking tests?
a. Yes
b. No

Q21. (For raters who have rating experience) How long have you been a rater of English speaking test?

Q22. Approximately how many hours do you use a computer in a week?
a. 30 hours or more per week
b. 20-29 hours per week
c. 10-19 hours per week
d. 5-9 hours per week
e. 0-4 hours per week

Q23. How often do you use the computer to video chat (using Skype, Zoom, GoogleHangouts, FaceTime, WeChat, WhatsApp, etc.)?
a. Every day
b. More than once a week
c. Once a week
d. Once a month
e. Never

Q24. If you use video chat apps, how many people do you usually video chat with at one time during conversation? (check all that apply)
a. More than 3 people
b. 3 people
c. 2 people
d. 1 person

Q25. What language do you use when you video chat?
a. Second language (English/if native speaker of English – other L2)
b. First language
c. Mostly second language
d. Mostly first language

Q26. If you use video chats, what is the purpose of using it (check all that apply)?
a. For work
b. For school work
c. To video chat with friends
d. To video chat with family

Q27. (For L2 English raters) From scale of 1 to 5, please indicate how you think of your own English language proficiency.

	1 (not good at all)	2	3 (average)	4	5 (very good)
Listening					
Speaking					
Reading					
Writing					
Pronunciation					
Grammar					
Vocabulary					

-----Thank you-----

APPENDIX C: Interview Protocol with Test-Takers

Interview protocol with test-takers (dataset 1)

1. Which speaking test was more difficult for you – the audio-only one or the one with video? Please elaborate more on your thoughts.

2. Which speaking test do you feel gave you more opportunity to speaking English – the audio-only one or the one with video? Please elaborate more on your thoughts.

3. Which speaking test did you prefer to take – the audio-only or the one with video? Please elaborate more on your thoughts.

4. If you were to take a speaking test in the future, which speaking test format would you take and why? Please answer in as much detail as you can.

APPENDIX D: Interview Protocol with Raters (Dataset 1)

Interview protocol with raters (dataset 1)

1. There were three rating modes: audio-only, video1, and video2. The video1 mode displayed both test-taker and test giver (examiner), and the video2 mode showed only test-takers' visual information. Which recording format do you think was straightforward for you to apply analytic categories? Please elaborate more on your thoughts.

2. Were test takers' nonverbal cues helpful when assigning scores? Please elaborate more on your thoughts.

3. Was listening only to test-takers' speech samples helpful when assigning scores? Please elaborate more on your thoughts.

4. Based on your rating experience, which recording mode best represents test-takers' actual English speaking proficiency? (video1, video2, audio)

APPENDIX E: Speaking Task Samples

Speaking task samples (dataset 1)

Within each format, there were four questions which differ in ACTFL Proficiency. Both formats started with a question that is Novice level, which is considered as the easiest stimulus. Following questions increase in its difficulty (Q2: Intermediate, Q3: Advanced, Q4: Superior). Regarding the purpose of this study to focus on the impact of test delivery modes, there were no follow-up questions that could potentially affect rating. Test-takers were expected to speak 1.5 min per each question. Thus, 12 minutes were expected for test-takers' completion of each format.

Format A

Q1. What do you like to do in your free time?

Q2. What is your normal routine at home?

Q3. I would like to know about the seasons in your country. How many seasons are there? How are they different? What is the weather like in each season?

Q4. Nowadays, communication through digital resources is increasingly maintained among young people through social media, texting, instant messaging, etc. In your opinion, what are the advantages and disadvantages of relying solely on digital resources for interpersonal communication? How do these changes in communication affect contemporary society?

Format B

Q1. Tell me about the place where you live.

Q2. Do you have a favorite singer? Can you introduce who she or he is?

Q3. I would like you tell me your favorite memory about a good friend. It can be a story that involves you and your friend together, or a story that you know about your friend. Talk about this memory in as much detail as you can.

Q4. Nowadays, with the advent of digital platforms, people have developed a ‘digital culture’ which refers to the way technology and the internet are shaping the way humans interact each other. This affects the way we act, think, and communicate with the society as well. In your opinion, what are the advantages and disadvantages of digital culture? How do these changes in communication affects contemporary society?

APPENDIX F: Codebook for Verbal Report

Table 44.

Codebook for verbal report (dataset 1)

Categories	Subcategories	Code Description and Data Samples	MAXQDA Codes
Q1.Straightforward mode for raters to assign scores	Audio-only mode	R12: "I think the audio-only helped me better to rate the test-takers. I sometimes found the visual aid or visual cues could be kind of distracting."	"better with less visual distraction" "Able to focus on rubric"
	Video1 mode	R1: "I felt it was a bit more real in the sense … that [test-]taker was trying to engage somehow with [the examiner]."	"test-taker may be more comfortable in video1" "more authentic, real-life based communication form" "test-taker's willingness to engage with examiner" "provides fuller picture"
	Video 2 mode	R7: "whereas in video2 mode, I knew that I could use a nonverbal cue so it helped me more."	"at a stage with gained experience to focus on nonverbal cues" "visual distraction" "rater's individual difference"
	Video modes	R5: "I felt it easier to rate when I was able to see test-takers' faces. Between video1 and video 2 modes, I think video 2 was easier for me to apply analytic categories but I do not think there is a big difference."	"had to switch between rubric and video" "no difference between video1 and video2" "felt connected to test-taker" "easier rating when seeing nonverbal cues"

Table 44. (cont'd)

Q2. Usefulness of nonverbal cues in video modes	Useful	R5: "I think it helped me when I wasn't able to understand what the test-taker was saying. The non-verbal information kind of gave me additional information about the content test-takers were trying to say."	"willingness and effort to communicate/engage" "eye gaze" "facial expressions"	
			Fluency	"distinguish mid-range bands in rubric" "helped understand sudden pause"
			Helps understanding the speech	"overcome background noise" "guess the content" "pronunciation"
	Distracting (not helpful)	R9: "there was one student who always looked away from [the examiner]. I remember it very well just because it was very distracting … and then a cat came across the screen and that I felt like wow this is really distracting"	"high proficiency test-takers" "noticing weird behavior"	
			"not looking at camera"	"bad impression"
			Distracting	"unnecessary laughing" "mask on" "background noise" "no guide on rubric"

Table 44. (cont'd)

Q3. When rating, was audio-only mode sufficient?	Sufficient	R2: "I didn't have much trouble with it because I've done some speech sample rating before and it was always just audio so that wasn't that unusual for me."	"more familiar/more experience" "able to focus more on language itself/less visual distraction" "audio mode shows test-takers' effortless speech"
	Insufficient	R1: "I didn't know exactly where to look at. I was just trying to look at the rubric while I was listening.	"didn't know where to look at (rater's eye gaze)"
Q4. Best representation of test-takers' English speaking proficiency	Audio-only	R12: "I can focus on their actual production"	"able to focus on linguistic production"
	Video1	R12: "I will say that the video 1 that has both examiner and test-taker because if you are going from a more communicative approach, I think non-verbal cues should play a big role in communication tasks like this."	"helpful to see nonverbal info" "real-life/authentic/communicative goals"
	Video2	R5: "I think it was video2 mode, where examiner's visual information is not displayed. This format makes me feel like I am the examiner who is talking to the test-taker. I could focus better because it felt like the test-taker was talking to me."	"feels like talking to test-taker" "lack of engagement"
	No difference in video modes	R7:"I think in terms of rating, I think it was probably equal on whether I could or could not see the test giver,"	"no difference in rating across video modes" "focused only on test-taker" "able to notice when test-takers, but still no difference across video modes"

Table 44. (cont'd)

Q5. Preferred test-takers when rating	Preferred test-taker	R7: "I always find it easier to rate people that are higher proficiency just because I feel like the definitions are a little bit easier to decipher in the analytic rubric on the higher levels"	"eye gaze: looking straight to the camera" "clear display" "engaging speaker" "high proficiency" "confident speaker"
		R5: "I think the screen display also affected my preference. For example, if a test-taker had a bright screen and I could clearly see their faces. In some cases, there were test-takers who looked the other way, didn't look straight to the camera, I guess because of the camera setting. In that case, I didn't feel comfortable because I had to put more energy. So I personally preferred the clear visual display of a test-taker."	
	Non preferred test-taker	R5: "I had a test-taker whom I didn't like, because it was really difficult to rate that person"	"pronunciation-wise, repetition" "difficulty of rating: low proficiency" "funny/messy background looked childish/distracting" "eye gaze: looking away/down/avoiding camera"

APPENDIX G: Test-Taker Summary Statistics

Table 45.

