문학박사 학위논문

Developing and Validating a Mobile Augmented Reality (MAR)-Mediated English Speaking Assessment for Korean EFL High School Learners

한국 고등학교 영어 학습자를 위한
모바일 증강현실 기반 말하기 평가
개발과 타당화 연구

2023년 8월

서울대학교 대학원
영어영문학과 어학 전공

변 정 희

Developing and Validating a Mobile Augmented Reality (MAR)-Mediated English Speaking Assessment for Korean EFL High School Learners

지도교수 이 용 원

이 논문을 문학박사 학위논문으로 제출함

2023년 4월

서울대학교 대학원
영어영문학과 어학 전공

변 정 희

변정희의 문학박사 학위논문을 인준함

2023년 7월

위 원 장 _____
부위원장 _____
위 원 _____
위 원 _____
위 원 _____

Abstract

Developing and Validating a Mobile Augmented Reality (MAR)-Mediated English Speaking Assessment for Korean EFL High School Learners

Byun, Jung Hee
Department of English language and literature
Seoul National University

This dissertation explores the feasibility of using mobile-based, context-aware augmented reality (MAR) technology as a new mode of second language (L2) speaking assessment. It describes in detail the efforts made to develop, and validate an MAR-based English speaking test for high school students in the domestic context.

Accordingly, a mobile AR-mediated English speaking test (hereafter, MARST) was developed using "Eco English Test," a mobile application. The test was administered to 200 Korean high school students, (110 males, 90 females) aged between 16 to 17 years. The test comprised four semi-direct speaking tasks on topics related to global environment.

MARST validation was conducted using both the conceptual frameworks of Assessment Use Argument and Interpretation/Use Argument framework, focusing especially on determining how the innovative testing mode provided by MAR technology can be

characterized and evaluated as part of the overall validation process of the whole test.

Four research questions were posed for this study: (1) To what extent are the MAR-mediated speaking test scores and the test's underlying structure comparable to those of several other measures of the same and other traits? (2) To what extent do the assessment settings (e.g., rater, task, and rating categories) affect test scores? (3) What are test-users' perceptions toward the use of MARST, and do they differ according to individual characteristics, such as gender and general English proficiency? (4) What are the linguistic features of MAR-mediated communication, and how do they inform MAR-mediated test validation?

For data analysis, test scores from several measures of speaking skills and other traits were collected, along with questionnaire and interview responses. These were analyzed using a mixed method approach that included psychometric approaches, such as the Classical Test Theory (CTT), the Multi-Trait Multi-Method (MTMM) and the Many-Facet Rasch Model (MFRM), as well as corpus and discourse analyses of test-takers' speaking responses.

The MTMM analysis not only showed positive correlations between the MARST score and other speaking measures, but also revealed the unidimensional internal factor structure of the MARST. The results are empirical evidence supporting the validation argument that MARST test scores contribute to a common construct of the target speaking ability

and interpretations and can be interpreted as a good indicator of test-takers' speaking ability.

The MFRM analysis offers empirical evidence to support the following validation arguments: (a) The observed scores of the MARST were reliable estimates of expected scores; (b) The separate analytic rating scales contributed to the target construct; (c) There was no task redundancy, nor was there a need for revision or deletion; (d) Test-takers performed significantly differently across various aspects of speaking; and (e) Interpretations of the test construct were consistent across different groups of test-takers.

The bias (interaction) analysis indicated that the rating behaviors of raters did not vary by gender, test-takers' region of residence, or the rating criteria. Regarding the mode effect, no significant differences were found across test-takers' gender, general English proficiency level, or region of residence. However, a statistically significant interaction was found between the scoring behaviors of two raters and both Task 1 (dialogue completion) and Task 3 (explaining the sequence of events). Yet, following the guideline suggested in the literature and drawing upon subsequent interviews, it was concluded that this interaction did not appear to have a substantial impact on the measurement of test-takers' ability to perform.

The results of the test-taker questionnaires revealed that the MAR-based testing was comfortable and engaging. Respondents generally agreed that the test input, presented through the MAR mode, was

authentic and provided clear instructions and guidance for crafting their responses. They highly rated the items inquiring about the suitability of the test tasks for presentation in the MAR mode and their relevance to classroom learning. Test users saw the MARST as a valuable alternative for L2 speaking assessment in English as a Foreign Language (EFL) contexts.

In the subsequent corpus and discourse analyses of 150 sampled responses to Task 3, which required test-takers to describe the procedure of an event to a simulated interlocutor, the immersive effects of the MAR mode on the linguistic features became apparent. These included an increased perception of the interlocutor's presence, heightened awareness of the speaker's role identity, and a sense of urgency given the task situation. MAR technological features appeared to encourage interactions with a simulated interlocutor, revealing the interactive linguistic features of test-takers' responses in tasks typically limited to a monologue. These factors suggested that the MARST did not underrepresent the intended speaking construct in Task 3.

Subsequently, three key issues were addressed in the validation arguments: (a) the contextualization of integrating MAR technology in second language assessment; (b) the investigation of the mode effects on the assessment construct, tasks, and test-takers; and (c) the investigation of control of variabilities in test conditions.

Several technological and pedagogical implications for MARST can be drawn from this study. Rating behaviors and strategies involved in the

speaking process should be further investigated in the MARST contexts. Language testers and technology experts should continue to work together to design and develop more authentic language learning and testing contexts for language learners.

Keyword : test mode, technology-mediated language assessment, Many-Facet Rasch Measurement(MFRM), Multitrait Multimethod (MTMM), mobile augmented reality

Student Number : 2013-30014

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1 Statement of the problem

In our modern digital society, there is increasing demand for exploring novel, innovative and alternative modes of language assessment that properly correspond with the rapidly growing new manner of communication in our everyday life, sometimes characterized as “untact” and “multi-modal”. Advances in technology, further driven by the extended COVID19 pandemic, have unprecedentedly changed the way we communicate. Such advances have established new norms of interaction not only for non-face-to-face communication that do not involve direct physical contact (e.g., online videoconferencing), but also communication in virtual reality (VR). Nowadays, we also find it quite natural to communicate in multi-modal contexts in which online text messages are combined with animated images or video clips.

If language assessment practices and systems are to evolve to meet learners’ diverse communicative needs and help them thrive in future societies, it is worthwhile to explore emerging technologies that have been drastically changing the ways we communicate and interact with other people, particularly in terms of their potential impact on, and implications for, second language assessment. Given that assessing communicative (second) language ability requires establishing authenticity by way of representing how actual use of language in

communication occurs in language users' real lives, it is a timely venture to conduct research that illuminates potentially viable new test platforms in which language assessment can be undertaken, and reflect contemporary and/or newly emerging communication modes that meet different language users' needs in various contexts.

As demonstrated in Mislay, Almond, and Lucas (2003), testing experts utilize a variety of platforms for test delivery. Paper-and-pencil tests and oral exams have a long history of use. Although computer-based tests were introduced later, these have already become a dominant format of assessment. Moreover, new ways to deliver tests continued to appear as well: over the web (Ockey, Gu & Keehner , 2017), via hand-held devices, like the mobile phone.

The instructional and learning benefits of mobile phones have been extensively studied in the field of general education. It is surprising, however, that among a number of test format candidates, there has yet been little research on the use of mobile devices in (second) language assessment so far. In fact, mobile phones have been used so extensively by people nowadays had such an immense influence on human beings that Choi (2019) coined the term "phono-sapience" (human beings making essential use of mobile phone as if it were part of their bodies. In addition to mobile devices' functions as teaching and learning tools, they are also widely recognized as a platform to provide simulated experiences, which are highly similar to, or completely different from, the real world, via VR and/or augmented reality(AR).

Advances in technology have made it possible to capture more complex performances in assessment settings by including, for example, simulation, interactivity, collaboration, and constructed response that we envisage as the future of assessment (Mislevy et al., 2003). These complex assessment data can serve as evidence, laying the foundation for the inferences a test developer wishes to make, with validity being defined as the basis for the inferences drawn from the assessment data. In this sense, it is worthwhile to investigate the effects of the test mode on the target construct, test scores, test tasks, and test-takers' perceptions. This is because, during the test validation process, assessment validators need to take into account the fact that a change in test format may influence score interpretation and a series of decisions based on it, which will also make it necessary to reexamine a number of related issues, ranging from the scope of the target construct components to be measured and the types of expected performance, to the kinds of tasks that MAR can best accommodate. It is possible that interpretations of test scores and inferences about test-taker's language proficiency in a mobile-based test may differ from those delivered in different formats such as paper-based or interview tests.

With these as a backdrop, the current study attempts to make validation endeavors for a new technology-based speaking assessment, and more particularly, seeks to address to what extent innovations or capabilities of the mobile AR test format can offer the appropriate means for informing the score-based interpretations about test-takers'

speaking proficiencies and fulfilling the desired test purpose and outcomes. In this sense, the current study attempts to contribute to encouraging relevant and meaningful theoretical and methodological discussions on the integration of technology in language assessment.

1.2 Context of the study

Since the 1960s when modern language testing began, practitioners have endeavored to make the testing process more efficient and innovative via various language assessment technologies. In the field of L2 assessment, a variety of multimedia and information and communication technologies (ICT) have enhanced the efficiency and effectiveness of the existing language assessment system, as evidenced by video-conferenced speaking tests, computer or web-based testing (CBT), and automated essay scoring."

Two of the most promising, and closely related ICTs for language learning and teaching are VR and AR. VR recreates or simulates a real-life environment or situation on a computer using computer-generated graphics, images, and sounds. In contrast, AR overlays computer-generated realities (or VR) onto an existing reality. In other words, VR provides a digital recreation of a reality, while AR embeds digital objects into real environment.

One important advantage of both AR and VR is their capacity to provide immersive learning environments in which learners, by

interacting with the virtual environment, can experience feelings and emotions similar to (or the same as) those that they might experience in the real world through interacting with the virtual environment (Liu, 2009). An immersive learning environment is effective if it cognitively, emotionally, and even physically engages learners using a combination of AR techniques (Whiteside, 2002). Since interaction and communication are key elements in language learning and acquisition (Nunan, 1989; Johnson & Johnson, 1994; Ellis, 2003), AR technology applications may hold promise for teaching and assessing productive language skills. In particular, situation-based language learning and assessment enable learners to develop a more immersive perception and multiple perspectives toward spatial objects and shapes and to increase interactions in both physical and interpersonal dimensions (Blagg, 2009).

As AR/VR-technology can simulate authentic features of real-life communication tasks in testing situations, AR-integrated language assessment provides a viable means to enhance interactions with test tasks and interlocutors, which may address some key factors of concerns in L2 assessment development and validation, such as authenticity. The adoption of speaking assessment can be facilitated in EFL instructional settings where teachers are relatively reluctant to implement speaking assessments owing to practical issues such as the lack of authentic tasks, longer testing time associated with the face-to-face mode, and compromised reliability caused by on-site scoring.

In the literature on language assessment, few studies have focused

on the affordances of AR digital access to traditional materials, online simulation, and context-sensitive reference support (Mayer, 2006) as alternative means of L2 assessment. Numerous researchers in the domestic context (Kim, 2010; Kim, 2016; Lee, 2017; Lee, 2010; Lee, 2018; Pang, 2008) express concerns that the research area is restricted to developing instructional contents and educational programs for primary school students or adult learners, primarily focusing on English language components such as vocabulary and writing skills. The scope of AR- or VR-based language education research should be expanded to the pedagogical evaluation of these programs and the validation of their systems.

It wasn't long ago that a validation study of MAR-based speaking performance assessment was conducted for EFL secondary school learners in the Korean domestic context. Byun(2020) created an AR-based art guide of Korean traditional genre paintings via the AR authoring tool, 'HP Reveal' (Hewlett Packard, 2018), reporting that AR integration was useful to create learner-centered assessment that enhanced student performance during the test.

Similarly, in the international context, mobile-based assessment studies have focused on formative assessments with elementary students and in STEM subjects (Nikou and Economides, 2018). Centering on the effect of the mobile device or virtual environment upon affective aspects of students such as less stressful atmosphere and learning motivation in the literature of language learning and testing

(Chen, Gu, & Wong, 2017; Ockey, et al., 2017; Redondo, Cózar-Gutiérrez, González-Calero, & Ruiz, 2020; Wang, Song, Xia, & Yan, 2009).

However, little research exists on how technological affordances are useful in improving or upgrading the quality of test-takers' language performance, which might in turn lead to more efficient development of communicative competence. For example, York(2019) found that the virtual modality had a positive effect on oral performance, particularly in terms of fluency and accuracy. Further, increasing task complexity appeared to benefit language learners, with virtual environments offering greater advantages when dealing with more complex tasks.

For this reason, the aims of the current study include investigating the specific task types best served by the new MAR test mode. As mobile AR systems are often associated with informal learning, diverse constraints of classrooms, such as time, space, discipline, or curriculum should be explored as well (Cuendet, Bonnard, Do-Lenh & Dillenbourg, 2013).

Language assessment in general serves several significant purposes. Some of these primary purposes include screening/selection, diagnosis/feedback, placement, program evaluation achievement (Henning, 1987). Language tests are also classified as aptitude, proficiency and achievement tests. The speaking test developed for this dissertation research was initially developed as a means of achievement assessment evaluating the spoken communicative abilities of students in a low-stakes classroom setting. To be more specific, its primary purpose

was to evaluate the extent to which students have achieved the targeted skills and knowledge required for spoken communication on environment-related issues and topics, as prescribed by the first-grade high school learning standards of the national curriculum of the English subject.

Before the achievement assessment, the students had learned and practiced useful expressions, background knowledge and grammatical structures in classes, upon which performing various communicative tasks would be based.

Regarding some major validity issues in L2 assessment, whether a change in the testing format will influence item types, test tasks, and quality of the oral test tasks and the quality of the oral performance to be elicited from those items/tasks is a grave concern. How technology intersects with the construct definition can not only impact test development but also the interpretation of test scores and the justification of test use for specific purposes.

Considering the pure construct perspective, for example, test-takers' performance on a language test in a CBT format may be affected by their language ability but also by their ability to navigate the elements on the computer screen, which is not relevant to their language ability. There is an opposing view that criticizes such a perspective as being too limited for the various test purposes and the communicative contexts of interest to test users. For example, when introducing new delivery modes for L2 speaking tests such as computer, online web, and

videoconferencing, it is common that validation studies (Kim & Craig, 2012; Zhou, 2015; Nakatsuhara, Berry, Inoue, & Galaczi, 2017; Ockey et al., 2017; Berry, Nakatsuhara, Inoue, & Galaczi, 2018) have commonly viewed mode effects as a facet of task conditions. They discussed how psychometric properties of the new test were comparable with those of face-to-face speaking tests, and thus showing that the two modes as equivalent and minimizing the mode effect could be minimized to a statistically negligible level.

The research agenda for the current study is clear: MAR-delivered speaking assessment needs to be contextualized in a test validation framework. This framework should provide two key components. First, it should offer empirical validation evidence highlighting psychometric qualities that uphold the quality of a language assessment. Second, it should present new considerations for test-takers' test performance, task type, and score interpretation. These considerations are particularly important for understanding test-takers' speaking abilities, an area of interest for both test users and language testers. Evidential support will be collected from this validation process to justify claims regarding score interpretations and inferences about test-takers' language ability in technology-mediated assessment.

1.3 Objectives of the study and research questions

The current study explores the viability of mobile and context-

aware AR(MAR) technology as a new delivery mode of language assessment to more authentically and interactively engage learners in the testing context. In this study, attention is paid to the validation process, which investigates the comparability of the MAR-delivered test scores and its underlying factor structure to other measures of the same ability. In addition, we examine the test mode effect (or interaction) on test-taker ability, individual features (i.e., gender, general proficiency level), task type, and the target construct to be measured.

The current research seeks to answer the following questions.

Can the MAR - mediated speaking test serve as an appropriate test platform for assessing language learners' oral proficiencies?

1. To what extent are the test scores and the test's underlying factor structure comparable to those of other measures of the same speaking traits (i.e., oral translation and interviews) and of other traits (i.e., listening, reading, and writing)?
2. To what extent do the assessment settings (e.g., test-taker, rater, task, and rating category) affect test scores?
3. What are the test-users perceptions of the MARST and do they differ according to individual characteristics such as gender and general English proficiency?
4. What are the linguistic features of MAR-mediated communication and

how do they inform MAR-mediated test validation?

To answer these questions, a MAR speaking test platform was developed in the form of a mobile application, named 'Eco English Test', which embedded AR for task-based performance assessment in a Korean EFL high school classroom setting.

In the next sections, the relationship between learning content and assessment features were explored, along with how test tasks aligned with the English Language Achievement Standards, stipulated in the 2015 national curriculum that was in effect during the test's development. This is because investigating the relevance of the newly developed MAR-mediated speaking assessment to its purpose and usefulness is an important part of the test's validation framework.

Quantitative analysis of the MTMM and factor analysis were conducted to investigate score comparability among different measures. MFRM analysis (Linacre, 2006) was also conducted to investigate the effect of multiple aspects, including task, rater, and rating category on the test scores used for inferring test-takers' oral proficiency. This might raise some empirical issues, such as whether a change in the testing format would influence the quality of the oral performance to be elicited, whether the selection of tasks adequately reflected test-taker's abilities, and whether the performance ratings were reliable.

The research incorporated both quantitative and qualitative approaches. First, it focused on the test-taking process by analyzing

sampled spoken responses produced during the test administration. Post-test questionnaires and interviews were also examined. The study employed a mixed method approach. This integrated both quantitative and qualitative analyses. This approach was instrumental in providing insights into several aspects.

Finally, this study sheds light on the effect of applying MAR technology to language testing. The study examined its effect on the qualities of language abilities being measured and the efficiency of test administration. It also provided information on the psychometric qualities of test scores.

1.4 Organization of the dissertation

This dissertation is composed of seven chapters. Chapter 1, the Introduction, outlines the statement of the problem, the context of the study, and concludes with the objectives and research questions. Chapter 2 provides a review of L2 speaking assessments, critiques existing models, introduces augmented reality's potential in testing, and presents a validation framework for the new Mobile Augmented Reality Speaking Test (MARST).

Chapter 3 delineates the research methods used for the current study, including the participants of the study, assessments/tests developed and used in the study, scoring and rating procedure, and methods of both quantitative and qualitative data analyses. Chapter 4

reports the results of data analyses.

Building on the findings from Chapter 4, Chapter 5 synthesizes the findings from Chapter 4 and articulates the validity arguments for the interpretation and use of scores from the MARST. Chapter 6 initiates the discussion phase. It not only summarizes the findings that answer research questions 1 to 4, but also addresses significant validity issues surrounding the application of the MARST. Lastly, Chapter 7 concludes the dissertation by discussing its implications and limitations, setting a path for future research in this field.

Chapter 2. Literature review

In this chapter, a comprehensive literature review is presented. It commences with the history of L2 (second language) speaking assessment and culminates with a description of the recent shift towards language assessment via virtual technologies. This shift highlights a critique of various models of L2 speaking test performance, pointing out their failure to consider a component of increasing importance – the test mode, followed by the presentation of potential affordances of augmented reality technology that may suggest an alternative speaking test performance model. It then details the task characteristic framework and test method characteristics, laying the groundwork for the validation framework of the newly-developed Mobile Augmented Reality Speaking Test (MARST) presented in this dissertation.

2.1 Historical overview of speaking assessment

It is clear that the development of L2 speaking assessment has attempted to follow technological advances in our society. In the next section, existing work on how professional knowledge of L2 speaking assessment has evolved is briefly discussed, with a focus on test modes and validation issues followed by the emerging research on applying mobile AR technology into language education and assessment.

2.1.1 From interviewer-led to group interview

Since the second World War, new technology (e.g., short-wave radio) has been used to evaluate the military personnel's foreign language skills, which allows them to perform in wartime situations. Accordingly, the Foreign Service Institute's oral proficiency interview (OPI) comprises tasks that require test-takers to describe pictures or speak on a certain topic.

As Leaper (2014) mentioned about the history of testing speaking, the Educational Testing Service (ETS) and American Council on the Teaching of Foreign Languages (ACTFL) in mid 1980s established proficiency guidelines, on which the OPI, the most widespread speaking assessment method across the US, was based. Later, the language proficiency interview (LPI), which was developed by the Foreign Service Institute (FSI) of the US Department of the State, and similar to ACTFL's OPI, was described as the "a face-to-face interview test with one or two trained interviewers".

Language researchers, however, had pointed out the shortcomings of interviewer-led tests, particularly heavy dependence on interviewers 'elicitation skills (van Lier, 1989) and the fact that they could dominate the topic of discussion (Johnson, 2001). As a solution, interaction between two or more participants was suggested, which are called the paired format and the group oral discussion format respectively.

Unlike paired tests, the positive aspects of group oral discussion

tests (GOTs) include its efficiency in testing individuals on a large scale (Bonk & Ockey, 2003; Hidsdon, 1991), capability to elicit a wide range of more natural or conversational language (Fulcher, 1996; Gan, 2010; Gan, Davison & Hamp-Lyons, 2009) and positive washback on teaching and learning (Fulcher, 1996; Van Moere, 2006). In addition, it does not make test-takers feel as nervous (Fulcher, 1996; He & Dai, 2006).

On the other hand, test developers need to hand over some control to the test to test-takers, who, in turn, must take responsibility for demonstrating their own and their peers' language skills. Moreover, potential threats to the validity of the test such as the influence of personality traits (i.e., shyness and assertiveness) (Bonk & Van Moere, 2004; Ockey, 2009, 2011), extraverts on interaction (Nakatsuhara, 2011), and difficulties in scoring reliability are not easy to overlook (Van Moere, 2006). To address these restrictions of direct speaking assessment, more attention should be paid to indirect speaking assessment via digital technology.

2.1.2 Assessment via multimedia

Research on the role of test mode in speaking assessment has been conducted with the advent of new technologies that seek to replace face-to-face speaking tests, which have served the needs of interactional communications over a century. Significantly, various limitations arising from the high resource requirements of in-person speaking tests

conducted by trained examiners, in a live context, have prompted research into alternative test formats. These alternative modes involve semi-direct speaking tasks designed to elicit speech from test-takers.

2.1.2.1 Computer-based testing (CBT)

The advantages of CBT include increased opportunities for learning by monitoring students' work, reducing the amount of time students spend on each test item, and providing prompt feedback to them (Dunkel, 1999). Meanwhile, researchers have also begun to understand the important issues in CBT design and development, namely the evaluation of CBT for the intended types of inferences and purposes (Norris, 2001) and technical and conceptual issues pertinent to assessing the construct of L2 reading proficiency (Chalhoub-Deville, 1999).

From the early 2000s, comparability studies of two modes commenced to validate relatively novel tests with various formats (e.g., computer vs. web-based, video-conferenced speaking tests vs. conventional face-to-face oral tests), thus highlighting the effects of the new test modes of web and videos. Kenyon and Tschirner (2000) and Shohamy (2004) compared simulated and actual test scores of OPI (Oral Proficiency Interview) speaking tests and proved that there was no difference in mean scores between the simulated and actual interview speaking tests. However, comparisons between the two tests were subject to some limitations as they varied in task type and content.

O' Loughlin (2001) made direct comparisons between tape-based and live versions of the Australian Assessment of Communicative English Skills (AACES). Some noteworthy findings here are that a single dimension of speaking ability could not be constructed from the combined data. To be more specific, the data gathered from the tape version showed more lexical density than those of the live version. This finding suggests that the live version measured interactive ability, whereas the tape version tapped into test-takers' monologic ability.

Zhou's (2015) validation study examined the psychometric qualities and underlying factor structures of computer-delivered L2 monologic speaking tasks in comparison to face-to-face tasks. There was no significant difference between the overall test scores awarded by the two modes, and the following exploratory factor analysis revealed a single factor with a similar pattern in the two modes. This study is different from previous ones because it specifies two types of monologic tasks – narrative and opinion – to examine the mode effects on task type. The opinion task seems more susceptible to the mode effect than the narrative task, given that, unlike in grammar, vocabulary, and fluency for the former task, the latter found a significant mode effect only in pronunciation. With regard to identifying speaking ability, which a single factor may represent, it is tentatively considered monologic speaking ability, although further study is required.

Suggestions for the current validation study include the need to (1) further explore the effect of test mode and task type (monologic and

interactive) on test construct, as in O' Loughlin(2001) and Zhou's(2015) studies, and (2) maintain a balance in the number of participants to form a group by gender and proficiency level that allows us to investigate the effect of individual characteristics on speaking performances.

2.1.2.2 Video-conferenced (VC) testing

Another kind of test mode that has been much researched is video-conferencing. Kim and Craig (2012) focused on the process of developing validation arguments for VC low-stakes oral interviews with 39 Korean college students. Evidence was gathered from the test scores of face-to-face interviews and VC tests with one month apart and post interview. The validity argument relied on discussing test usefulness which elaborates the positive and negative theoretical and empirical rationales in terms of reliability, construct validity, authenticity, interactiveness and practicality. The test results proved there was no significant difference between the two modes and the Korean test-takers regarded face-to-face and VC interview modes as similar.

Nevertheless, there seems to be some room for improvement in their validation argument. First, Kim and Craig (2012) echo Zhou's (2015) suggestion for the need to further research the relationship between task types and test takers' English proficiency level. Second, their study did not specify what components of the speaking constructs can be empirically identified in the interpretation of the test results, which is

probably due to the small sample size, limited proficiency grouping and tasks in the interviews. Lastly, the authors reported that the small screen size limited test-takers' view of interviewers' non-linguistic gestures and facial expression cues. Accordingly, technical affordances that the test mode can support to meet the demands for test-takers' testing performance should be analyzed and properly integrated when designing and operationalizing a test as they can serve as a critical source of evidence to support test validation.

At a larger scale, validation studies of a newly developed VC(Video-Conferencing) speaking test were conducted in comparison with the IELTS face-to-face oral interview test (Nakatsuhara et al., 2017; Berry et al., 2018). MFRM analysis provided evidence for scoring validity in terms of four facets - test-taker, task version, rating scale, and rater. In the quantitative analysis, over 200 test-takers' perceptions of the VC speaking tests, including sound quality and examiner training, were satisfactory, as revealed from the questionnaires and focus group discussion. The test results suggest that if test-takers get more used to the VC test, they would barely find any difference between the two modes.

As for the implications of the study, Nakatsuhara and colleagues (2017) elicited more explicit language to negotiate meaning, which means the nature of VC communication does not always allow for subtle ways of establishing mutual understanding and negotiating turns. Thus, the speaking construct to be measured in the VC test should be operationalized in the form of more explicit negotiation of meaning and

turn management to embrace these aspects of test-taker language as part of the construct. Lastly, the VC mode studies call for devising an alternative measure in case of online disconnection.

2.1.3 Virtual environment (VE)-based testing

A web-based VE, made possible by advances in interactive computer technology, has become an increasingly promising area of language assessment that can facilitate productive or interactive speech communication, as well as task engagement and authenticity in L2 assessment and learning. Immersive VE resembles a physical place in real life via three-dimensional graphics, motion-speech synchronization, and video communication platforms (Ockey et al., 2017). Users have the unique opportunity to not only observe their self-representation embodied as an avatar within a 3D environment but also engage in real-time oral communication with multiple individuals through text or voice. This communication is facilitated by internet-connected computers or devices, allowing users to interact with others from anywhere at any time.

Ockey, Gu, and Keehner (2017) concluded from their initial testing of different nationalities that the test-takers' level of English competence and the role of affective factors (e.g., test anxiety) involved in the experience of the new test mode should be taken into account. When test-takers participated in the semi-direct test, they expressed feelings of nervousness and a sense of lacking control. This was primarily

attributed to the fact that the machine controlled the test-taker's role, and as a result, they did not receive any assistance or support when they encountered difficulties during the test.

York (2019) aimed to investigate how different modalities (virtual and face-to-face) and task complexity levels influenced oral performance in language learners who were assigned to either virtual or face-to-face tasks with varying levels of complexity. It was suggested that the virtual modality had a positive impact on oral performance, particularly in terms of fluency and accuracy. Furthermore, the effects of task complexity on oral performance were more prominent in the virtual group compared to the face-to-face group. This implies that the virtual environment might be particularly beneficial for learners when dealing with complex tasks.