Test-taker summary statistics from dataset 1

Test-taker ID	Observed (raw score) average	Fair average	Difficulty measure (in logits)	Model S.E.	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
084	6.78	6.64	1.68	0.23	2.30	2.31	-0.01	0.46
078	6.67	6.34	1.05	0.58	2.15	2.16	-0.16	0.79
006	5.83	5.92	0.14	0.24	2.11	2.07	-0.08	0.36
110	4.11	4.11	-3.85	0.27	2.04	1.98	-0.25	-0.10
096	5.94	5.82	-0.09	0.24	1.95	1.95	0.11	0.36
087	7.58	7.39	3.04	0.20	1.78	1.82	-0.09	0.51
082	7.03	6.87	2.15	0.22	1.80	1.75	0.46	0.37
083	7.53	7.33	2.95	0.20	1.63	1.68	-0.19	0.38
003	5.42	5.51	-0.74	0.24	1.48	1.47	0.53	0.74
099	7.42	7.23	2.79	0.21	1.41	1.44	0.55	0.31
005	6.69	6.77	1.94	0.23	1.37	1.44	0.60	0.32
094	7.14	6.97	2.34	0.22	1.43	1.40	0.72	0.23
008	7.03	7.10	2.56	0.22	1.29	1.36	0.62	-0.29
097	6.00	5.88	0.03	0.25	1.30	1.31	0.72	0.26
101	6.42	6.43	1.24	0.24	1.24	1.28	0.72	0.49
103	5.75	5.80	-0.14	0.24	1.24	1.24	0.76	0.48
092	6.89	6.74	1.89	0.23	1.18	1.23	0.75	0.04
079	5.28	5.10	-1.58	0.34	1.22	1.22	0.75	0.62
107	5.75	5.80	-0.14	0.24	1.19	1.20	0.77	0.65
018	5.25	5.34	-1.08	0.24	1.16	1.17	0.77	0.78
001	4.92	5.00	-1.79	0.24	1.15	1.14	0.83	0.40
105	4.44	4.44	-3.01	0.26	1.07	1.14	0.86	0.14
010	5.94	6.03	0.38	0.24	1.13	1.13	0.87	0.06
088	6.06	5.93	0.15	0.25	1.07	1.07	0.93	0.53
102	7.22	7.18	2.71	0.22	1.06	1.01	1.10	0.60
090	5.17	5.04	-1.71	0.24	1.05	1.06	0.92	0.70
014	5.89	5.98	0.26	0.24	1.05	1.04	0.94	0.24
095	7.94	7.75	3.56	0.20	1.03	1.02	1.20	0.62
016	5.94	6.03	0.38	0.24	1.03	1.01	0.98	-0.98
012	4.42	4.48	-2.92	0.26	0.98	0.97	1.04	0.75
015	5.67	5.76	-0.21	0.24	0.96	0.96	1.03	0.59
104	6.86	6.84	2.08	0.22	0.97	0.95	1.10	0.73
013	5.50	5.59	-0.56	0.24	0.92	0.92	1.07	0.32
091	5.89	5.76	-0.20	0.24	0.90	0.90	1.10	0.40
089	5.11	4.98	-1.83	0.24	0.89	0.90	1.09	0.53
020	5.11	5.20	-1.37	0.24	0.89	0.89	1.17	0.69
106	5.56	5.60	-0.55	0.24	0.82	0.84	1.17	0.54
098	5.11	4.98	-1.83	0.24	0.78	0.78	1.25	0.36

Table 45. (cont'd)

004	6.69	6.77	1.94	0.23	0.77	0.79	1.20	0.27
002	5.44	5.54	-0.68	0.24	0.77	0.77	1.25	0.37
085	6.03	5.90	0.09	0.25	0.76	0.76	0.20	0.26
069	6.50	6.18	0.70	0.59	0.75	0.75	1.23	-0.93
081	6.64	6.50	1.40	0.24	0.75	0.76	1.22	0.06
007	6.53	6.60	1.61	0.24	0.74	0.75	1.23	0.40
051	5.83	6.01	0.34	0.35	0.74	0.74	1.25	0.23
056	5.28	5.47	-0.83	0.34	0.73	0.73	1.30	0.42
108	6.31	6.33	1.02	0.24	0.73	0.74	1.22	0.81
060	5.11	5.30	-1.17	0.34	0.72	0.72	1.32	0.48
080	7.50	7.25	2.82	0.29	0.73	0.71	1.49	0.62
009	6.19	6.28	0.91	0.24	0.70	0.71	1.27	0.58
100	3.89	3.80	-4.75	0.28	0.69	0.68	1.31	-0.08
086	4.61	4.49	-2.89	0.25	0.67	0.68	1.32	0.40
017	4.89	4.97	-1.85	0.24	0.66	0.67	1.34	0.63
045	6.00	6.18	0.69	0.35	0.64	0.64	1.33	0.75
011	6.81	6.88	2.16	0.23	0.64	0.65	1.37	-0.15
077	6.50	6.33	1.01	0.34	0.64	0.64	1.33	0.42
093	8.75	8.69	5.26	0.34	0.66	0.63	1.08	0.40
109	7.75	7.81	3.65	0.23	0.64	0.62	1.54	0.91
057	5.44	5.63	-0.48	0.34	0.60	0.61	1.40	0.46
050	5.61	5.80	-0.13	0.34	0.59	0.60	1.39	0.85
019	7.94	8.10	4.06	0.21	0.55	0.60	0.68	-0.02
048	5.78	5.96	0.22	0.34	0.54	0.54	1.44	0.77
044	5.78	5.96	0.22	0.34	0.52	0.52	1.46	0.37
054	5.61	5.80	-0.13	0.34	0.52	0.53	1.47	0.62
047	5.39	5.58	-0.60	0.34	0.52	0.52	1.49	0.60
065	7.67	7.21	2.75	0.49	0.47	0.47	1.41	-0.59
059	5.72	5.91	0.10	0.34	0.46	0.46	1.54	0.12
043	5.56	5.74	-0.25	0.34	0.46	0.45	1.58	0.79
046	5.78	5.96	0.22	0.34	0.43	0.43	1.55	0.59
053	5.22	5.41	-0.94	0.34	0.40	0.40	1.65	0.64
042	6.39	6.57	1.53	0.34	0.39	0.39	1.58	0.51
052	6.22	6.40	1.17	0.35	0.39	0.39	1.58	0.72
058	5.61	5.80	-0.13	0.34	0.38	0.38	1.63	0.50
064	5.50	5.17	-1.44	0.59	0.34	0.34	1.69	0.93
049	5.83	6.01	0.34	0.35	0.31	0.31	1.67	0.77
068	6.83	6.50	1.38	0.57	0.25	0.26	1.67	0.26
041	5.67	5.85	-0.02	0.34	0.24	0.24	1.76	0.81
073	7.17	6.79	1.99	0.53	0.23	0.23	1.75	0.18
055	5.78	5.96	0.22	0.34	0.20	0.20	1.78	0.73
066	7.50	7.07	2.51	0.50	0.20	0.20	1.86	0.93
076	7.50	7.07	2.51	0.50	0.20	0.20	1.86	0.93
067	8.00	7.51	3.22	0.49	0.03	0.03	1.79	0.00
071	8.00	7.51	3.22	0.49	0.03	0.03	1.79	0.00

Table 45. (cont'd)

072	8.00	7.51	3.22	0.49	0.03	0.03	1.79	0.00
063	7.00	6.65	1.69	0.55	0.03	0.03	1.96	0.00
075	5.00	4.68	-2.47	0.59	0.02	0.02	2.04	0.00
074	6.00	5.68	-0.38	0.60	0.02	0.02	1.92	0.00
062	4.00	3.77	-4.84	0.69	0.02	0.02	1.91	-0.01
061	9.00	8.90	6.39	1.81	Maximum			0.00
070	9.00	8.90	6.39	1.81	Maximum			0.00

Note. The statistics are sorted by infit mean-square values.

APPENDIX H: Test-Taker Summary Statistics

Table 46.