On the other hand, in the VE context, which can provide the notion of social presence via avatar representation, it is probable that participants feel less anxiety and stress than in a face-to-face environment owing to the greater sense of anonymity in VE interactions (Wang et al., 2009; Liou, 2011). However, it seems necessary to further investigate the degree of the difference in test-takers' affective responses to the immersive VE that are due to individual characteristics.

The second concern revolves around whether using Virtual Environments (VE) may lead to an underrepresentation of the oral communication construct. This is because VE might not necessitate test-takers to demonstrate certain pragmatic competences. There is a

question about the validity of generalizing the findings from direct speaking tests to a test-taker's ability to engage in collaborative interactions in a real face-to-face setting. However, languages for more specific purposes and contexts (i.e., a pilot communication with a control tower; Douglas, 2000) can be easier to simulate with a VE than in a face-to-face situation. The key to the challenge of VE-delivered testing is to create an authentic context by simulating a more accurate portrayal of the TLU, the context or situations on the screen where a test-taker is using a target language.

Specifically, the so-called "immersive actions" in the VE testing environment can be achieved by making virtual experiences as close as possible to being with others in the same place. Thus, if strong audio-visual cues, context-appropriate objects, and background materials are provided, test-takers will implicitly recognize and produce context-appropriate behavior without any external distractions (Dede, Salzman, Loftin, and Ash, 2000). In this way, the validity of inferences from test scores in the VE environment regarding test-takers' language abilities can be strengthened.

Recently, demand for mobile devices for educational purposes is growing as they support multiple types of learning including, but not limited to: (a) experiential, situated, and context-aware learning; (b) self-regulated and hands-on project learning, AR mobile learning; and (c) inquiry learning (Chiang, Yang, & Hwang, 2014; Swanson, 2018; Traxler, 2010). As a channel that mediates AR and the real world, mobile devices

are a time-and location-independent medium for delivering personalized and context-aware learning content; creating proper environments of ubiquity, and interactivity; facilitating collaboration among learners; and providing seamless bridging between contexts in both formal and informal learning (Sung, Chang, & Liu, 2016; West & Vosloo, 2013).

Therefore, mobile devices can effectively support new and innovative question types and assessment activities augmented with virtual or real physical elements (Santos, Hernádez-Leo, Pèrez-San Agustín, & Blat, 2012) to evaluate EFL students' learning needs. Nikou and Economides' (2018) meta-analysis summarizes numerous studies addressing a wide range of applications that mobile devices can support in assessments such as self- and peer-assessments (Chen, 2010; Lai & Hwang, 2015), formative assessments (Hwang & Chang, 2011), performance-based assessments (Campbell & Main, 2014), adaptive and personalized assessments (Sung, Chang, & Liu, 2016; Triantafillou, Georgiadou & Economides, 2008), game-based assessments (Wang, 2015) and assessments with AR features (Chao, Chang, Lan, Kinshuk, & Sung, 2016).

Among the various applications of mobile devices, this dissertation seeks to bring attention to a new test delivery mode: AR technology in a mobile platform. The theoretical basis for the AR system is situated learning, which emphasizes the importance of integrating human-technology-context interactions. Multi-modal and contextual information mediated by AR in a hand-held device is generally known to help learners understand how knowledge can be used in a certain situation and

feel highly engaged and motivated in a learning activity, thus facilitating understanding.

In language learning context, a number of teaching and assessment studies highlight the cognitive effects of AR on participants' individual factors such as proficiency level and learning style. Wang (2017) reported that AR techniques aid the intermediate-level students the most in their writing performance in the aspects of content control, structure, and wording. This study shows that AR technology may provide learners suitable learning scaffolding to transform their thinking into specific words and assist them in recalling their experiences related to the writing topics.

Hsu (2017) investigated the effect of AR on learning styles — sequence or non-sequence oriented — by comparing two AR educational game systems for third graders that help them learn English vocabulary in free and situated surroundings. Two systems were devised, one self-directed learning approach that did not restrict the learning sequence and the other a task-based learning approach that controlled the learning sequence. The results showed that both approaches were highly effective in promoting learning. In addition, students with a serial learning style expended lower mental effort and had less foreign language learning anxiety regardless of the systems used, although the challenge level and control of the system was matched to the students' proficiency.

Chao et al.(2016) conducted a validation study of performance assessment that integrated mobile AR technologies into a cooking course. In the action research process, students completed their work in three

modules: (a) authentication (assessor / assessee); (b) AR context awareness; and (c) AR interaction. The three validation framework methods were: (a) a questionnaire covering the effect of performance assessment, mobile service, and AR technology; (b) test results compared with pen-and-paper assessment; and (c) interviews.

The results indicated significant score differences between the two assessments with scores higher in the mobile AR-supported performance assessment. Most of the students and teachers agreed that the mobile AR technology allowed for high autonomy and provided good visual effects, making students more attentive to presentations, interactions and feedback in the assessment process.

To be brief, while a summary of the aforementioned evaluative research on different types of test mode is presented in Table 1, several considerations are made from the extant literature in terms of significant validation issues : comparability of MARST with assessments of other language test modes, the extension or constraints of the language construct to be measured due to the integration of the new MAR mode, and the interaction of MAR with task conditions (i.e., task type and complexity) and test-takers' individual traits, raters (because raters use the same mobile application with test-takers for different reasons) as well as various testing contexts (i.e., local performance condition and assessment purpose). And this may imply the potentially sizeable influence of test mode upon test performance and thus expanded modelling of speaking test performance with test mode is necessary, as claimed in 2.2.4 for test

development and empirical validity studies.

Table 1. Summary of selected evaluative research on test mode

Mode	Reference	Feature
Face-to-face	Fulcher(1996) Johnson(2001) Van Moere (2006)	· Measure interactional communication · Mode effect of interviewer's elicitation skills and test-takers' personality traits
Tape	O' Loughlin (2001)	· Measure monologic ability · More lexical density
Computer	Zhou(2015)	· Mode effect on grammar, vocabulary, fluency in opinion task · Measure general language skills
Video-conferencing	Nakatsuhara et al. (2017) Berry et al. (2018)	· More explicit language from negotiation of meaning · Lack of mutual negotiation and turns · Measure speaking skills
Virtual environment	Ockey, Gu, and Keehner (2017) York(2019)	· Less stressful atmosphere for speaking · Mode effect of task complexity on oral performance · Beneficial for complex speaking task
(Mobile) AR	West & Vosloo, (2013) Sung, et al. (2016) Chao et al.(2016) Hsu(2017) Wang(2017) Weng et al.(2020)	· as a mode of performance assessment, self- and peer assessment · Beneficial for intermediate learners in writing · Improved visual effect and attention to task presentation · Providing interactive environment

2.2 Models of L2 speaking test performance

Several models describing L2 speaking performance are reviewed in this section, and the model for the development and validation of the MAR-based speaking assessment will be suggested. According to Fulcher (2015), after Kenyon (1992) developed the first model, a series of research — McNamara(1996), Skehan(1998), Fulcher (2003) — have

established the procedural framework for describing the process of speaking test performance, factors affecting test-takers' performance (i.e., language ability, test conditions, tasks and rating criteria) and the relationship between them.

2.2.1 McNamara's model(1996)

McNamara constructs the model which largely consists of candidates' test-taking and raters' rating processes, as illustrated in Figure 1. Based on the communicative language ability model proposed by Bachman(1990), this model highlights the interactional features of performance assessment with a particular focus on the rating process.

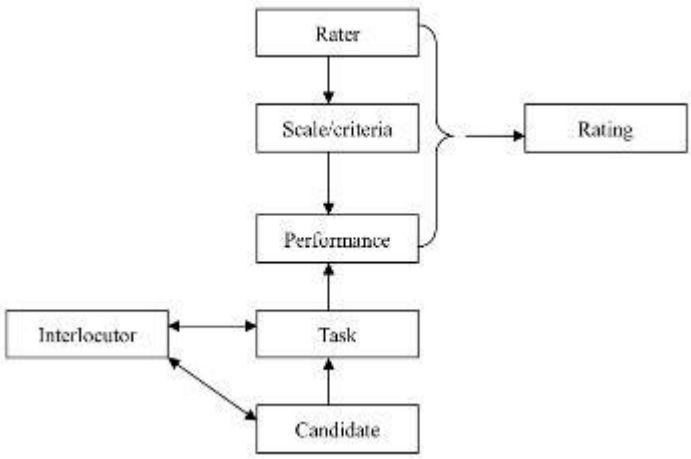


Figure 1. Proficiency and its relationship with performance (McNamara, 1996)

A few major factors of speaking performance are identified as the tasks that bring about performance, which the rater judges via scale criteria. It seems, however, that performance is the only single factor

leading to the score decision and interpretation without further taking into account other contextual factors of the test-taking conditions and its interactions with candidates.

2.2.2 Skehan's model(1998)

Skehan(1998) proposes a model of oral test performance where the task dimensions and candidates' ability are further analyzed. The task dimensions are divided into task qualities and task implementing conditions, while the candidate dimension is separated into ability for use(dual-coding) and underlying competence, as illustrated in Figure 2.

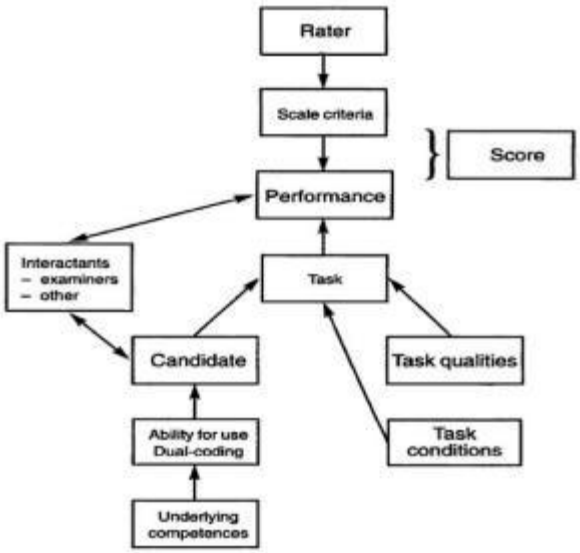


Figure 2. Model of oral test performance (Skehan, 1998)

Fulcher(2003) states that there are three main factors that impact test scores in Skehan's model: the interactive conditions of the

performance, the test-taker's abilities, and the task conditions and qualities used to elicit performance. In O'Sullivan et al.(2002), however, the candidate's abilities seems to have been less discussed while factors involved in test tasks were analyzed in detail.

Skehan's model sheds light on test tasks with some factors that might affect the task difficulty, and thus, test scores and outcomes. For example, a couple of studies (McNamara, 2002; Norris, 2002) proved that task difficulty could be adjusted by setting different task conditions.

In this sense, it can be inferred that describing task characteristics and conditions as required for specific contexts is a highly important part of the test validation process, helping language test developers manage the task quality.

2.2.3 Fulcher's extended model(2003)

Fulcher devises a model of speaking test performance that specifies an extensive range of factors that have been investigated in speaking assessment research, as illustrated in Figure 3 : raters, rating scale, test takers, and test tasks (Fulcher, 2015).

As one of the most distinctive features of this model, the overall elements necessary for the development and validation of a speaking assessment are specified interactively, for instance, making connection among score inferences, decisions/consequences and test taker.

In addition to providing a detailed description of task conditions, there are

several test taker factors derived from other than language abilities on constructs. The ones that are more relevant to the situation of testing performance include task-specific knowledge and skills, and real-time processing, and individual characteristics, all of which the previous models seem not to have taken seriously.

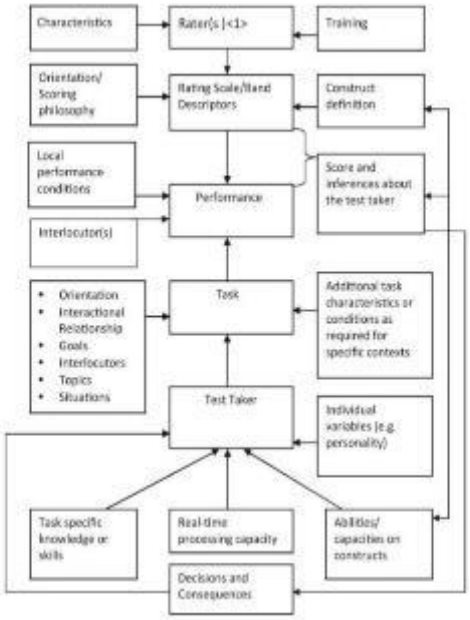


Figure 3. Extended model of speaking test performance (Fulcher, 2003)

Nevertheless, the development of speaking assessment cannot be discussed without mentioning the test mode or instrument. The choice and use of test mode affects the articulation of the validity of speaking assessment as the impact of test mode on individual test takers, tasks, and test-taker's performance, and the intended test constructs should not be underestimated by test developers and researchers (as demonstrated by the test method research outlined in 2.1.2). Therefore,

it is suggested that the “test mode” by which a language assessment is delivered to the test-taker be the next component to include in the revised model.

2.2.4 Implication on MAR-based speaking test

Clearly, the review of the models of speaking test performance indicates the expansion of the past decades in developing language acquisition and language assessment. It provides test developers with important implications for understanding and defining the speaking construct and interactions among variables that affect test-takers' performance.

As far as MAR-based speaking tests are concerned, a revised framework model that can be proposed here to integrate the test mode in the center of the framework as a test platform, where considerations of test-taker, technical affordances of the test mode and decision-making involved in the entire process of assessment stay in balance for validation. The MAR test platform is where test administration, performance and scoring take place altogether. Test-takers are expected to perform by interacting with the tasks, and raters obtaining access to test-takers' performance data and score them with reference to the rating scales.

Thus, the MAR test mode can immediately influence the test-taker, task, and performance, as can be seen in Figure 4 suggesting how to contextualize the MAR technology for assessment design. Examples of

impactful elements include: (a) individual's cognitive and affective variables; (b) test mode and local performance conditions created by the test mode during performance; (c) a number of decisions associated with tasks, target constructs, and scoring.

Usually it becomes possible to find evidence of characterizing the impact of the test mode on test-takers' performance in the process of observing and analyzing the speaking performance during the test. Investigating the relationship among test method, test task and test-taker's performance is a primary consideration in conducting the construct validation research. Considering the test method factor is important in that factors that affect the use of language in language tests to a large extent serves as a subset of factors that determine language use in general (Bachman, 1990). In other words, the MAR test mode can be now referred to as a construct-relevant factor.

The next section outlines the cognitive features of MAR technology, from which the affordances and implications of the VE that it can offer for making fundamental decisions in assessment design and practice can be derived.

2.3 Affordances of augmented reality (AR) technology

Technology can create a learning and assessment platform where learner's affective and cognitive processes intersect (Lajoie, 2014). What matters here is what attributes the media have and how they are

used. With this in mind, some cognitive principles and concepts of AR are presented to provide a glimpse of how AR media can be adopted to support test-takers dealing with the cognitive demands of task complexity and assessment difficulties.

2.3.1 Integration of text and picture comprehension

The integrated model of text and picture comprehension (ITPC), which is also termed the "modality effect", highlights the positive effects of using a combination of text and pictures rather than text or pictures alone. It is assumed that combining these two can expand effective working memory capacity, thereby reducing the effects of excessive cognitive load (Low & Sweller, 2005). A mixed mode of information presentation is more beneficial than a single-mode for poor readers and readers with low prior knowledge. Moreover, it is preferable that the words and pictures are semantically related and presented close together in space and time and that the text is spoken rather than written (Mayer, 2014a). For example, spoken text with pictures results in better learning outcomes than written text with pictures, which also holds true to animation, as explained in 2.3.3.

Since AR is known as an environment generating a composite view combining a real scene viewed by the learner and additional information generated by the computer, it is very suitable for the just-in-time presentation of procedural information (Mayer, 2014a). Such semantic

coherence and contiguity (or proximity) features in integrating verbal and visual information creates opportunities for the use of AR as a mode of assessment by which a task requires task-takers to assume a cognitive burden when processing a certain degree of topical knowledge and transferring it to achieve test performance.

In the literature, transfer is presumably enhanced when learning and application environments are similar. Multimedia environments allow more sensitive and accurate assessment of learner knowledge by increasing authenticity in multimedia materials used for testing contexts and activities and by getting test-takers to engage in deeper processing (Kopriva, 2008). Similarly, the adoption of the AR mode in language testing is more likely to assess a test taker's language ability in various simulated contexts generated by virtual integration of digital and real-world environments than it is in conventional assessment situations.

2.3.2 Social cues: personalization, embodiment and voice

To effect progress in a multimedia-based program or platform, test developers must also take into account social considerations that affect learners' motivation to engage in cognitive processing (Mayer, 2014b). Social cues that increase social presence, that is to say, a feeling of interacting with another social being, are designed to promote interactivity which involves mutual actions and reactions with a learner. The effect of social cues are based on the so-called "personalization

principle," which suggests people learn more deeply when the words in a multimedia presentation are in a conversational style rather than a formal one; for instance "you" and "I" or making self-revealing comments rather than relying solely on third-person pronouns.

Animated pedagogical agents or intelligent virtual tutors are examples of virtual characters employing verbal and facial expressions and gestures to create affective learning experiences, known as the embodiment principle. When on-screen characters display human-like movements (for example, directing learners' attention through pointing), eye contact, and facial expressions, people are said to learn better, yielding a small to medium effect size (Mayer, 2014b). There is also solid evidence in support of the voice principle, which contends that human voices serve as social cues.

2.3.3 Animation

The next most prominent technology-driven change in how information is presented in education is probably the use of animation. As summarized as the animation processing model, animation has two important functions: representing and directing. The representing function means that animation displays the spatial and temporal structure of objects and events, such as changes in position, color, size, and form, as well as the three-dimensional shape of static objects, which permits a virtual walk-around of objects (Lowe & Schnotz, 2014).

Animation can be used to direct learners' visual attention to task-relevant features of the displayed information by omitting irrelevant aspects of information and depicting aspects that may not otherwise be visible. On the other hand, static pictures are more effective comparing different states than an animation because a learner's perception is selective and her cognitive processing capacity is limited, causing a trade-off between the processing of spatial and temporal patterns (Lowe & Schnotz, 2014). It should be noted here that the use of AR technology allows for more opportunities to select different functions of both animation and static pictures depending on the purpose and features of test tasks.

2.3.4 Implication: Connection to language assessment design

Features of AR technology are briefly summarized as multi- or mixed-modal presentation of virtual and physical materials, socialization via simulated human-like embodiment and voice cues, and animation processing functions. They are presumably useful for: (a) stimulating learners' cognitive process of information and in-depth understanding; (b) providing them increased feelings of task engagement as well as interactiveness; and (c) creating simulated situations that are as close to real life as possible.

Combined with a hand-held mobile device, the MAR platform adds such features as mobility, immediacy, and autonomy. This contrasts with

several serious constraints of conventional speaking tests that may impair test validity and practicality. For example, in a face-to-face oral test, test-takers and administrators must be in the same place at the same time. Test-takers are concerned about rating biases associated with interaction with interlocutor, on-site rating, testing order, and time concerns and expenses. Score reports are also delayed which creates a distance between testing and the provision of positive feedback regarding learning.

Although some of these issues have been addressed to a great extent by alternative forms of assessment; for example, CBT and VC tests, it is still necessary for test designers to expand the kinds of testing tools and platforms they employ. By doing so, test developers can accommodate both various test purposes and test-takers' needs and assess what they want to see from test-takers' performance in appropriate conditions capturing important elements of knowledge and skills required in the assessment.

Consequently, the question emerges: how can we ensure that the MAR-delivered test environment provides the right kind of environment and affordances? Designing and evaluating the MAR test platform is almost compatible with making assessment design and validation. While the affordances of the MAR platform increased interactions in both physical and interactional dimensions, vivid and immersive contexts/situations via visual information are theoretically concerned with such aspects of test qualities as authenticity, interactiveness,

practicality, and fairness, all of which need to be discussed in the validation process.

We can start creating the connectivity between MAR technology and L2 speaking by first incorporating several key elements: technological affordances, the test-taker, assessment decisions, and, most importantly, the central platform where both technology and assessment components converge, the connectivity between MAR technology and L2 speaking assessment can be proposed. Building upon this connectivity and the review of various models of L2 speaking test performance presented in 2.2, the proposed MAR-integrated L2 speaking assessment model is depicted in Figure 4.

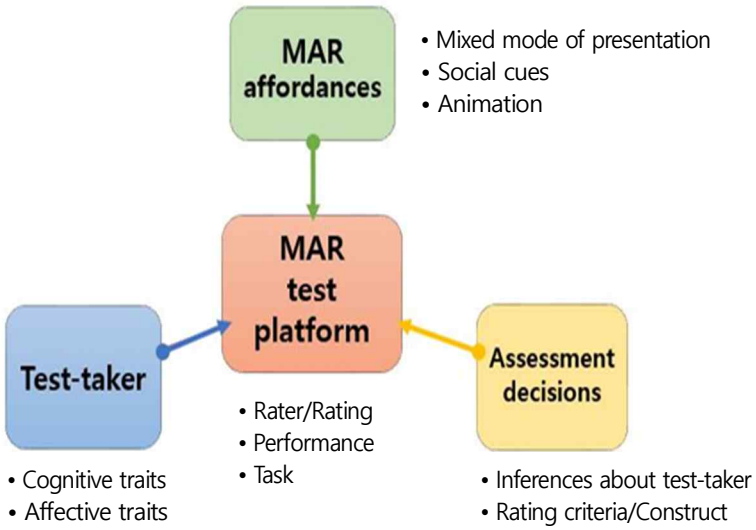


Figure 4. Framework of the MAR-integrated L2 speaking assessment

In order to explore the role of AR technology in eliciting relevant features of speaking skills, the first thing to do is to examine the construct

of speaking and the challenges associated with its assessment. Therefore, the next section will outline the theoretical areas of concern, which include the nature of L2 speaking assessment, encompassing skills, knowledge, and speech processing.

2.4 Task Characteristics Framework for Test Design

Second language tasks and assessments have evolved in parallel with increasingly multicomponent models of language ability. Building on Bachman' s (1990) model of language ability, Bachman and Palmer (1996) articulates the "task characteristics framework", a detailed framework of task characteristics intended as the basis for both test design and test-related research as the characteristics of tasks are considered the only factors that test developers have direct control upon (Shohamy, Or, and May, 2017) .

Three main activities are involved in the framework: 1) describing TLU (Target Language Use) tasks as a basis for designing language test tasks, 2) describing a variety of test tasks to assess comparability and reliability, and 3) comparing the characteristics of TLU (Target Language Use) and test tasks to assess authenticity (Bachman and Palmer, 1996).

The task characteristics are analyzed in five aspects : (a) the setting; (b) the test rubrics; (c) the input to the task (both in terms of format and language input); (d) the expected response (in terms of format and language); and (e) the relationship between the input and the response.

Each of these aspects of the MARST task characteristics for the current research is delineated from the following sections.

Technology that affects the delivery of the input also affects the way test-takers respond. The reciprocal interaction that occurs in the MARST is qualitatively different from that in conversations among humans. Speaking responses are audio-recorded while a picture of their face picture is being taken. The interaction between the input and response is characterized by improved interactivity and authenticity supported by the affordances of AR, that is, digital access to traditional materials and context-sensitive reference support. Presumably, this will reduce test-takers' cognitive load while promoting communicative contexts, resulting into test-takers' enhanced task engagement.

2.4.1 Characteristics of the setting

The setting characteristics consist of physical setting, participants, and the time allotted to perform the task. The MARST can reduce the administrative burden by having test-takers download the speaking assessment platform from the app store onto their mobile phones. All test materials are electronically accessible with a set of markers to activate AR on the mobile app. As the test is taken in an online format, stable access to internet for test delivery is of particular interest. Participants in this study are students in the first grade of high school who are 16 to 17 years old with varying English proficiency levels. They are very

familiar with mobile communication; thus, the testing equipment is unlikely to affect their use of linguistic knowledge and affects during test administration than tasks delivered via other test formats, such as human interlocutors in face-to-face interviews. In the test under study, test-takers decide where and when to take the test for themselves within a designated period of time. In the absence of a human examiner or proctor, test-takers must agree to have their face videorecorded in real time during the test via mobile camera.

2.4.2 Characteristics of the test rubrics

A rubric refers to the information given to the test-taker about how to proceed in taking the test including instructions, time allocation, test organization, and responses to the test tasks. The focal points of speaking test criteria are the control of language knowledge and accurate use and, the delivery of information with fluent and correct pronunciation in well-organized utterance while taking into account a simulated interlocuter. As Luoma (2004) explains, success in communication-oriented tasks and criteria partly depends on the content and sequence of the test taker's speech in an information-relationship talk, particularly narratives or explanations.

The tasks and instructions are presented in an automatic and consistent manner to enhance test fairness. When test-takers start to run the application, an animated and friendly-looking social agent named "Eco Bear" appears on the screen to interact with the individual test-

takers, providing them with specific contexts and instructions on test tasks in a simple and clear target (English) language.

With the relationship between rating categories and assessment features outlined in Table 2, these tasks assess three aspects of speaking ability: (a) language use, including range and accuracy of lexico-grammatical use; (b) fluency, which focuses on flow of test takers' speech, pronunciation, and intonation; and (c) content/ topic development, which assesses the degree to which responses are appropriate to task requirements, and well-organized.

Table 2. Relationship between rating categories and assessment features

Rating category	Assessment features
Accuracy	Knowledge of vocabulary & syntax ° lexico-grammatical range and use ° structural complexity of sentence Textual knowledge ° use of connectives and conversational organization
Fluency	Knowledge of phonology ° comprehensible pronunciation, liaison and intonation Length of response
Content	Task fulfillment and pragmatic knowledge ° topical understanding and development in terms of specificity and relevance ° ideational and manipulative functions & register

Tasks assigned to test-takers have to revolve around the common theme of the environment crisis. A semi-direct (simulated) role-play needs to be developed utilizing AR technology, wherein test-takers engage in a brief conversation with a virtual friend regarding the creation of recycled items to celebrate World Environment Day. The role-play

task, designed to evaluate test-takers' speaking skills in making suggestions and demonstrating their ability to take action, can be preceded by a warm-up activity that offers valuable hints on how to approach the task. Next, monologic tasks can be developed to assess informational components of speaking competence. These tasks require test-takers to describe the content depicted in the animated AR scenes of an environment poster and express their personal opinion about its purpose.

An additional monologic task can be created where test-takers assume the role of an IT company employee and provide an explanation of the process involved in operating a recycled device after watching an AR-based video clip. Presumably, however, this monologic task can be characterized as semi-interactional in that affecting the test-takers' cognitive processing, the MAR-based input makes them aware of the presence of the interlocutor or a group of rainforest rangers on the mobile screen during task performance.

In other words, this semi-interactional monologic task type would serve as an example of how the technical aspects of MAR mode enhanced authenticity and interactivity by stimulating test-takers' cognitive and affective characteristics as described in 2.2.2. This, in turn, may impact test-takers' language performance and the nature of the monologic task, ultimately enhancing the quality of the construct being measured—the language ability assessed in the test—and expanding the construct itself. Later, this argument, useful for the MARST validation, can be supported

by the qualitative approach involving the analysis of test-takers' speaking process during test administration.

To ensure that MARST fulfills its purpose as a learning assessment and to evaluate its usefulness, the validation process involved establishing a clear relationship between the rating categories and assessment features, as outlined in Table 2. Additionally, it was crucial to establish a connection between the assessment features and the class contents covered, as outlined in Table 3.