Test-taker summary statistics for dataset 2

Test-taker ID	Observed (raw score) average	Fair average	Difficulty measure (in logits)	Model S.E.	Infit mean square	Outfit mean square	Estimated discrimination	Point measure *r*
210	49.17	49.00	3.80	0.50	2.96	3.01	-0.17	0.08
070	40.00	39.40	-1.02	0.55	2.57	2.56	0.39	-0.31
125	49.58	49.21	3.92	0.49	2.33	2.47	0.12	0.37
114	46.67	46.74	2.68	0.43	2.37	2.36	-1.12	-0.24
165	51.25	51.38	5.32	0.49	2.22	2.21	0.15	0.50
109	50.42	50.10	4.51	0.52	2.20	2.20	0.41	0.60
143	50.00	49.58	4.16	0.49	2.07	2.11	0.32	0.44
268	47.92	47.94	3.22	0.45	1.96	2.10	0.04	0.25
177	41.25	41.71	0.57	0.51	1.99	2.05	0.36	0.51
221	41.67	42.17	0.81	0.48	1.91	1.91	0.26	0.59
038	49.58	50.05	4.48	0.51	1.84	1.89	0.55	0.07
279	40.00	39.19	-1.18	0.55	1.82	1.85	0.62	0.09
154	45.83	44.61	1.85	0.41	1.83	1.82	-1.18	-0.44
046	48.33	48.35	3.43	0.47	1.70	1.82	0.45	0.58
151	52.50	52.74	5.99	0.44	1.76	1.82	0.02	-0.07
092	51.25	50.23	4.60	0.48	1.82	1.74	0.43	0.40
161	49.58	48.84	3.70	0.52	1.80	1.80	0.59	0.23
162	47.50	48.99	3.79	0.44	1.64	1.80	0.27	0.62
062	40.00	40.22	-0.41	0.53	1.60	1.75	0.82	0.23
254	36.67	36.53	-2.49	0.41	1.38	1.75	0.22	0.89
219	59.17	59.64	10.18	0.76	1.13	1.73	0.84	-0.03
230	51.67	51.47	5.37	0.47	1.56	1.68	0.54	0.24
084	50.00	50.72	4.92	0.52	1.63	1.65	0.68	0.08
222	48.75	48.41	3.46	0.49	1.64	1.64	0.56	-0.14
248	51.67	51.54	5.40	0.47	1.64	1.63	0.45	0.00
099	40.42	39.83	-0.71	0.55	1.57	1.63	0.70	-0.24
231	53.75	53.60	6.36	0.41	1.60	1.58	-0.75	0.17
144	47.50	46.54	2.60	0.44	1.49	1.56	0.38	0.44
239	55.83	55.34	7.04	0.41	1.53	1.49	-1.81	0.45
156	51.67	51.87	5.58	0.47	1.46	1.49	0.62	0.37
214	54.17	56.42	7.48	0.42	1.41	1.47	-0.03	0.09
229	53.75	53.55	6.34	0.42	1.46	1.47	0.02	-0.02
149	52.92	53.18	6.18	0.43	1.42	1.46	0.34	-0.13
075	59.58	59.39	9.63	1.02	1.04	1.45	0.95	-0.21
153	48.33	47.02	2.80	0.47	1.42	1.45	0.63	-0.44
139	42.92	43.56	1.44	0.43	1.38	1.44	0.32	0.69
002	50.00	50.16	4.55	0.51	1.42	1.42	0.80	0.45
250	39.58	40.23	-0.41	0.54	1.42	1.38	0.80	0.23
183	49.58	49.83	4.33	0.51	1.35	1.42	0.78	-0.06
173	46.67	45.04	2.01	0.42	1.40	1.40	0.33	-0.07
167	52.50	52.66	5.96	0.44	1.32	1.40	0.58	0.71
102	49.17	48.45	3.48	0.50	1.37	1.35	0.80	0.37

Table 46 (cont'd)

098	46.67	45.60	2.23	0.42	1.35	1.34	0.32	-0.09
174	46.25	46.05	2.40	0.42	1.32	1.35	0.56	-0.30
211	47.08	45.79	2.30	0.43	1.35	1.32	0.48	-0.12
096	50.42	50.10	4.51	0.52	1.35	1.35	0.81	-0.02
127	45.42	44.72	1.89	0.42	1.31	1.34	0.58	-0.08
107	50.00	49.60	4.18	0.51	1.34	1.30	0.84	0.51
181	45.00	46.23	2.47	0.41	1.33	1.32	0.06	0.08
074	41.25	40.56	-0.16	0.51	1.23	1.33	0.83	-0.12
090	41.25	40.70	-0.06	0.49	1.32	1.19	0.77	0.45
135	47.08	47.09	2.83	0.43	1.28	1.32	0.53	0.03
228	51.67	51.87	5.58	0.47	1.32	1.29	0.74	0.61
052	44.17	44.01	1.62	0.41	1.28	1.31	0.17	0.12
251	48.75	48.99	3.79	0.48	1.29	1.29	0.82	0.60
256	51.25	51.44	0.48	1.29	1.23	1.23	0.81	0.57
087	39.17	38.96	-1.33	0.49	1.21	1.28	0.86	0.39
273	52.50	52.04	5.66	0.44	1.28	1.26	0.63	0.01
227	44.58	44.41	1.77	0.41	1.23	1.26	0.61	-0.18
158	44.17	42.55	1.00	0.41	1.19	1.22	0.65	-0.30
209	48.75	47.76	3.13	0.49	1.16	1.22	0.88	-0.14
031	54.58	53.27	6.22	0.41	1.21	1.22	-0.07	0.35
091	43.33	41.97	0.71	0.43	1.21	1.17	0.61	0.33
189	44.17	43.79	1.53	0.42	1.18	1.20	0.74	-0.06
103	55.83	55.67	7.17	0.42	1.17	1.20	0.34	-0.33
025	52.92	51.78	5.54	0.43	1.17	1.20	0.69	0.38
007	51.25	51.07	5.14	0.49	1.17	1.18	0.87	-0.16
088	42.92	42.16	0.81	0.45	1.17	1.10	0.79	0.42
110	47.92	47.80	3.15	0.45	1.13	1.16	0.87	0.32
045	46.25	46.14	2.44	0.41	1.16	1.15	0.59	0.41
249	51.67	51.46	5.36	0.47	1.14	1.15	0.85	0.12
047	57.92	58.19	8.40	0.51	1.09	1.15	0.88	-0.02
022	44.58	44.47	1.80	0.41	1.13	1.14	0.78	-0.09
101	52.50	51.59	5.43	0.44	1.14	1.06	0.81	0.40
133	42.50	42.42	0.94	0.45	1.13	1.11	0.81	0.51
064	45.00	44.46	1.79	0.40	1.13	1.13	0.42	0.53
108	58.75	58.99	8.97	0.63	1.06	1.13	0.93	0.07
258	50.00	49.60	4.18	0.51	1.12	1.12	0.91	-0.09
184	48.33	48.01	3.25	0.47	1.11	1.07	0.91	0.24
104	57.92	57.61	8.05	0.51	1.06	1.11	0.91	-0.11
146	53.33	52.53	5.90	0.44	1.08	1.11	0.85	0.51
097	43.75	42.70	1.07	0.41	1.10	1.07	0.69	0.56
215	57.50	57.58	8.03	0.48	1.00	1.10	0.96	0.20
253	56.25	55.23	7.00	0.43	1.05	1.09	0.81	0.00
116	50.00	49.69	4.24	0.52	1.08	1.09	0.96	0.10
202	50.83	52.34	5.81	0.50	0.99	1.08	1.00	0.02
282	42.92	43.16	1.27	0.44	1.04	1.08	0.97	-0.09
203	44.17	42.55	1.00	0.41	1.05	1.08	0.92	-0.02
060	43.75	43.20	1.29	0.42	1.08	1.07	0.91	-0.07
172	55.42	56.18	7.38	0.41	1.05	1.06	0.80	-0.01
170	45.83	44.72	1.89	0.41	1.05	1.06	0.92	-0.18
006	54.58	54.53	6.73	0.42	1.05	1.06	0.91	0.16

Table 46 (cont'd)