According to Table 3, the major learning points of the role-play task type focus on language expressions essential for the communicative functions of making suggestions and expressing the ability to do something. Test-takers can be aware of the proper sequencing of dialogue to fulfill these functions. Additionally, the task requires test-

Table 3. Relationship between assessment features and learning contents

Achievement standards[1]	Task type	Learning contents
Ask about personal life experiences or plans and make suggestions	AR-based warm-up activity & Role-play (Dialogue completion)	Knowledge of vocabulary & syntax ° proposal expressions "How about + ing, Why don't we, Let's + base form of the verb " ° ability expressions "I can, I am going to + base form" ° make A(recycled items) from/out of B (used items) Textual knowledge ° sequence of conversation (suggestion-ability) Knowledge of phonology ° comprehensible pronunciation & flow Task fulfillment and pragmatic knowledge ° proper matching of used and recycled items, formality

[1] These are part of English Language Speaking Achievement Standards in the 2015 revised national curriculum of high school(The Ministry of Education, 2015)

Illustrate drawings, photographs, and diagrams on familiar general topics	Poster description	Knowledge of vocabulary & syntax ° use of sentences with compound, complex, and modifier structures (prepositional phrases), conjunctions ° use of verbs: look/seem/appear to infinitives or adjectives Textual knowledge ° use of connectives Knowledge of phonology ° comprehensible pronunciation & flow Task fulfillment and pragmatic knowledge ° specific descriptions of objects followed by inference about the state of the planet, ideational function (description)
Express opinions and feelings about everyday life or familiar general topics	Expressing personal opinion	Knowledge of vocabulary & syntax ° use of sentences with compound, complex, and modifier structures (relative clauses) and conjunctions Textual knowledge ° use of connectives Knowledge of phonology ° comprehensible pronunciation & flow Task fulfillment and pragmatic knowledge ° clear interpretations on the poster message and personal idea, ideational function (explanation)
Explain or ask and answer the order of events (context or circumstances) dealing with everyday life or familiar general topics	Explaining the order of events	Knowledge of vocabulary & syntax ° use of sentences with compound, complex, and modifier structures (relative clauses, prepositional phrases, imperatives and passives) ° IT-related technical terms Textual knowledge ° use of connectives, word repetition, and pronouns Knowledge of phonology ° comprehensible pronunciation & flow Task fulfillment and pragmatic knowledge ° clear stepwise explanation of the order of events ° ideational(explanation)/manipulative(request, persuasion) functions

takers to apply their knowledge to match used items with their corresponding recycled counterparts.

The primary learning points in the picture description task have a specific focus. They are aimed at assessing the language functions of expressing factual information. This information is presented in the AR-

mediated environment poster. Following this, the task also assesses test-takers' personal opinions. These opinions are in response to the issue depicted in the poster.

Consequently, the important language features targeted for assessment include the use of prepositional phrases and copular verbs such as look, seem, and appear. These verbs do not typically denote action but instead express a state of being or perception.

The focal learning points addressed in the opinion asking task include the use of sentence structures with compound and complex, and modifier structures (relative clauses, prepositional phrases), as well as conjunctions. They are essential to fully express opinions and feelings based on what test-takers see in the picture of the poster.

The learning points of the sequential order task bring attention to the use of prepositions, passives, pronouns, and cohesive devices necessary for explaining dynamic events in a sequential manner, particularly when referring to non-person objects. The knowledge and terminology used during the task are derived from what the test-takers had learned in class. These linguistic elements play a crucial role in fulfilling the communicative functions of explaining stepwise how the device works in the rainforest area and providing guidance to simulated interlocutors who may potentially use the device.

2.4.3 Characteristics of the input and expected response

The inputs indicate the materials contained in a given test task. Notably, the audio and visual inputs are combined through a novel channel of AR. The AR channel includes rich contextual information consisting of images, sounds, and full-motion video, thus potentially enhancing authenticity in both the input and response.

Multi-modal input has the potential to increase the intrinsic interest of the test tasks, thereby strengthening the possibility for greater interaction between test-takers' communicative language abilities and the test tasks. Therefore, it is expected that persistent concerns involved with EFL language testing will be relieved to some extent in that many EFL speaking tests have utilized input and tasks that are too decontextualized in comparison with target language use tasks.

The digital materials serve as input support context-sensitive references in either the mobile app or AR embedded on the app. One example is the animated social agent that appears via either the app screen or AR and interacts with individual test-takers, giving them cues and prompts in replacement of a human interlocutor. Additionally, some visual and/or aural information such as a short text, animated pictures or a video clip serve as clues to construct their responses.

Expected responses include test-takers' language use in limited and extended production formats. The limited production may be as long as a single sentence in response to dialogue completion task, while extended production may be as long as a 4 to 5-sentence paragraph in response to the rest of the tasks that follow the dialogue completion task. Test-

takers need to be trained regarding how to use the MAR app platform, access the inputs and record their responses prior to the test administration. Each task time varies from 4 to 8 minutes, including a preparation time of 20 minutes in total.

2.4.4 Relationship between input and response

This section describes how the input and expected response are related to each other, with regard to the reactivity, scope, and directness of the relationship. According to Bachman and Palmer (1996), reactivity refers to the extent to which the input or the response directly affects subsequent input and responses. The reactivity of tasks can be characterized into three aspects: reciprocal, non-reciprocal and adaptive. In reciprocal tasks, test-takers engage in language use with another interlocuter and, receive feedback so that the exchange can affect subsequent language use. On the other hand, a non-reciprocal relationship between input and response is characteristic of reading, taking a dictation and writing a composition as these tasks have neither feedback nor interaction.

The MARST is characterized as semi-reciprocal in that a simulated social agent or interlocutor appears to interact with test- takers during the test, such as the response that can be expected in Task 1 and 3. The interlocutor's reactions and responses to Task 1, however, are pre-programmed without exchanging instant messages with each other. In

this sense, relative to the face-to-face interview format, the MARST is less reciprocal. Task 1 and 3 with simulated interlocutors are designed to determine whether a semi-direct (or simulated) speaking task delivered via the MAR mode can be a viable and sound alternative to direct speaking tests delivered via the face-to-face interview mode.

The scope of the relationship indicates the amount or range of input that must be processed for test-takers to respond. A broad scope of test task requires test-takers to process a lot of input, whereas tasks with a limited amount of input are characterized as having a narrow scope. By nature, the input of speaking tests do not require as broad a scope as reading tests, which usually request test-takers to read long passages.

However, MARST tasks, in which materials and instructions are highly contextualized, may expect test-takers to process more inputs delivered via multiple modes – audio, visual and animated – than conventional speaking tests. In other words, MARST tasks which are capable of embedding rich and specific contexts expand the aspects of the construct or the mastery of linguistic knowledge and skills to be addressed in speaking assessment. Thus, MARST tasks are characterized as having a broad scope relative to other speaking test forms.

Directness of relationship indicates the degree to which an expected response can be based primarily on information provided in the input. While conventional speaking tests include relatively indirect tasks that ask test-takers to give their opinion about a certain topic without heavily relying on the information in the input, the MARST includes tasks in

which responses are based on information in the input to a greater degree. In summary, Table 4 illustrates that compared with other formats of speaking tests, the MAR speaking test format has the distinguishing feature that the input and the response of a task are closely related.

Table 4. Features of the relationship between input and response of speaking test formats

Relationship between input and response \ Format	Interview	Computer	MAR
Reactivity (− nonreciprocal → + reciprocal)	++	−	+
Scope (− narrow → + broad)	−	+	++
Directness (− indirect → + direct)	−	+	++

2.5 Test Method Characteristics

Apart from the relationship between input and response, however, attention should be paid to the overall test method characteristics since one important purpose of the task characteristic framework is to modify certain test method characteristics to create new testing methods.

It is because how technology may affect interpretations of language test performance is based on the understanding of how MAR is likely to be different from other means of presenting language test tasks. Accordingly, it is necessary to discuss test method differences using a test method framework including the physical and temporal circumstances of the test, the test rubrics, input, the expected response, the interaction between input and response, and the characteristics of assessment.

In this respect, based on Chapelle and Douglas' (2006) discussion of the advantages and limitations of computer-assisted language testing (CALT), Table 5 outlines those of the MARST in terms of test method characteristics. A significant point here is that test developers should describe and analyze the characteristics of the test method to make appropriate technology choices that can affect how and what a test measures, which is also what we refer to as "where technology is integrated into test method."

Table 5. Test method characteristics and advantages and limitations of MARST (adapted from Chapelle and Douglas, 2006, p.23)

Test method characteristics	Advantages	Limitations
° Physical and temporal circumstances ° Location & Time ° Personnel	The MARST can be taken at convenient locations and time if wireless Internet is available; IT expertise and human intervention are less required for establishment and maintenance than the CALT.	Different device models may disturb stable conditions of test administration; security is a critical issue.
° Rubric/Instructions ° Procedures for responding	Test tasks presented in a dynamic manner make the test-taking experience engaging.	AR scenes that pop up under test-takers' control may disturb instructions or inputs from being uniformly represented.
° Input and expected response ° Features of the context: setting, participants, tone ° Format: visual/aural/video	Multimedia capabilities allow for a variety of input types and formats enhancing the contextualization and authenticity of test tasks.	It takes some time for test-takers to learn how to use the MAR-embedded app and produce their responses using it.
° Interaction between input and response ° Reactivity: semi-reciprocal	The MAR app provides test -takers immediate feedback from interlocuters. More interactive than CALT	Less control of testing conditions than CALT because of its mobility
° Characteristics of assessment ° Construct definition ° Criteria for correctness ° Scoring procedures	The MAR app allows for making, storing test-takers' responses, and reporting test scores; contextualized AR-mediated task may affect the construct	The MAR is a new, expensive and limited technology, thus creating potential problems for construct definition and

	definition and scoring criteria to be assessed. validity.

2.6 Validation framework

2.6.1 Historical overview

From the late 1970s to early 1980s, construct validity was introduced in language testing (Palmer et al., 1981). Messick's revolutionary concept of validity (1989) attempts to integrate contents, criterion, and construct validity into a unitary model of validity centering on construct validity and also incorporates social consequences and values into the newly-proposed unitary model. Messick's unified view of validity, defined as an integrated judgement of the degree to which empirical evidence and theoretical rationales support arguments on the adequacy of interpretations and actions based on test scores (Chapelle, Jamieson & Hegelheimer, 2003), made seminal influence on conceptualizing the validation process in educational measurement and language testing.

In 1990s and early 2000s major advancement in developing validation frameworks is manifest in Test Usefulness (Bachman and Palmer, 1996), its revision of Assessment Use Argument (Bachman and Palmer, 2010), Interpretation/Use Argument and Validity Argument (Kane, 1992; 2006; 2013) and Chapelle and her colleagues' (Chapelle, Enright, & Jamieson, 2008) argument-based approach drawing on Kane's framework.

In the 1990s, Bachman and Palmer (1996) proposed the notion of test usefulness to make Messick's work more accessible to language testers (Xi & Sawaki, 2017). They proposed the so-called test usefulness

with respect to five qualities : construct validity, reliability, authenticity, interactiveness, impact, and practicality. As the test usefulness prioritizes the investigations of the five qualities depending on assessment contexts and purposes, there are criticisms that it lacks a conceptual system or structure to prioritize the five qualities and evaluate overall test usefulness.

In early 2000s, Kane' s validation framework proposes the two stages of validation process – formulating an interpretive argument, the so-called 'Interpretation/Use Argument' and evaluating a validity argument when evidence is found from validation research. In the interpretive argument, a coherent analysis of all the positive and negative evidences on a score-based interpretation is used through the chain of inferences of seven types as illustrated in Figure 5 of the next section.

It is acknowledged that Kane's approach provides a transparent framework for language testing researchers and practitioners to prioritize different sources of evidence, integrate them in the process of evaluating the strength of a validity argument, and gauge the progress of the validation efforts (Xi & Sawaki, 2017). Later, to meet the needs of a wider range of language testers and users in the language testing area, efforts to adapt Kane's framework continue.

To be specific, Chapelle, Enright, and Jamieson (2008) refine the Kane's framework, constructing inferential links from performance to a score-based interpretation and use, and making a more elaborate discussion of the pertinent assumptions than Kane's work (Xi & Sawaki,

2017). On the other hand, Bachman and Palmer(2010) intend to simplify Kane's framework to construct what is called 'assessment use argument' for those without professional knowledge. Their test validation process is built upon considerations about intended assessment use and consequences. But, the two approaches commonly call attention to test use and consequences.

2.6.2 Constructing an Assessment Use Argument (AUA)

The structure of an AUA features a series of claims and warrants through which test developers articulate or clearly states their intended inferences from the bottom of test takers' performance to their test scores, to interpretations about their ability, to the decisions that are made, and finally to the consequences of test use as shown in 'Interpretation and Use' in Figure 5. The claim resulting from one inferential link becomes the data that serve as the basis for the next inference in the chain.

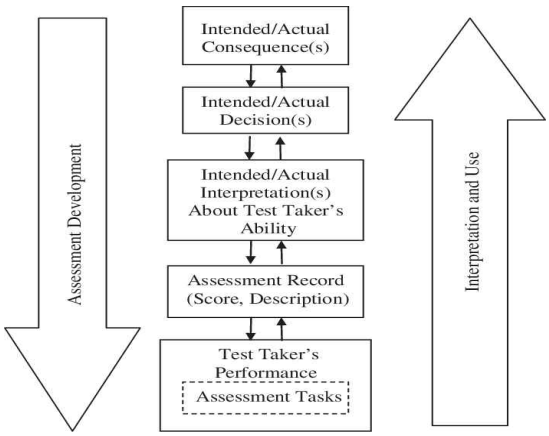


Figure 5. Inferential links from consequences to assessment performance

(Bachman and Palmer, 2010)

Meanwhile, in the opposite direction lies the 'Assessment Development' process from the top considerations of educational and societal consequences and decisions to be made, to interpretations about test takers' ability all the way down to the development of specific assessment tasks. These claims and warrants are subject to rebuttals or counterclaims. A rebuttal means negative evidence weakening the claim or warrant. When a rebuttal is articulated, test developers or stakeholders may investigate appropriate actions to mitigate the negative impact of the rebuttal on test use or to provide evidence that weakens the rebuttals and thus strengthens the claims and warrants in the AUA.

According to Bachman and Palmer (2010), test developers should develop and articulate a specific AUA for each intended use of a test, thereby forming a judgement about the extent to which the uses of a given test are justified. In this justification process, it is important to take into account various contextual factors including but not limited to the types of stakeholders (e.g., test takers, parents, admission officers), the test stakes, the availability of resources, and the cultural, societal and educational value systems of stakeholders.