124	48.75	49.57	4.16	0.48	1.02	1.05	1.00	0.28
261	44.58	43.47	1.40	0.41	1.05	1.05	0.92	-0.06
204	50.42	50.30	4.65	0.50	1.05	1.02	1.00	0.49
009	44.58	44.47	1.80	0.41	1.03	1.04	0.97	0.08
257	51.26	51.75	5.52	0.49	1.02	1.04	1.00	0.32
259	57.08	56.75	7.62	0.46	1.04	0.97	0.96	1.15
126	50.00	49.35	4.01	0.52	1.03	1.04	0.99	0.29
030	42.92	42.81	1.112	0.45	1.03	1.04	1.00	0.15
252	46.67	47.09	2.83	0.59	1.03	1.04	0.98	-0.26
187	42.50	43.54	1.43	0.45	0.99	1.03	1.03	-0.26
106	45.83	45.64	2.24	0.41	1.02	1.03	0.99	-0.05
272	40.00	39.42	-1.01	0.53	1.00	1.02	1.01	0.34
168	44.17	44.89	1.96	0.41	1.01	1.02	1.00	-0.06
035	45.83	46.22	2.47	0.41	1.01	1.02	1.01	-0.10
071	47.08	46.03	2.40	0.43	0.99	1.02	1.03	-0.20
264	44.17	42.47	0.96	0.41	0.99	1.01	1.04	0.08
281	45.42	45.71	2.27	0.40	1.01	1.01	1.02	-0.05
166	51.25	50.46	4.75	0.49	1.01	1.00	1.01	0.40
032	45.42	44.71	1.89	0.40	0.99	0.99	1.06	-0.01
005	45.83	46.15	2.44	0.41	1.00	0.99	1.04	0.09
266	55.00	56.19	7.38	0.41	0.99	0.99	1.06	0.08
242	46.67	47.36	2.95	0.42	0.99	1.00	1.04	-0.11
067	44.17	42.88	1.15	0.41	0.99	0.99	1.06	0.00
192	44.17	43.61	1.46	0.41	0.99	0.99	1.06	0.12
086	55.00	56.19	7.38	0.41	0.99	0.99	1.09	0.10
123	50.00	50.18	4.57	0.49	1.00	0.98	0.98	0.10
284	45.42	45.71	2.27	0.40	0.98	0.98	1.09	0.03
063	53.75	53.23	6.21	0.41	0.98	0.99	1.08	0.00
233	45.00	45.74	2.28	0.40	0.98	0.98	1.10	0.07
036	44.17	45.29	2.11	0.41	0.98	0.98	1.09	0.04
122	56.67	56.68	7.59	0.45	0.98	1.00	1.04	0.31
142	55.83	54.76	6.82	0.42	0.98	0.97	1.10	0.17
179	45.00	44.46	1.79	0.41	0.98	0.97	1.09	0.16
129	46.25	45.16	2.06	0.41	0.97	0.97	1.09	0.02
240	48.75	47.99	3.25	0.49	0.97	0.98	1.03	0.41
010	42.92	41.98	0.72	0.43	0.97	0.98	1.07	-0.04
018	44.17	44.03	1.63	0.41	0.96	1.00	1.10	0.19
277	47.08	46.85	2.72	0.43	0.95	0.96	1.09	-0.05
058	44.17	44.28	1.72	0.41	0.97	0.95	1.11	0.17
217	57.08	55.90	7.27	0.45	0.95	0.96	1.10	0.27
020	49.17	48.45	3.48	0.50	0.94	0.97	1.04	0.19
137	42.92	43.55	1.43	0.43	0.94	0.95	1.10	0.03
113	47.08	47.42	2.97	0.44	0.94	1.00	1.06	0.19
132	46.67	45.60	2.23	0.42	0.94	0.94	1.12	0.07
171	57.50	57.41	7.94	0.48	0.94	0.98	1.08	0.27
130	45.42	44.29	1.73	0.40	0.94	0.94	1.19	0.19
275	43.33	42.68	1.06	0.42	0.94	0.94	1.13	0.09
017	35.42	34.88	-3.05	0.38	0.94	0.94	1.06	-0.04
081	42.08	43.05	1.22	0.46	0.93	1.00	1.06	-0.27
004	56.25	56.34	7.45	0.44	0.93	0.95	1.15	0.35

Table 46 (cont'd)

285	44.17	42.72	1.08	0.41	0.94	0.93	1.18	0.17
044	45.00	43.65	1.48	0.40	0.93	0.93	1.23	0.21
188	47.08	46.85	2.72	0.43	0.93	0.95	1.12	0.01
265	45.83	45.75	2.29	0.42	0.93	0.92	1.15	0.28
111	43.33	43.17	1.28	0.43	0.93	0.92	1.13	0.26
186	59.17	59.03	9.12	0.74	0.99	0.92	1.02	0.16
245	54.17	54.02	6.53	0.41	0.93	0.92	1.26	0.23
262	55.42	54.93	6.88	0.42	0.92	0.94	1.27	0.32
100	52.92	53.18	6.18	0.43	0.96	0.92	1.10	0.09
078	45.83	45.61	2.23	0.41	0.92	0.92	1.21	0.22
283	45.00	43.47	1.40	0.40	0.91	0.91	1.29	0.28
057	47.08	47.82	3.16	0.43	0.93	0.91	1.12	0.17
224	42.92	42.24	0.85	0.44	0.91	0.90	1.14	0.20
267	52.92	51.42	5.34	0.43	0.90	0.91	1.16	0.15
255	45.42	46.63	2.63	0.40	0.90	0.90	1.28	0.28
033	42.08	42.34	0.90	0.46	0.90	0.93	1.10	-0.13
053	47.50	47.09	2.83	0.44	0.90	0.90	1.13	0.03
039	42.92	43.56	1.44	0.43	0.90	0.91	1.16	0.18
077	52.92	52.66	5.96	0.43	0.91	0.90	1.16	0.11
178	49.17	48.85	3.71	0.50	0.91	0.90	1.06	0.31
095	45.83	45.31	2.12	0.41	0.90	0.89	1.27	0.28
134	49.17	48.45	3.48	0.50	0.89	0.89	1.08	0.39
117	42.50	42.38	0.92	0.46	0.93	0.89	1.10	0.22
079	45.83	45.61	2.23	0.41	0.89	0.89	1.27	0.30
115	42.08	42.62	1.03	0.46	0.89	0.90	1.12	-0.05
026	46.25	44.95	1.98	0.41	0.91	0.89	1.20	0.26
270	53.33	52.32	5.80	0.42	0.89	0.89	1.24	0.24
066	43.33	44.00	1.61	0.42	0.90	0.88	1.20	0.24
246	47.08	45.83	2.32	0.43	0.90	0.88	1.17	0.12
011	43.75	43.52	1.42	0.41	0.88	0.88	1.26	0.32
131	45.00	45.74	2.28	0.40	0.87	0.87	1.37	0.38
236	56.67	56.59	7.55	0.44	0.89	0.87	1.31	0.45
160	42.50	41.20	0.27	0.45	0.90	0.87	1.14	0.15
105	50.42	50.44	4.74	0.49	0.91	0.87	1.09	0.53
119	46.25	45.75	2.28	0.41	0.88	0.87	1.28	0.31
094	55.42	53.93	6.49	0.42	0.86	0.87	1.45	0.45
195	58.33	58.22	8.42	0.55	0.92	0.86	1.09	0.36
232	41.25	41.41	0.40	0.50	0.98	0.86	1.02	0.49
278	42.92	44.02	1.62	0.43	0.88	0.86	1.20	0.24
226	55.00	55.73	7.20	0.41	0.86	0.86	1.63	0.49
185	48.33	48.01	3.25	0.47	0.86	0.92	1.10	-0.24
205	43.33	42.99	1.20	0.42	0.87	0.85	1.24	0.32
072	43.33	42.31	0.88	0.42	0.86	0.84	1.26	0.35
198	41.67	40.57	-0.16	0.48	0.84	0.88	1.12	-0.06
019	47.08	46.09	2.42	0.43	0.88	0.84	1.19	0.21
241	57.92	57.68	8.09	0.51	0.94	0.84	1.10	0.34
056	43.75	43.20	1.29	0.42	0.86	0.83	1.29	0.39
213	47.50	46.01	2.38	0.44	0.85	0.83	1.19	0.18
001	52.08	51.92	5.61	0.46	0.86	0.83	1.16	0.11
191	43.33	44.45	1.79	0.42	0.87	0.83	1.24	0.34

Table 46 (cont'd)

136	46.67	47.40	2.97	0.42	0.85	0.83	1.27	0.33
073	43.75	42.68	1.06	0.41	0.85	0.83	1.34	0.44
263	47.50	48.48	3.50	0.44	0.84	0.83	1.20	0.17
076	37.92	37.79	-1.96	0.45	0.83	0.84	1.14	0.04
199	42.50	42.94	1.18	0.46	0.87	0.83	1.15	0.32
244	47.92	48.50	3.51	0.45	0.82	0.83	1.18	0.04
138	46.09	45.03	2.01	0.42	0.83	0.82	1.38	0.46
014	45.83	44.73	1.90	0.41	0.83	0.82	1.39	0.44
147	52.92	51.95	5.62	0.43	0.84	0.82	1.26	0.31
216	46.25	47.76	3.13	0.43	0.83	0.81	1.24	0.46
175	41.67	41.68	0.44	0.48	0.81	0.82	1.15	-0.02
150	48.75	47.53	3.02	0.48	0.81	0.92	1.10	-0.50
260	47.50	46.92	2.75	0.44	0.82	0.81	1.23	0.25
037	47.92	48.24	3.37	0.45	0.84	0.80	1.17	0.19
225	46.25	45.62	2.23	0.42	0.82	0.80	1.33	0.43
034	42.08	41.94	0.70	0.46	0.82	0.80	1.19	0.18
120	42.50	41.86	0.65	0.45	0.85	0.79	1.20	0.28
069	38.33	38.91	-1.36	0.47	0.79	0.82	1.14	-0.05
121	44.58	44.67	1.87	0.43	0.83	0.79	1.29	0.50
083	46.25	47.40	2.96	0.41	0.80	0.79	1.42	0.55
016	45.00	45.74	2.28	0.41	0.81	0.79	1.40	0.50
269	47.50	47.28	2.91	0.44	0.82	0.79	1.23	0.26
082	37.92	38.83	-1.41	0.45	0.79	0.79	1.17	0.13
218	42.08	40.88	0.06	0.46	0.80	0.77	1.20	0.24
200	46.25	47.76	3.13	0.43	0.83	0.77	1.25	0.47
093	43.75	42.32	0.89	0.42	0.78	0.77	1.40	0.52
276	43.33	43.83	1.55	0.45	0.84	0.77	1.20	0.48
041	51.67	51.92	5.60	0.47	0.77	0.78	1.19	-0.04
089	54.17	53.47	6.31	0.43	0.79	0.76	1.37	0.54
068	42.50	41.92	0.69	0.45	0.79	0.76	1.26	0.44
280	43.33	43.83	1.55	0.45	0.80	0.76	1.23	0.51
208	47.50	48.17	3.33	0.44	0.77	0.75	1.28	0.38
049	42.08	42.26	0.86	0.46	0.78	0.75	1.23	0.34
054	44.58	43.10	1.25	0.41	0.76	0.75	1.59	0.64
148	59.17	59.69	10.34	0.75	0.96	0.74	1.05	0.27
027	47.92	48.77	3.66	0.45	0.77	0.74	1.23	0.25
061	51.67	52.19	5.74	0.47	0.74	0.73	1.22	0.08
238	51.67	52.56	5.91	0.47	0.75	0.73	1.22	0.11
024	48.33	47.74	3.12	0.47	0.74	0.73	1.22	0.14
235	55.00	55.39	7.06	0.41	0.73	0.73	1.90	0.69
163	50.83	51.70	5.49	0.48	0.74	0.72	1.20	0.38
182	48.33	48.61	3.57	0.47	0.75	0.72	1.22	0.10
048	48.33	49.12	3.87	0.47	0.72	0.71	1.23	0.14
207	45.00	46.49	2.58	0.43	0.73	0.70	1.40	0.61
223	52.50	52.71	5.98	0.44	0.75	0.70	1.33	0.46
023	47.08	46.09	2.42	0.43	0.73	0.69	1.39	0.58
029	44.17	42.86	1.14	0.41	0.69	0.68	1.64	0.73
196	42.50	42.26	0.86	0.45	0.75	0.68	1.30	0.46
247	50.42	50.82	4.98	0.50	0.67	0.67	1.20	0.32
140	58.33	58.37	8.52	0.55	0.84	0.66	1.19	0.68

Table 46 (cont'd)

085	38.75	39.50	-0.96	0.50	0.66	0.67	1.20	0.05
051	52.08	52.35	5.82	0.45	0.71	0.66	1.31	0.50
206	52.50	54.58	6.74	0.44	0.73	0.65	1.34	0.46
050	37.92	38.18	-1.77	0.45	0.71	0.64	1.24	0.47
237	51.67	52.19	5.74	0.47	0.67	0.64	1.28	0.34
243	52.92	53.47	6.31	0.44	0.70	0.63	1.44	0.58
180	41.67	42.55	1.00	0.48	0.68	0.62	1.26	0.47
055	48.33	48.25	3.37	0.47	0.68	0.61	1.28	0.29
043	38.75	38.79	-1.43	0.49	0.64	0.59	1.23	0.17
028	41.67	41.84	0.64	0.48	0.64	0.58	1.29	0.45
145	48.33	48.63	3.58	0.46	0.65	0.56	1.32	0.34
271	48.75	48.58	3.55	0.49	0.58	0.55	1.29	0.26
059	49.17	49.38	4.04	0.50	0.55	0.58	1.26	-0.21
220	48.33	48.91	3.75	0.47	0.61	0.55	1.32	0.50
159	50.00	49.28	3.97	0.52	0.53	0.53	1.24	0.14
155	48.33	49.64	4.21	0.47	0.59	0.52	1.35	0.43
201	50.87	49.99	4.44	0.51	0.52	0.53	1.28	-0.09
197	39.17	39.85	-0.70	0.52	0.52	0.52	1.25	0.00
015	51.25	51.12	5.17	0.48	0.55	0.51	1.33	0.32
040	48.33	48.63	3.58	0.46	0.60	0.50	1.36	0.41
003	51.25	51.12	5.17	0.48	0.55	0.50	1.33	0.32
021	50.00	50.43	4.73	0.50	0.50	0.49	1.28	0.20
274	47.92	48.90	3.74	0.45	0.61	0.49	1.41	0.50
118	49.17	49.63	4.20	0.50	0.49	0.49	1.29	-0.05
141	47.92	48.24	3.37	0.45	0.58	0.48	1.41	0.55
080	50.83	51.63	5.45	0.51	0.47	0.47	1.31	0.08
194	51.25	50.94	5.06	0.48	0.53	0.45	1.36	0.36
190	49.17	49.00	3.80	0.50	0.43	0.41	1.33	0.27
157	40.83	40.87	0.05	0.52	0.45	0.40	1.32	0.32
065	40.42	40.80	0.01	0.55	0.33	0.35	1.31	-0.23
164	49.17	50.61	4.85	0.49	0.38	0.33	1.39	0.36
176	49.58	49.45	4.08	0.50	0.33	0.35	1.39	-0.07
008	40.42	40.77	-0.01	0.53	0.31	0.32	0.05	-0.18
234	50.42	50.28	4.63	0.52	0.30	0.32	-0.18	0.48
042	50.83	50.97	5.07	0.50	0.36	0.30	0.48	0.20
013	50.42	50.82	4.98	0.50	0.25	0.23	0.20	0.28
112	50.42	50.85	5.00	0.52	0.24	0.22	0.28	0.31
152	49.58	51.01	5.10	0.50	0.22	0.19	0.31	0.00
169	50.00	50.43	4.73	0.50	0.06	0.05	0.00	0.00
193	50.00	50.09	4.50	0.52	0.03	0.03	0.00	0.00
128	50.00	50.42	4.79	0.52	0.02	0.02	0.00	0.00
212	50.00	49.60	4.17	0.52	0.01	0.01	0.00	0.00
012	60.00	59.82	(10.88	1.84)	Maximum		0.00	0.00

Note. The statistics are sorted by infit mean-square values.

REFERENCES

REFERENCES

Allen, D. E., & Guy, R. F. (1977). Ocular breaks and verbal output. *Sociometry*, *40*, 96–99. https://doi.org/10.2307/3033550

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing* (2nd ed.). American Educational Research Association.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411–423. https://doi.org/10.1037/0033-2909.103.3.411

Bachman, L. F. (1990). *Fundamental considerations in language testing* (Oxford). Oxford University Press.

Bachman, L. F. (1991). What does language testing have to offer? *TESOL Quarterly*, *25*(4), 671–704. https://doi.org/10.2307/3587082

Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.