2.6.3 Interpretation/Use Argument(I/UA) structure

The argument-based validation approach, which consists of an interpretation/use argument (I/UA; Kane, 1992, 2006, 2013) and a validity argument, has long been adopted by language testing researchers to

identify the specific assumptions that require investigation in validation research. A validity argument involved here serves as a means of systematically presenting the positive and negative theoretical and empirical rationales that address the validity of testing outcomes. Rationales are drawn from the validity considerations, test design, and the validation process.

The argument for interpretation and use of the MARST constitutes a chain of seven inferences that include domain definition, evaluation, generalization, explanation, extrapolation, utilization, and consequence implications as illustrated in Figure 6. The domain definition inference connects the TLU (Target Language Use) domain of general English speaking with the observation of performance on the MARST. It aims to investigate whether features of a target domain can be modeled and key knowledge, skills, and abilities can be identified.

The warrant supports this inference by analysis of the curriculum so as to identify some required linguistic knowledge, skills, and abilities – making suggestions/description, expressing one's abilities /opinions about a general topic and procedural information– , all of which Korean EFL high school students are expected to learn in both daily life and academic contexts according to the national curriculum of high school English subjects.

During the test development process, domain analysis legitimizes the assumption about the required skills and knowledge by referring to the national curriculum and the textbook in use that is assumed to abide by what is specified in the national curriculum. Evaluation inference

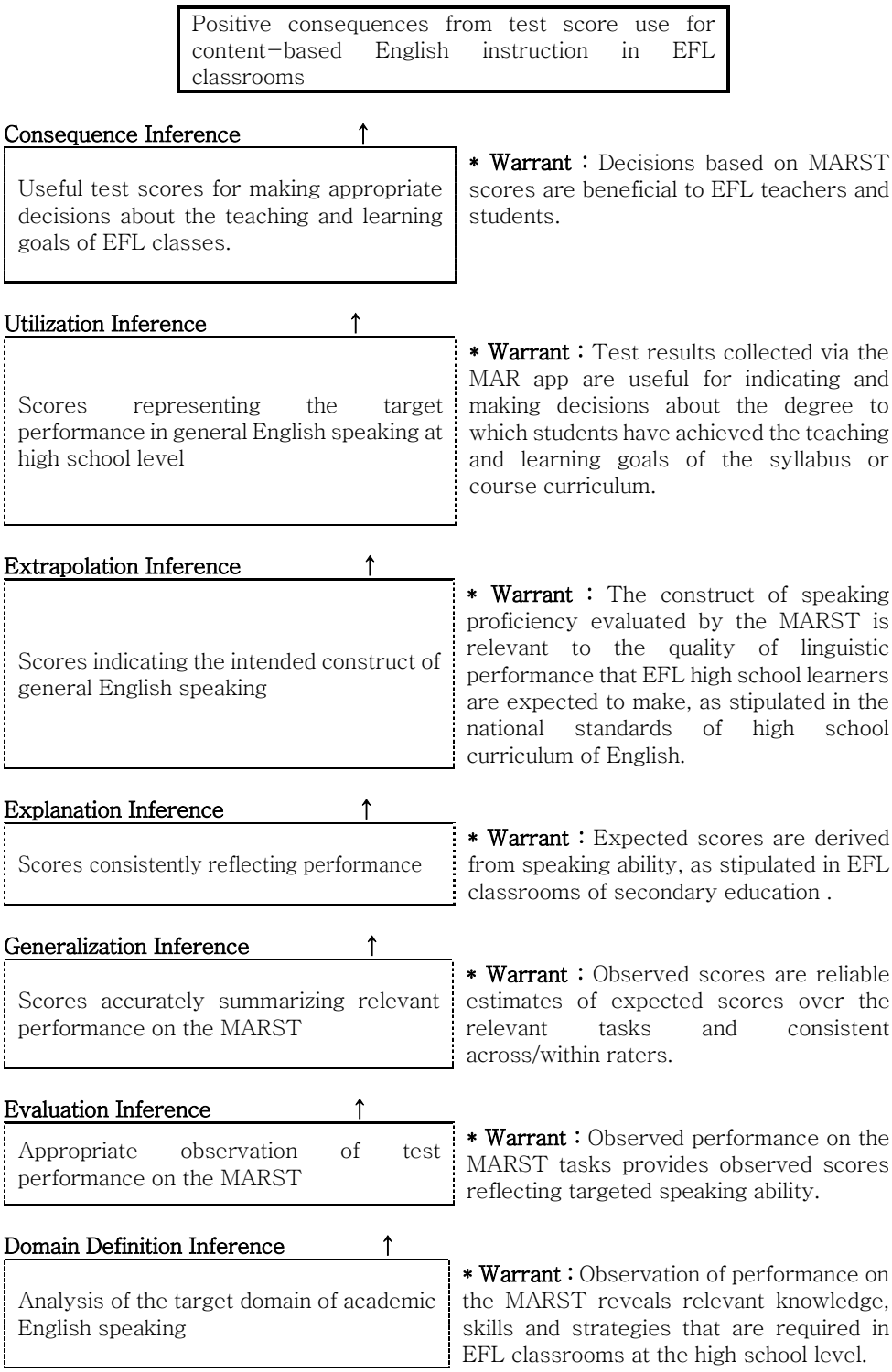


Figure 6. Sketch of the MARST interpretive argument

connects the observed performance with raters' rating outcomes such as observed scores and performance descriptors, aiming to investigate whether the scoring rubric is implemented appropriately, accurately and consistently.

This inference is justified by a warrant that the observed performance on the MARST provides observed scores reflecting the targeted speaking ability. This warrant can be upheld by appropriate task administration conditions, rubric development and rating procedures in which raters perceive that what they assess reflects evidence of students' mastery of targeted abilities. Second, it is necessary that observed scores from student' performance on the MARST adequately differentiate students' abilities across various levels, which Research question 2 focuses on via MFRM analysis.

Generalization inference connects the observed scores to expected scores, which refer to the scores test takers would expect to obtain across different tasks, test forms, occasions, and rating conditions. This inference is supported by the warrant that observed scores are reliable estimates of expected scores for relevant tasks, and consistent across/within raters.

The warrant will be evidenced by investigating whether task, test, and rating specifications are well-defined so that parallel tasks and test forms are created. It also needs to be proved that scoring of test-takers' proficiency is consistent within and across raters. To do so, test task specifications were fully developed in the test design, and rater reliability and consistency was investigated across raters. These two sources of

evidence were extracted from the MFRM analysis to estimate rater severity and task difficulty and analysis of the rating process.

Explanation inference means the degree to which the expected scores reflect the intended construct. It can be used to investigate, whether the scores can be explained using proper theories. It should be proved that the linguistic knowledge, processes, and strategies are relevant to general English speaking for global communication as defined in the national curriculum of high school English.

The warrant also needs to be supported by the assumption that the scores obtained in the MARST are related to scores on other speaking assessments that measure the same general English proficiency by comparing test takers' performance on the MARST with those of other comparable speaking measures such as oral interviews, oral translations, and indirect multiple-choice speaking tests. A comparability study was undertaken using MTMM analysis, followed by factor analysis, which is the focus of Research question 1.

Extrapolation inference is concerned with whether the construct of speaking proficiency used on the MARST is relevant to the quality of linguistic performance in real target (non-test) contexts where general English proficiency is required; for example, high school classroom instruction in EFL situations. This warrant was upheld by examining the relationships between the MARST scores and other indicators of language performance on non-test tasks in high school classroom instruction, which Research question 1 holds to be true.

Utilization inference is concerned with the test use of the target score. The warrant is that test scores collected on the MARST are useful for making decisions about teaching and learning in the EFL classroom. This is supported by instructors and test takers' opinions about the potential use of the MARST in EFL speaking classes, which Research question 3 taps into by means of test user surveys and interviews.

Consequence implication concerns the impact of test use on stakeholders. It should be warranted that the test results collected from the MARST have a positive influence on the course curriculum, test development, and diverse stakeholders (Yang, 2021). This means that the test results collected from the MARST are assumed to provide useful information to stakeholders — test-takers, instructors of English speaking classes, and test administrators — and contribute to the development of an effective curriculum (Yang, 2021). Research question 3 investigates these areas by examining test-takers and instructors' perceptions of the MARST, who are its potential users and stakeholders.

2.6.4 Validation framework for MARST

The major source used to situate the MARST in the validation framework of L2 assessment is a combination of Bachman and Palmer's AUA framework and Kane's Interpretation/ Use Argument and Validity Argument. It is because there is general agreement in the literature of language testing that they are the two dominant conceptual frameworks or

tools to operationalize Messick's definition of validity as a unitary but multifaceted concept.

Both are found useful to develop a test development process or discuss issues on existing validation studies. Thus, it is necessary to be equipped with clear and sufficient or sometimes, nuanced understanding on not only the conventions of AUA and Kane's validity argument but also how each of them functions in what specific way. For example, Chapelle and Lee (2017) mention that the former is intended for a general audience in need of simplified structure expressing core concepts of validation, while the latter is for more research-oriented authors who report development efforts of validation argument.

In the meantime, the two approaches are common in specifying three components – claims, warrants to support claims and backing which states evidence upholding warrants. In addition to them, Kane's validity argument includes seven inferences which create a chain that connects one claim with another as its ground. Thus, Kane's work is often used to clarify the meaning of each component of validity arguments in detail including the meaning of the intended ability or construct, even with the test development process reflected in a domain definition inference.

Comparison of the two validation arguments in Table 6 suggests how this dissertation research combines them by grouping parts that are seen as mutually pertinent in the two systems. It also illustrates detailed specification of warrants and sources of evidences including validation methods that the MARST validation framework actually rests on.

Table 6. Comparison of the validation frameworks developed by Bachman and Palmer (2010) and Kane (1992; 2006; 2013) (adapted from Chapelle and Lee, 2021, pp.34−35)

AUA (Assessment Use Argument)	I/UA (Interpretation/ Use Argument) and Validity Argument	Warrant	Source of evidence
Target domain	Domain definition	Link b/w general English proficiency and observed test performance	Analysis of the national curriculum and textbook
Assessment records	Evaluation & Generalization	Link b/w observed performance and raters' rating Link b/w observed score and expected score	Analysis of task administration and rating procedure MFRM analysis of rater consistency across tasks (rater severity & task difficulty) Analysis of unusual responses in MFRM
Test interpretations · meaningfulness · impartiality · generalizability · relevance · sufficiency	Explanation & Extrapolation	Link b/w expected score and the intended construct Link b/w intended construct and performance in the target context	Analysis of the national curriculum MTMM analysis of relevant speaking measures Factor analysis of internal structure of the MARST scores Discourse analysis of spoken responses Analysis of similarities with other indicators of TLU
Decision/Use	Utilization	Test use of target score	Survey & interviews about test users' perceptions of the MARST
Consequences	Consequence	Impact of test use upon stakeholders	

Chapter 3. Methodology

Validation frameworks are known to specify the process used to prioritize, integrate, and evaluate evidence collected using various methods from three areas: 1) psychometric and statistical methods in education; 2) qualitative methods in language testing by Second Language Acquisition (SLA), conversation and discourse analysis, and cognitive psychology; and 3) the influence of cognitive demands of tasks on task complexity and difficulty (Xi & Sawaki, 2017).

This chapter brings the development process of the MARST into focus. Some essential considerations are made, including the test purpose, the speaking construct to be assessed, the target language use (TLU) domain, the specific tasks involved, and the test structure. It also outlines the methodology of data collection and analysis along with test design to collect and analyze the quantitative data, followed by qualitative data.

3.1 Test development

The test purpose is to examine the extent to which classroom-based assessment intends to assess test-takers' spoken language use based on what they learn in the English classroom. Students are asked to talk about one of the most interesting issues in the contemporary world and to make inferences about their speaking abilities on the basis of the test scores.

3.1.1 Domain analysis

The specific speaking skills to be assessed in the test include making suggestions, expressing one's ability to do something, and providing a visual description. This is followed by expressing one's opinion based on the description, and explaining sequential information. All of these tasks are assumed to tap into discrete but interrelated components of the speaking construct. Moreover, these abilities are currently included as achievement standards in the 2015 revised national curriculum of English for high schools as summarized in Table 7. The publications of most accredited textbooks including the one which the current dissertation based the development of the MARST upon abide by a number of achievement standards including those in Table 7 in different contexts throughout the period over one year or a semester.

The TLU domain, the context or situations on the screen where a test-taker is using a target language, and tasks are concerned with global

Table 7. Relevant English Language Speaking Achievement Standards in the 2015 revised national curriculum of high school(The Ministry of Education, 2015)

Task	Speaking Achievement Standards (excerpted)
1	Ask about personal life experiences or plans and make suggestions
2-1	Illustrate drawings, photographs, and diagrams on familiar general topics
2-2	Express opinions and feelings about everyday life or familiar general topics
3	Explain or ask and answer the order of events (context or circumstances) dealing with everyday life or familiar general topics

environment issues; for instance, World Environment Day, held annually on June 5th, aims to raise global awareness about environmental issues and encourage participation in activities to protect environment. The test tasks include three virtual social agents: 1) a virtual friend to have conversation about a way to celebrate the day, 2) a virtual polar bear, one of the life species most affected by global warming, which prompts test-takers to talk about an environment poster and their opinions, and 3) a virtual rainforest ranger who prompts test-takers (simulated as technical staff in a recycling IT company) to explain how a device for rainforest protection works.

To collect evidence for evaluating reliability, which is one of the qualities of usefulness, an analysis of each task's characteristics, which is addressed in 3.1.4, is critical. Moreover, corresponding scores among different raters is also necessary, which makes pre-scoring rater training an essential component of executing a reliable test procedure. In this study, several raters are asked to score the same MAR-mediated test and calculate the score consistency across the different raters.

Collecting evidence about the test-takers' ability that a test intends to measure - regarded as construct validity - can be drawn from multiple sources. These include studies that seek to determine the relationship of the MART scores with the scores of other measures of the same speaking construct (i.e., oral translations and oral interviews) and the scores of other related language constructs (i.e., listening, reading and writing). Additionally, the extent to which test tasks correspond to real-

life tasks and their engagement of test-takers' language ability can be improved by analyzing the characteristics of test tasks in the next section (3.1.4) and by piloting a demo version of the MARST prototype to a group of non-test taker students, and some teachers and asking their perceptions of the qualities of construct validity, authenticity, and interactiveness of the MARST test. To collect information for evaluating authenticity after test administration, test-takers and raters were asked to describe their perceptions of the authenticity of the test tasks by questionnaire survey and interviews.