Bachman, L., & Palmer, A. (2010). *Language assessment in practice*. Oxford University Press.

Batty, A. O. (2014). A comparison of video-and audio-mediated listening tests with many-facet Rasch modeling and differential distractor functioning. *Language Testing*, *32*(1), 3–20. https://doi.org/10.1177/0265532214531254

Beattie, G. (1978). Floor apportionment in conversational dyads. *British Journal of Clinical and Social Psychology*, *17*, 7–16. https://doi.org/10.1111/j.2044-8260.1978.tb00889.x

Beattie, G., & Shovelton, H. (1999). Mapping the range of information contained in the iconic hand gestures that accompany spontaneous speech. *Journal of Language and Social Psychology*, *18*(4), 438–462. https://doi.org/10.1177/0261927X99018004005

Bejar, I., Douglas, D., Jamiesone, J., Nissan, S., & Turner, J. (2000). *TOEFL 2000 listening framework: A working paper* (TOEFL Monograph Series No. 19). Educational Testing Service.

Bejar, I. I. (1983). *Achievement testing: Recent advances*. Sage.

Beltrán, J. (2016). The effects of visual input on scoring a speaking achievement test. *Columbia University Working Papers in TESOL & Applied Linguistics*, *16*(2), 1–23. https://doi.org/10.7916/D8795GKM

Berry, V., Nakatsuhara, F., Inoue, C., & Galaczi, E. (2018). *Exploring the use of video-conferencing technology to deliver the IELTS Speaking Test: Phase 3 technical trial* [IELTS Partnership]. IELTS Partners: British Council, Cambridge Assessment English and IDP.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model* (3rd ed.). Routledge.

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). The Guilford Press.

Burch, A. R., & Kasper, G. (2016). Like Godzilla. In M. Prior & G. Kasper (Eds.), *Emotion in Multilingual Interaction* (pp. 57–85). John Benjamins.

Burgoon, J. K., & Bacue, A. E. (2008). Nonverbal communication skills. In J. O. Greene & B. R. Burleson (Eds.), *Handbook of communication and social interaction skills* (pp. 179–220). Lawrence Erlbaum Associates.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series No. 20). ETS.

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*. Routledge.

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333–342). Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Cassell, J., McNeill, D., & McCullough, K. E. (1999). Speech-gesture mismatches: Evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics & Cognition*, *7*, 1–33. https://doi.org/10.1075/pc.7.1.03cas

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Choi, S., & Lantolf, J. P. (2008). The representation and embodiment of meaning in L2 communication. Motion events in the speech and gesture of advanced L2 Korean and L2 English speakers. *Studies in Second Language Acquisition*, *30*(2), 191–224. https://doi:10.1017/S0272263108080315

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Clark, J. L. D., & Hooshmand, D. (1992). “Screen-to-screen” testing: An exploratory study of oral proficiency interviewing using video-conferencing. *System*, *20*(3), 293–304. https://doi.org/10.1016/0346-251X(92)90041-Z

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, R. L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*, 113–139. https://doi.org/10.1080/09541449208406246

Conlan, C. J., Bardsley, W. N., & Martinson, S. H. (1994). *Study of intra-rater reliability of assessments of live versus audio-recorded interviews in the IELTS Speaking component* [Unpublished study]. International Editing Committee of IELTS.

Conway, C. A., Jones, B. C., DeBruine, L. M., & Little, A. C. (2008). Evidence for adaptive design in human gaze preference. *Proceedings of the Royal Society of London Series B: Biological Sciences*, *275*, 63–69. https://doi.org/10.1098/rspb.2007.1073

Craig, D. A., & Kim, J. (2010). Anxiety and performance in videoconferenced and face-to-face oral interviews. *Multimedia-Assisted Language Learning*, *13*(3), 9–32.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE Publications.

Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. (2003). Finite sampling properties of the point estimates and confidence intervals of the RMSEA. *Sociological Methods and Research*, *32*, 208–252. https://doi.org/10.1177/0049124103256130

Dahl, T. I., & Ludvigsen, S. (2014). How I see what you’re saying: The role of gestures in native and foreign language listening comprehension. *The Modern Language Journal*, *98*, 813–833. https://doi.org/10.1111/modl.12124

Davis, L., Timpe-Laughlin, V., Gu, L., & Ockey, G. J. (2018). Face-to-face speaking assessment in the digital age: Interactive speaking tasks online. In J. M. Davis, J. M. Norris, M. E. Malone, T. H. McKay, & Y.-A. Son (Eds.), *Useful assessment and evaluation in language education* (pp. 115–130). Georgetown University Press.

Dorans, N. J., & Lawrence, I. M. (1999). *The role of the unit of analysis in dimensionality assessment* (ETS Research Report RR-99-14). ETS.

Ducasse, A., & Brown, A. (2009). Assessing paired orals: Raters’ orientation to interaction. *Language Testing*, *26*(3), 423–443. https://doi.org/10.1177/0265532209104669

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Engle, R. A. (1998). Not channels but composite signals: Speech, gesture, diagrams, and object demonstrations are integrated in multimodal explanations. In M. A. Gernsbacher & S. J. Derry (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 278–298). Springer-Verlag.

Feldman, R. S., Philippot, P., & Custrini, R. J. (1991). Social competence and nonverbal behavior. In R. S. Feldman & B. Rimé (Eds.), *Fundamentals of nonverbal behavior* (pp. 329–350). Cambridge University Press.

Feyereisen, P., & de Lannoy, J.-D. (1991). *Gestures and speech: Psychological investigations*. Cambridge University Press.

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social condition, and individual differences. *Psychological Bulletin, 133,* 694–724. https://doi.org/10.1037/0033-2909.133.4.694

Fulcher, G. (2003). *Testing second language speaking*. Longman/Pearson Education.

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly, 5*(2), 89–119. https://doi.org/10.1080/15434300801934702

Galaczi, E. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*(5), 553–574. https://doi.org/10.1093/applin/amt017

Galaczi, E., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

Gary, O., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). Exploring the potential of a video-mediated interactive speaking assessment. (ETS Research Report RR-19-05). ETS.

Gergle, D., Kraut, R. E., & Fussell, S. (2013). Using visual information for grounding and awareness in collaborative tasks. *Human-Computer Interaction, 28,* 1–39. https://doi.org/10.1080/07370024.2012.678246

Gergle, D., Kraut, R., & Fussell, S. (2004). Language efficiency and visual technology minimizing collaborative effort with visual information. *Journal of Language and Social Psychology, 23*(4), 491–517. https://doi.org/10.1177/0261927X04269589

Goffman, E. (1963). *Behavior in public places*. The Free Press of Glencoe.

Goldin-Meadow, S. (2003). *Hearing gesture: How our hands help us think*. The Belknap Press.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*, 516–522. https://doi.org/10.1111/1467-9280.00395

Green, B. A., & Lung, Y. S. M. (2021). English language placement testing at BYU-Hawaii in the time of COVID-19. *Language Assessment Quarterly*, *18*(1), 6–11. https://doi.org/10.1080/15434303.2020.1863966

Groen, M., Ursu, M., Michalakopoulos, S., Falelakis, M., & Gasparis, E. (2012). Improving video-mediated communication with orchestration. *Computers in Human Behavior*, *28*(5), 1575–1579. https://doi.org/10.1016/j.chb.2012.03.019

Gullberg, M. (1998). *Gesture as a communication strategy in second language discourse: A study of learners of French and Swedish*. Lund University Press.

Gullberg, M. (2006). Handling discourse: Gestures, reference tracking, and communication strategies in early L2. *Language Learning*, *56*(1), 155–196. https://doi.org/10.1111/j.0023-8333.2006.00344.x

Gullberg, M., de Bot, K., & Volterra, V. (2008). Gestures and some key issues in the study of language development. *Gesture*, *8*, 149–179. https://doi.org/10.1075/gest.8.2.03gul

Hair, J., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson Educational International.

He, A., & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. E. Young & A. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 1–24). John Benjamins.

Huang, B. H., Bailey, A. L., Sass, D. A., & Chang, Y. S. (2020). An investigation of the validity of a speaking assessment for adolescent English language learners. *Language Testing*, 1–28. https://doi.org/10.1177/0265532220925731

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin.

Ijuin, K., Umata, I., Kato, T., & Yamamoto, S. (2018). Difference in eye gaze for floor apportionment in native- and second-language conversations. *Journal of Nonverbal Behavior*, *42*, 113–128. https://doi.org/10.1007/s10919-017-0262-3

Isbell, D. R., & Kremmel, B. (2020). Test review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *Advance online publication*, 1–20. https://doi.org/10.1177/0265532220943483

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience and Biobehavioural Reviews*, *33*, 843–863. https://doi.org/10.1016/j.neubiorev.2009.02.004

Iwashita, N., May, L., & Moore, P. J. (2021). Operationalising interactional competence in computer-mediated speaking tests. In R. Salaberry & A. R. Burch (Eds.), *Assessing speaking in context: Expanding the construct and its applications* (pp. 283–302). Multilingual Matters.

Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. *The Modern Language Journal*, *87*(1), 90–107. https://doi.org/10.1111/1540-4781.00180

Kampe, K. K., Frith, C. D., Dolan, R. J., & Frith, U. (2001). Reward value of attractiveness and gaze. *Nature*, *413*, 589. https://doi.org/10.1038/35098149

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). American Council on Education and Praeger.

Kelly, S. D., Ö zyürek, A., & Maris, E. (2010). Two sides of the same coin. *Psychological Science*, *21*, 260–267. https://doi.org/10.1177/0956797609357327

Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychologica*, *26*, 22–63. https://doi.org/10.1016/0001-6918(67)90005-4

Kendon, A. (1981). Introduction: Current issues in the study of 'nonverbal communication. In T. A. Sebeok & J. Umiker-Sebeok (Eds.), *Nonverbal Communication, Interaction, and Gesture: Selections from Semiotica* (pp. 1–53). Mouton Publishers.

Kendon, A. (2000). Language and gesture: Unity or duality? In D. McNeill (Ed.), *Language and gesture* (pp. 47–63). Cambridge University Press.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology*, *5*(2), 60–83.

Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal*, *84*(1), 85–101. https://doi.org/10.1111/0026-7902.00054

Kim, J., & Craig, D. A. (2012). Validation of a videoconferenced speaking test. *Computer Assisted Language Learning*, *25*(3), 257–275. https://doi.org/10.1080/09588221.2011.649482

Kim, M., & Crossley, S. A. (2020). Exploring the construct validity of the ECCE: Latent structure of a CEFR-based high-intermediate level English language proficiency test. *Language Assessment Quarterly*, *17*(4), 434–457. https://doi.org/10.1080/15434303.2020.1775234

Kita, S., Alibali, M. W., & Chu, M. (2017). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*, *124*(3), 245–266. https://doi.org/10.1037/rev0000059

Kleinke, C. L. (1986). Gaze and eye contact: A research review. *Psychological Bulletin*, *100*, 78–100. https://doi.org/10.1037/0033-2909.100.1.78

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press.

Knight, L. V., & Sweeney, K. (2007). Revealing implicit understanding through enthymemes: A rhetorical method for the analysis of talk. *Medical Education*, *41*, 226–233. https://doi.org/10.1111/j.1365-2929.2006.02681.x

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal*, *70*(4), 366–372. https://doi.org/10.1111/j.1540-4781.1986.tb05291.x

Lam, D. M. K. (2018). What counts as "responding"? Contingency on previous speaker contribution as a feature of interactional competence. *Language Testing*, *35*(3), 377–401. https://doi.org/10.1177/0265532218758126

Lam, D. M. K. (2021). Don't turn a deaf ear: A case for assessing interactive listening. *Applied Linguistics*, *42*(4), 740–764. https://doi.org/10.1093/applin/amaa064

Larson, J. W. (1984). Testing speaking ability in the classroom: The semi-direct alternative. *Foreign Language Annals*, *17*(5), 499–507. https://doi.org/10.1111/j.1944-9720.1984.tb01738.x

Lavolette, E. B. (2013). Effects of technology modes on ratings of learner recordings. *The IALLT Journal*, *43*(2), 1–27. https://doi.org/10.17161/iallt.v43i2.8524

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. *Language Testing*, *13*(2), 149–170. https://doi.org/10.1177/026553229601300202

Linacre, J. M. (2012). *Many-Facet Rasch Measurement: Facets Tutorial*. https://www.winsteps.com/a/ftutorial2.pdf

Linacre, J. M. (2019). *A user's guide to FACETS*. https://www.winsteps.com/a/Facets-Manual.pdf

Linacre, J. M. (2021). *A user's guide to FACETS*. https://www.winteps.com/a/Winsteps-Manual.pdf

Lippa, R. (1975). Expressive control and the leakage of dispositional introversion-extroversion during role-played teaching. *Journal of Personality*, *44*, 541–559. https://doi.org/10.1111/j.1467-6494.1976.tb00137.x

Little, T. D., Cunningham, W. A., & Shahar, G. (2002). To parcel or not to parcel: Exploring the question, weighting the merits. *Structural Equation Modeling*, *9*(2), 151–173. https://doi.org/10.1207/S15328007SEM0902_1

Liu, T., Aryadoust, V., & Foo, S. (2022). Examining the factor structure and its replicability across multiple listening test forms: Validity evidence for the Michigan English Test. *Language Testing*, *39*(1), 142–171. https://doi.org/10.1177/02655322211018139

Mason, M. F., Tatkow, E. P., & Macrae, C. N. (2005). The look of love: Gaze shifts and person perception. *Psychological Science*, *16*, 236–239. https://doi.org/10.1111/j.0956-7976.2005.00809.x

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly*, *8*(2), 127–145. https://doi.org/10.1080/15434303.2011.565845

McCafferty, S. G. (2002). Gesture and creating zones of proximal development for second language learning. *Modern Language Journal*, *86*(2), 192–203. https://doi.org/10.1111/1540-4781.00144

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. https://doi.org/10.1038/264746a0

McNamara, T. (1996). *Measuring second language performance*. Longman.

McNamara, T. (1997). Interaction in second language performance assessment: Whose performance? *Applied Linguistics*, *18*(4), 446–466. https://doi.org/10.1093/applin/18.4.446

McNeill, D. (1992). *Hand and mind. What gestures reveal about thought*. University of Chicago Press.

McNeill, D. (1998). Speech and gesture integration. In J. M. Iverson & S. Goldin-Meadow (Eds.), *The nature and functions of gesture in children's communication* (pp. 11–27). Jossey-Bass.

McNeill, D. (2005). *Gesture and thought*. University of Chicago Press.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods*, *9*(3), 369–403. https://doi.org/10.1177/1094428105283384

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). American Council on Education and Macmillan.

Montero Perez, M. (2020). Multimodal input in SLA research. *Studies in Second Language Acquisition*, *42*, 653–663. https://doi.org/10.1017/S0272263120000145

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2016). *Exploring performance across two delivery modes for the same L2 speaking test: Face-to-face and video-conferencing delivery* (IELTS Partnership No. 1). Cambridge English Language Assessment.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (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. https://doi.org/10.1080/15434303.2016.1263637

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

Nakatsuhara, F., Inoue, C., & Taylor, L. (2020). Comparing rating modes: Analysing live, audio, and video ratings of IELTS speaking test performances. *Language Assessment Quarterly*. https://doi.org/10.1080/15434303.2020.1799222

Nakatsukasa, K. (2016). Efficacy of recasts and gestures on the acquisition of locative prepositions. *Studies in Second Language Acquisition*, *38*, 771–799. https://doi.org/10.1017/S0272263115000467

Nambiar, M. K., & Goon, C. (1993). Assessment of oral skills: A comparison of scores obtained through audio recordings to those obtained through face-to-face evaluation. *RELC Journal*, *24*(1), 15–31. https://doi.org/10.1177/003368829302400102

Neu, J. (1990). Assessing the role of nonverbal communication in the acquisition of communicative competence in L2. In R. C. Scarcella, E. S. Andersen, & S. D. Krashen (Eds.), *Developing communicative competence in a second language* (pp. 121–138). Newbury House.