According to Bachman and Palmer (1996), areas of language knowledge that are involved in test-takers' responses include organizational (i.e., grammatical and textual knowledge) and pragmatic knowledge (i.e., knowledge of ideational functions of descriptions and explanations, and manipulative functions of suggestions). Organizational knowledge refers to how utterances are organized, comprising grammatical knowledge and textual knowledge. Grammatical knowledge is relevant to knowledge of vocabulary, syntax, and phonology. Textual knowledge concerns how utterances are organized to form texts, and can be categorized into knowledge of cohesion and conversational organization. Pragmatic knowledge refers to how utterances are related to the communicative goals of language users. Notably, a major interest of the MARST speaking test lies in the knowledge of ideational and manipulative functions, i.e., descriptions, explanations, and suggestions.

Construct validation studies of language tests have examined not

only the product of language tests – test scores, but also the process utilized in test-taking (Bachman, 1990). Collecting information on the examination of the process that test-takers engage at the level of individuals is considered highly productive to understand what strengthens the MARST validity.

To collect information for evaluating interactiveness, test-takers were asked to give their opinions on the extent to which their language knowledge, topical knowledge, and metacognitive strategies (i.e., goal setting, assessment and planning) were engaged in taking the test. Lastly, to evaluate impacts, raters and test takers were asked to comment on the fairness of the test, its appropriateness for decision-making, and its specific use or application.

3.1.2 Speaking construct

The construct to be assessed encompasses several aspects of performance – accuracy, fluency and content – in the speaking construct. '*Accuracy*' is a rating criterion for measuring the degree of accuracy and the range of grammatical and vocabulary use. '*Fluency*' measures the degree to which test-takers' spoken responses reflect a natural flow with comprehensible pronunciation and proper intonation without repeated pauses or hesitations. Finally, '*Content*' deals with organizational knowledge regarding how responses are relevant and specific to given tasks in terms of topic development and understanding.

The specific speaking skills to be assessed in the test include making suggestions, expressing one's ability to do something, and providing a visual description. This is followed by expressing one's opinion based on the description, and explaining procedural information. All of these tasks are assumed to tap into discrete but interrelated components of the speaking construct. Moreover, these abilities are currently included as some of the achievement standards in the national English curriculum for high schools as indicated in Table 5. Thus, most accredited textbooks deal with them in different contexts in the three high-school years.

3.1.3 Test structure

The purpose of the test is to measure speaking abilities of test-takers, which are required to communicate topical information about protecting the global environment. Task 1, a role-play, aims to assess test-takers' interactional knowledge and skills to express ability and suggestions about making recycled items on World Environment Day in a conversation with a simulated friend. Prior to Task 1, there is a warm-up activity which requests test-takers match some used items with recycled ones on the AR screen; this activity hints at how to construct their response.

Task 2 aims to assess test-takers' abilities to: 1) describe an environmental poster (Task 2-1), and 2) provide their opinion about its message (Task 2-2). There is also AR-based input that vividly presents

the poster contents in animated form. Task 3 aims to assess test-takers' abilities to explain in ordered steps, how a recycled device called 'RFCx' works after watching an AR-based short video clip on the app screen.

As specified in Table 8, the test comprises three tasks. Task 1 asks test-takers to participate in a role-play with a social agent that appears in the MAR application to complete a dialogue that should include making a suggestion on how to celebrate World Environment Day and expressing their ability to participate or engage in it. Task 2 requires test-takers to describe an environment poster and express their opinions about its intention. Task 3 requests test-takers explain the working process of a device called RFCx (Rainforest Connection), an invention recycled from used cell phones, to rainforest rangers who will use the device on a daily basis.

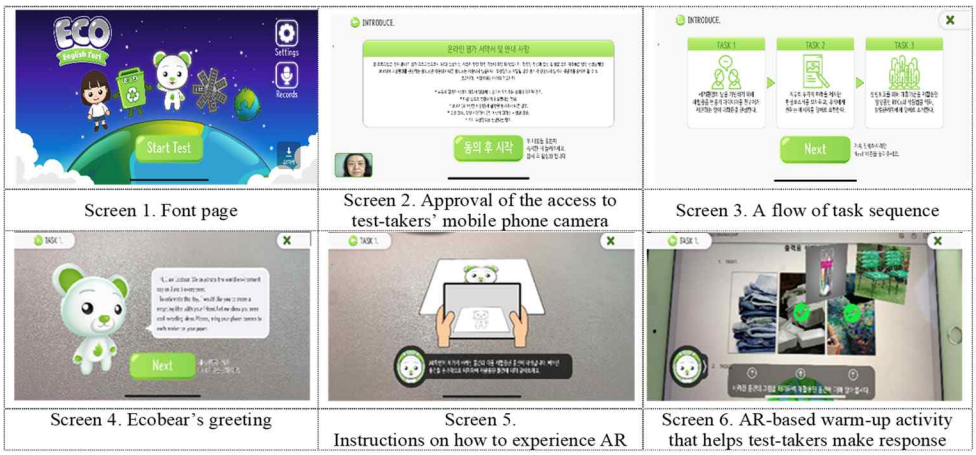
Table 8. Test structure

Task	Type	Time (prep+test)	Cognitive demands	Required language competence	Topic
1	Dialogue completion	3+1mins	Exchanging information (suggestion/ expressing one's ability)	Organizational knowledge & Pragmatic knowledge & Topical knowledge	Recycling
2-1	Poster description (monologue)	5+3mins	Providing specific information about objects	Organizational knowledge & Pragmatic knowledge & Topical knowledge	Environment Poster
2-2	Opinion expression (monologue)		Elaborating/ Justifying opinions		
3	Sequence explanation (monologue)	5+3mins	Explaining topical information in sequential order	Organizational knowledge & Topical knowledge & Pragmatic knowledge	An invented device for saving rainforests

Before the test, test-takers were trained on how to use the mobile

application, which serves as the assessment platform. This training included learning how to access inputs and make responses using AR. The MARST was administered at a place and time of the test-takers' choosing. Security issues, such as cheating or proxy examination, were addressed through a technical measure that recorded the test-takers' faces in real-time during the test.

The screenshots in Figure 7 show the preparation stage before the actual tasks. In this stage, test-takers sign up and grant the application access to their mobile camera, which captures real-time images of them during the test administration.



Screen 1. Font page | Screen 2. Approval of the access to test-takers' mobile phone camera | Screen 3. A flow of task sequence

Screen 4. Ecobear's greeting | Screen 5. Instructions on how to experience AR | Screen 6. AR-based warm-up activity that helps test-takers make response

Figure 7. Screenshots of the pre-test stage on the MAR app

3.1.4 Test task specifications

Task 1 involves a brief role-play with a social agent that appears in the application. Test-takers are required to provide a succinct spoken response to complete a dialogue. This is done by referring to a set of

recycling examples presented before the actual task. The purpose of this task is to assess test-takers' ability to express their own capabilities and make suggestions.

Figure 8 presents a set of screenshots that demonstrate what test-takers do in Task 1 on the MAR application. The task assesses both organizational knowledge – such as the grammatical and textual knowledge required to facilitate a natural conversation flow – and pragmatic knowledge, which is tied to communicative functions that are both ideational and instrumental. The task takes approximately 4 minutes to complete, with 3 minutes allocated for studying recycling examples and 1 minute for producing an AR-based response.

Screen 7. Giving task instructions | Screen 8. Recording the first response | Screen 9. Recording the second response

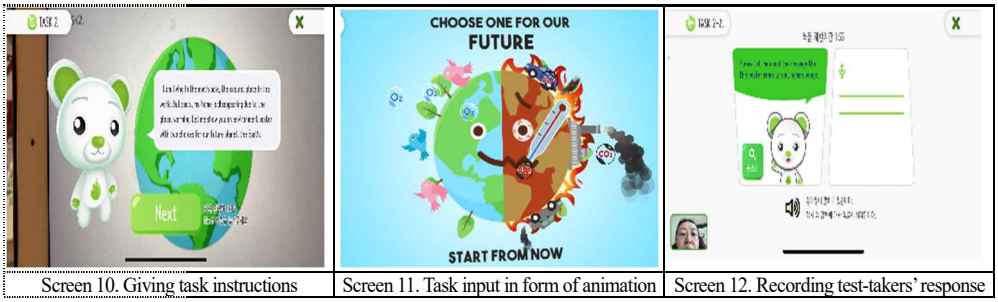
Figure 8. Screenshots of Task 1 on the MAR app

Detailed task specifications, including the scenario scripts, AR presentation designs, suggested answers, and scoring rubrics are illustrated in Appendix 1. Inputs are provided via a new AR channel employing both aural and visual (animation) mediums that enable test-takers to listen to instructions and watch each original item turn into recycled one.

The expected response for Task 1 limits test-takers' language use in terms of production formats to just two single sentences. Consequently, test-takers' language involves a relatively narrow range of vocabulary and syntactical structures needed to cohesively construct a short conversation. The pragmatic characteristics of the responses are informal in register and ideational, as the task requires an exchange of information about recycling between the interlocutors - the agent, Eco Bear, and the individual test-takers. Raters score test-takers' responses using an analytic scale, with the rating criteria including accuracy, fluency, and content (for more details, see Appendix 1).

Task 2 requires test-takers to describe an environmental poster and express their opinion about its message. The presentation of the input in this task is characteristic of AR mode, which brings the static picture of the poster to life with an animation effect describing two opposite choices for the Earth - with or without the effects of global warming. This task aims to assess test-takers' language abilities used to describe topical information presented in an environmental poster, both visually and verbally. The sequence of Task 2 is illustrated in Figure 9.

Inputs are provided by a new AR channel employing both aural and visual mediums (animation) that allow test-takers to hear the narration of the agent Eco Bear and recognize two contrasting options for the future



Screen 10. Giving task instructions	Screen 11. Task input in form of animation	Screen 12. Recording test-takers' response

Figure 9. Screenshots of Task 2-1 and 2-2 on the MAR app

of the earth, to make a choice, and give reasons for their choice. Task 2 intends to assess test-takers' organizational knowledge such as the grammatical and textual knowledge required to make a cohesive monologue and topical knowledge about the global warming effect.

Expected responses for Task 2 entail test-takers' use of both limited and extended production formats within approximately 8 minutes, which includes 5 minutes for preparation and 3 minutes for response. Specifically, the language characteristics of test-takers' responses should showcase a broad range of vocabulary and syntactical structures, with suitable cohesion to interconnect multiple sentences into a paragraph. The pragmatic characteristics of the responses are formal in register, comprising a mix of ideational and manipulative elements in terms of communicative function. This is because the task requires individual test-takers to describe the visual and verbal information presented in the poster about the effects of global warming and suggest appropriate actions. Raters assess test-takers' responses using an analytic scale with the same rating criteria as Task 1, although the

descriptions of each criterion differ slightly (see Appendix 1).

Task 3 requires test-takers to explain in sequential order how a device designed to save rainforests functions. After watching an animation delivered via the AR channel-without subtitles or audio-test-takers are expected to provide a relatively concise spoken response. The aim of this task is to evaluate the test-takers' abilities to relay topical information in a sequential manner, explaining the operation of the device based on the AR animation. Figure 10 provides a visual representation of how Task 3 is presented to the test-takers.

Screen 13. Giving task instructions | Screen 14. Video (nonverbal) task input | Screen 15. Recording test-takers' response

Figure 10. Screenshots of Task 3 on the MAR app

In Task 3, the expected responses from test-takers are characterized by a variety of language knowledge elements. This includes grammatical and textual knowledge used to express a logical sequence of operations for the device, facilitated by the use of cohesive devices. It also involves pragmatic knowledge relating to the ideational function of communication, such as explanation, and formal register, which incorporates technical vocabulary necessary to explain the sequence of the device's operation. Key terms include "solar-powered", "solar panel",

"illegal logging", "rangers", and "send the signal to a cloud". The task can be undertaken at a location and time convenient to the test-takers and is expected to last approximately 8 minutes, including preparation time (5 minutes) and response time (3 minutes).

3.2 Participants

The test-takers are 203 public high school students (110 males and 93 females), aged 16 to 17. Among them, all males and 37 females live in Jinju, while 56 females live in Changwon. The two cites belong to Gyeongsangnam-do, the province located in the south-eastern part of South Korea. Most of the test-takers do not have experience of studying abroad and demonstrate mixed levels of English proficiency, (i.e., low to high-intermediate), given that the general English proficiency level of the province is ranked relatively low nationally[2].

Having learned about with World Environment Day and the details of the device in class, however, they were aware of the topic of the test tasks to a large extent and some basic communicative functions and specific language knowledge necessary to serve these functions. The four raters consist of two female pre-service teachers, one female in-service teacher with over 20 years of teaching experience, and a male native speaker from the U.S. with 5 years of teaching experience.

[2] The Ministry of Education announced the result of the Korean SAT scores administered in 2021. Visit the following site and check the ratio of the English scores by province at page 21. https://www.moe.go.kr/boardCnts/viewRenew.do?boardID=294&boardSeq=89967&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N

Rater training was conducted in two groups: one group comprised the researcher and the native English speaker, and the other included the researcher and the two Korean raters. Both groups followed the same three-step process: a pre-rating session to discuss the rubrics and rating strategies, a rating session, and a post-rating session to reconcile any score differences that exceeded two levels or more.

Prior to the pre-rating session, the researcher prepared audio response samples that were representative of each score level. These samples were used to facilitate a discussion with the raters about the linguistic and paralinguistic features of the responses. Additionally, several other data samples were provided for training and discussion to ensure that the raters understood the rating scales and could identify the linguistic features corresponding to each level.

In the first rating round, each of the three raters evaluated the same 50 data sets, a quarter of the total 200 sets, with each set containing a test-taker's responses to four tasks. After this, in the post-rating session, the researcher had a discussion with any raters whose scores had discrepancies of two levels or more, in order to make necessary adjustments. Following this, the second rating round was carried out on the remaining data sets with the most reliable rater, as determined by their agreement with the researcher in the previous steps. The entire process is summarized in Table 9.