Ockey, G. J. (2014). The potential of the L2 group oral to elicit discourse with a mutual contingency pattern and afford equal speaking rights in an ESP context. *English for Specific Purposes*, *35*, 17–29. https://doi.org/10.1016/j.esp.2013.11.003

Ockey, G. J. (2021). An overview of COVID-19's impact on English language university admissions and placement tests. *Language Assessment Quarterly*, *18*(1), 1–5. https://doi.org/10.1080/15434303.2020.1866576

Ockey, G. J., Gu, Li., & Keehner, M. (2017). Web-based virtual environments for facilitating assessment of L2 oral communication ability. *Language Assessment Quarterly*, *14*(4), 346–359. https://doi.org/10.1080/15434303.2017.1400036

Ockey, G. J., Timpe-Laughlin, V., Davis, L., & Gu, L. (2019). *Exploring the potential of a video-mediated interactive speaking assessment* (Research Report ETS RR-19-05; pp. 1–27). Educational Testing Service. https://doi.org/10.1002/ets2.12240

O'Loughlin, K. J. (2001). *The equivalence of direct and semi-direct speaking tests*. UCLES/Cambridge University Press.

O'Sullivan, B., Weir, C. J., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing*, *19*, 33–56. https://doi.org/10.1191/0265532202lt219oa

Overoye, A. L. (2019). *When hands make memories: The retrieval and representation of gesture and speech* [Unpublished doctoral dissertation]. University of California, Santa Cruz.

Palanica, A., & Itier, R. J. (2012). Attention capture by direct gaze is robust to context and task demands. *Journal of Nonverbal Behavior*, *36*, 123–134. https://doi.org/10.1007/s10919-011-0128-z

Palermo, R., & Rhodes, G. (2007). Are you always on my mind? A review of how face perception and attention interact. *Neuropsychologia*, *45*, 75–92. https://doi.org/10.1016/j.neuropsychologia.2006.04.025

Pennycook, A. (1985). Actions speak louder than words: Paralanguage, communication, and education. *TESOL Quarterly*, *19*(2), 259–282. https://doi.org/10.2307/3586829

Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. *Language Testing*, *35*(3), 427–445. https://doi.org/10.1177/0265532218772325

Purpura, J. E., Davoodifard, M., & Voss, E. (2021). Conversion to remote proctoring of the community English language program online placement exame at Teachers College, Columbia University. *Language Assessment Quarterly*, *18*(1), 42–50. https://doi.org/10.1080/15434303.2020.1867145

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Affective effects on test takers. *Language Assessment Quarterly*, *6*(2), 113–125. https://doi.org/10.1080/15434300902800059

R Core Team. (2021). *R: A language and environment for statistical computing*. R Fondation for Statistical Computing. https://www.r-project.org/

Riggio, R. E. (1992). Social interaction skills and nonverbal behavior. In R. D. Feldman (Ed.), *Applications of nonverbal behavioral theories and research* (pp. 3–30). Psychology Press.

Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can(not) tell us about interactional competence. *Language Testing*, *39*(1), 1–23. https://doi.org/10.1177/02655322211003332

Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing*, *35*(3), 331–355. https://doi.org/10.1177/0265532218758128

Ross, S. (2018). Listener response as a facet of interactional competence. *Language Testing*, *35*(3), 357–375. https://doi.org/10.1177/0265532218758125

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1036. https://www.jstatsoft.org/article/view/v048i02

Rubio, D. M., Berg-Weger, M., & Tebb, S. S. (2001). Using structural equation modeling to test for multidimensionality. *Structural Equation Modeling*, *8*(4), 613–626. https://doi.org/10.1207/S15328007SEM0804_06

Salaberry, R., & Burch, R. (Eds.). (2021). *Assessing speaking in context: Expanding the construct and its applications*. Multilingual Matters.

Saussure, F. de. (1959). *Course in general linguistics*. Philosophical Library.

Sawaki, Y., & Sinharay, S. (2017). Do the TOEFL IBT® section scores provide value-added information to stakeholders? *Language Testing*, *35*, 529–556. https://doi.org/10.1177/0265532217716731

Schober, M. F. (1993). Spatial perspective-taking conversation. *Cognition*, *47*, 1–24. https://doi.org/10.1016/0010-0277(93)90060-9

Schober, M. F., & Clark, H. H. (1989). Understanding by addresses and overhearers. *Cognitive Psychology*, *21*, 211–232. https://doi.org/10.1016/0010-0285(89)90008-X

Sellen, A. J. (1995). Remote conversations: The effects of mediating talk with technology. *Human-Computer Interaction*, *10*(4), 401–444. https://doi.org/10.1207/s15327051hci1004_2

Senju, A., & Hasegawa, T. (2005). Direct eye gaze captures visuospatial attention. *Visual Cognition*, *12*, 127–144. https://doi.org/10.1080/13506280444000157

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing*, *11*(2), 99–123. https://doi.org/10.1177/026553229401100202

Slama-Cazacu, T. (1976). Nonverbal components in message sequence: "Mixed syntax." In C. McCormack & S. A. Wurm (Eds.), *Language and man: Anthropoglocial issues* (pp. 217–227). Mouton.

Song, J., & Hsu, W.-L. (2021). Desgin and implementation of a classroom-based virtual reality assessment. In *Assessing speaking in context: Expanding the construct and its applications* (pp. 265–282). Multilingual Matters.

Spolsky, B. (1985). What does it mean to know how to use a language? *Language Testing*, *2*(2), 180–191. https://doi.org/10.1177/026553228500200206

Styles, P. (1993). *Inter- and intra-rater reliability of assessments of "live" versus audio- and video-recorded interviews in the IELTS Speaking test*. British Council centre.

Sueyoshi, A., & Hardison, D. (2005). The role of gestures and facial cues in second language listening comprehension. *Language Learning*, *55*(4), 661–669. https://doi.org/10.1111/j.0023-8333.2005.00320.x

Taylor, L., & Falvey, P. (Eds.). (2007). *IELTS collected papers: Research in speaking and writing assessment. Studies in language testing 19*. Cambridge: UCLES/Cambridge University Press.

Tsunemoto, A., Lindberg, R., Trofimovich, P., & McDonough, K. (2021). Visual cues and rater perceptions of second language comprehensibility, accentedness, and fluency. *Studies in Second Language Acquisition*, 1–26. https://doi.org/10.1017/S0272263121000425

van Leeuwen, T. (2004). Ten reasons why linguists should pay attention to visual communication. In P. LeVine & R. Scollon (Eds.), *Discourse and technology: Multimodal discourse analysis* (pp. 7–19). Georgetown University Press.

van Lier, L. (1989). Reading, writing, drawling, stretching and fainting in coils: Oral proficiency interviews as conversations. *TESOL Quarterly*, *23*(3), 480–508. https://doi.org/10.2307/3586922

Vo, S. T. (2019). *Effects of task types on interactional competence in oral communication assessment* [Doctoral Dissertation]. Iowa State University.

von Grünau, M., & Anston, C. (1995). The detection of gaze direction: A stare-in-the-crowd effect. *Perception*, *24*, 1297–1313. https://doi.org/10.1068/p241297

Vuilleumier, P. (2002). Perceived gaze direction in faces and spatial attention: A study in patients with parietal damage and unilateral neglect. *Neuropsychologia*, *40*, 1013–1026. https://doi.org/10.1016/S0028-3932(01)00153-1

Wagner, S., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language*, *50*(4), 395–407. https://doi.org/10.1016/j.jml.2004.01.002

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. MacMillan Palgrave.

Weir, C. J., Vidakovic, I., & Galaczi, E. (2013). Measured constructs. In *Studies in Language Testing* (Vol. 37). Cambridge University Press.

Wieser, M. J., Pauli, P., Alpers, G. W., & Mühlberger, A. (2009). Is eye to eye contact really threatening and avoided in social anxiety? - An eye-tracking and psychophysiology study. *Journal of Anxiety Disorders*, *23*, 93–103. https://doi.org/10.1016/j.janxdis.2008.04.004

Wright, B. D., & Linacre, J. M. (1994). *Reasonable mean-square fit values*. Rasch Measurement Transactions. https://www.rasch.org/rmt/rmt83b.htm

Yan, X., Cheng, L., & Ginther, A. (2019). Factor analysis for fairness: Examining the impact of task type and examinee L1 background on scores of an ITA speaking test. *Language Testing*, *36*(2), 207–234. https://doi.org/10.1177/0265532218775764

Young, R. (2008). *Language and interaction*. Routledge.

Young, R. (2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426–443). Routledge.

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

Ziegler, M., & Hagemann, D. (2015). Testing the unidimensionlaity of items: Pitfalls and loopholes. *European Journal of Psychological Assessment*, *31*(4), 231–237. https://doi.org/10.1027/1015-5759/a000309