Table 9 . Rating procedure

Session	Pre-rating		Rating		Post-rating
	Group training	⇒	Individual scoring	⇒	Individual adjustment
Description	▶**Group 1**: the researcher and the native rater **Group 2**: the researcher and the two Korean raters ▶Each group discusses the rating categories and level descriptors by analyzing some samples of different levels. ▶Let the raters practice scoring other samples with the rating scale. ▶Compare their results with each other and the researcher.		▶**Round 1** : Raters are respectively given scoring assignments of spoken data sets of fifty test-takers. : During rating, the raters are asked to leave the researcher brief notes on what are thought uncertain or make direct contact with the researcher in order to resolve any issue occurring in the rating process.		▶Individual contact is necessary in case score discrepancies exceed two levels or more for final decisions. ▶**Round 2** : The remaining spoken data sets for scoring are assigned to the rater whose rating results are found most compatible with those of the researcher.

3.3 Data analysis

3.3.1 CTT and MTMM analysis

The proposed study employs a mixed-methods approach to data analysis, integrating both quantitative methods – including the Classical Test Theory (CTT), Multitrait-Multimethod (MTMM) analysis, and Multi-facet Rasch Measurement (MFRM) – and qualitative methods such as corpus and discourse analyses.

The quantitative methods of validation require data collection from multiple measures of different abilities or traits, as well as measures of

the same trait, as outlined in Table 10. In this study, the focus is on the trait of speaking ability. This approach is inspired in part by the Multitrait–Multimethod (MTMM) matrix hypothesis proposed by Campbell and Fiske (1959, as cited in Bachman, 2004).

Table 10. MTMM Design for quantitative analysis

Method / Trait	AR	Multiple choice	Interview	Translation
Speaking	●	●	●	●
Listening		○		
Reading		○		
Writing		○		○

●: measures of the same trait and different methods
○: measures of different traits and different methods

The data for the quantitative approach were collected from several sources: (a) ratings of test–takers' speaking performances in the MAR–based application, as illustrated in Figure 11; (b) scores from a general English proficiency test ; (c) scores from a listening test that was held nationwide in September, 2020; (d) scores from an achievement test that test–takers took at school in October, 2020; and (e) scores from two oral performance assessments (an oral translation and oral interview), and lastly 6) scores from a writing translation test held in December, 2020 in which test–takers translated Korean sentences into English.

The Classical Test Theory (CTT) yields a variety of data, including the descriptive statistics of the MARST test scores across tasks and scoring criteria, item difficulty and discrimination, as well as reliability measures. The

MTMM matrix is a technique used to check the validity of a measure by comparing it to multiple other measures. According to this theory, correlations between measures of the same trait (monotrait correlations) should be higher than correlations between measures of different traits using different methods (heterotrait–heteromethod correlations).

Figure 11. A screenshot of ratings in the MAR app

The Multitrait–Multimethod (MTMM) correlation matrix will subsequently be used as a foundation for developing a confirmatory factor analysis (CFA) model that includes several trait factors and multiple method factors. The CFA model posits that the different measures will have factor loadings on their corresponding trait and method factors and zero loadings on all other factors. This methodology will assist in validating the argument of this study regarding the extent to which test scores are influenced by the specific ability, or trait, and the method factor.

If the factor loadings on the trait factors are high, this would substantiate the claim that test performance is primarily governed by the language ability that the test is designed to measure. Conversely, if the

factor loadings on the test method factors are low, this would serve as evidence to reject the counterclaim that test performance is largely dictated by the method of testing. This approach provides a comprehensive way to assess the influence of both traits and methods on test performance, lending credibility to the overall validity of the test.

3.3.2 MFRM analysis

The MFRM software FACETS (version 3.83.6; Linacre, 2021) was employed to examine four facets – test-taker (ability), rater, rating category, and task – with two additional dummy facets – region and gender – used to explore test bias. Of these, the test-taker's ability emerged as the most influential facet, followed by the rater, the task, and the rating category. In the MARST under examination, raw scores were assigned on a scale of 1-4 for each of the three rating categories – Accuracy, Content, and Fluency.

The MFRM enables test developers to estimate the influence of each facet on the measurement process by gauging its difficulty or 'severity' (for example, the strictness of each rater). This difficulty estimate is then incorporated when calculating the probability of any given test-taker responding to any item, achieving any score category threshold, or being assessed by any rater (Bond & Fox, 2007).

The evaluation of test-takers' performance is carried out using the Rating Scale Model (RSM) (Andrich, 1978), which presumes the rating

scale operates similarly across all items. The RSM extends the binary Rasch model to accommodate items scored on a polytomous scale, meaning items that have multiple scoring categories. An item with 'k' possible score categories necessitates 'k−1' difficulty parameters, or thresholds, to distinguish between these score categories. In this study, the RSM is used to estimate three thresholds: one for achieving a score of 2 instead of 1, one for achieving a score of 3 instead of 2, and one for achieving a score of 4 instead of 3. Given that the RSM assumes the same step difficulty for all items, all tasks in this study are scored using the same number of score categories.

The assumption of unidimensionality should be met to the extent that data in language testing, which are generally considered as multidimensional constructs, still need to display adequate psychometric unidimensionality (Bonk & Ockey, 2003; Henning, 1992). Linacre (1998) finds it preferable to investigate multidimensionality in datasets by first conducting Rasch analysis and, then using PCA on the item standardized residuals since they are linear (as opposed to the raw scores), and yield more accurate estimates of the subsequent factors. The subsequent PCA indicates the structure of the underlying dimensions. Therefore, as long as the data fit the MFRM model to an acceptable extent based on the item fit values, unidimensionality is upheld.

Fit statistics are utilized to assess the alignment of data to the Rasch model. While the acceptability of fit is largely based on judgment, the infit and outfit mean square values serve as a practical benchmark as they are

known to be less sensitive to sample size and are weighted by the information in the response (Bonk & Ockey, 2003). This study employs infit and outfit mean square values as well, with the acceptable range being from 0.5 to 1.5.

3.3.3 Questionnaires and interview data

In addressing concerns related to test validation, specifically assessment interpretation and use, and in the development of a validity argument, language test researchers often employ a mixed methods approach that incorporates both quantitative and qualitative methodologies. Accordingly, this dissertation will utilize such an approach. Beyond the quantitative strategies outlined in the previous chapter, such as the MTMM and MFRM designs, this study will employ several qualitative techniques. These include online questionnaire surveys conducted immediately post-testing via the same mobile app, on-site interviews, and a discourse analysis of test-takers' speaking outcomes. This combination of methods will facilitate a comprehensive understanding of the process and outcomes of oral language testing.

How the survey questionnaires are presented appears in Figure 12. The test-takers' responses to the questionnaires are counted to provide descriptive statistics. A simplified version of the questionnaire is offered in Table 11 and the full version is available in Appendix 2. Also, interviews with a number of test-takers and teachers follow. The post-

test survey and interviews are mainly be concerned about their experiences of the MAR-based speaking test with respect to some features of interest : interactiveness, engagement, motivation and usefulness to back up validity arguments for the effectiveness and evaluation inferences.

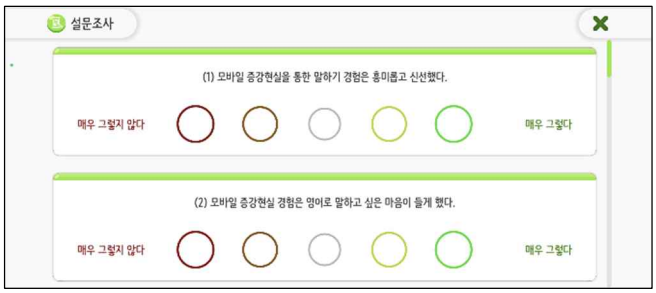


Figure 12. Questionnaires in the MAR app

Table 11. Post-test questionnaires

Category	MAR app			MAR-delivered test materials	
Item	Q1	Q2	Q3	Q4	Q5
Trait	Interest	motivation	comfort	authenticity	sufficient clue

Category	MAR-delivered test materials		MAR-delivered test		
Item	Q6	Q7	Q8	Q9	Q10
Trait	Interactiveness	clarity	relevance	appropriacy	usefulness

3.3.4 Speaking response data

In language testing, qualitative approach including verbal protocol, conversation and other discourse-based analysis (i.e., rhetorical, functional or linguistic analysis) of test language promotes research on test-taking processes and rater performances. Conversational analysis

is conducted to compare the conversational features of oral interview discourse with that of real-life interactions (Lazaraton, 2000), the features of group oral discussion tests (Leaper, 2014), raters performance and their qualitative differences in assigning scores (Cumming, 1997; Winke & Gass, 2013), and the comparability of oral interviews and semi-direct tests (Lazaraton, 2002).

The qualitative approach of the current study calls attention to the testing process, in particular, test-takers' oral talk in performing Task 3, a monologic task by discourse analysis. Some discourse-based analytic techniques includes linguistic analysis, functional analysis and structural analysis. As one of the qualitative research methods, discourse analysis involves examining the language used in speaking tests at various levels, from the micro-level of individual words and phrases to the macro-level of broader discourse patterns. This method can help researchers identify the linguistic features that characterize successful speaking performance and the criteria that raters use to evaluate test-takers.

In Task 3, a monologic task, test-takers were told to explain as an company employee to (simulated) rainforest rangers, who were prospective device users, how a device works, invented for protecting the rainforests by his/her company. What needed to be investigated here were test-takers' responses such as the identity of speaker in the

interaction and several linguistic resources that Young (2011) suggests such as the use of register[3].

These analyses intended to inform the understanding of the nature of test-takers' interaction with MAR-mediated speaking task, in other words, the effect of the MAR mode on test-takers' oral performance of a specific (monologic) type of task. Assuming that authentic interaction in language activities occurs when participants perceive social presence and engage with these activities, the analysis of the transcripts of audio-recorded speaking samples shed light on the nature of test-takers' oral talk in the MAR mode, which is known to afford social cues increasing a feeling of interacting with another social being, as explicated in 2.3.2; whether test-takers were sensitive to the context of the situation in Task 3 including the presence of the 'non-appearing but significant' simulated interlocutor.

To this end, audio-recordings of twenty test-takers in three test score groups – low, intermediate and high – were sampled and transcribed to search for discourse-based evidence that the test-takers' monologic talk demonstrated that the MAR mode facilitated test-takers' engagement and interaction with the context of the task. The evidence drawn from these qualitative methods will offer validation evidence in judging whether the MAR mode factor promotes or compromises score-based test interpretations and use.

[3] Young(2011) refers register to the use of pronunciation, vocabulary and syntax that characterize or are typical of the practice where interaction takes place.

Chapter 4. Results

This chapter reports the results of both quantitative and qualitative data analysis for the research questions of this dissertation. The former three sections of this chapter from 4.1 to 4.3 provide the results of the quantitative findings generated by statistical treatments of test score data, while the last section of 4.4 presents the results of the qualitative findings from analyzing survey and speaking response data.

4.1 Descriptive analysis

There is a summary of the descriptive statistics for the results of the four MARST tasks in Table 12. The intraclass correlation coefficient (ICC) was calculated to find out the inter-rater reliability. It is generally known as an indicator of the level of agreement between two or more raters rating a specific attribute including language ability. The ICC values range from −1 (perfect disagreement) to +1 (perfect agreement).

In this study, the ICC value was calculated from 90 test-takers' spoken responses to the four speaking tasks, Task 1, 2-1, 2-2, and 3, in which all the four raters altogether participated in rating according to the rating rubric. The rater group was set to be fixed, while test-takers' task scores remained randomly assigned. As illustrated in Table13, each task shows the four raters reaching significantly strong agreement throughout the overall tasks, along with Cronbach's alpha, the most

widely-known reliability index, being put in front.

Table 12. Item statistics for MARST tasks (N=194)

Task	Type	Min[4].	Median[5]	Max[6].	Mean	SD	Var[7].
1	Dialogue completion	2	16	20	14.39	5.52	30.42
2-1	Poster description	13	38	50	35.78	10.46	109.49
2-2	Opinion expression	13	38	50	37.21	10.82	117.14
3	Sequence explanation	13	38	50	34.90	11.85	140.43
	Sum[8]	41	124	170	122.27	34.65	1200.55

Table 13. Cronbach's alpha and ICC
for rater agreement of MARST tasks (N=194)

Task	Cronbach's alpha	ICC		
		ICC[9]	95% CI[10]	*P*
1	.96	.96	.94~.97	.000*
2-1	.95	.95	.93~.97	.000*
2-2	.96	.96	.93~.97	.000*
3	.96	.95	.93~.97	.000*
Sum	.84	.84	.80~.87	.000*

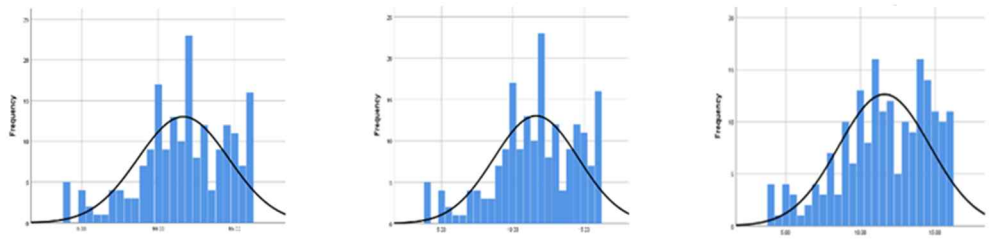
Regarding normal distribution, data is symmetrically distributed. The symmetrical distribution has zero skewness as all measures of a central tendency lies in the middle. According to the measures of skewness and kurtosis in Table 14 and Figure 13, the given score data set for each of the three rating categories – Accuracy, Fluency, and Content – was considered 'close to a normal distribution', using a less strict term. Skewness refers to a degree of asymmetry observed in a probability distribution that deviates from the symmetrical normal

[4] Min.: a minimum score of each task
[5] Median : a measure of central tendency that represents the middle value in a data set, separating the data into two equal halves
[6] Max.: a maximum score of each task
[7] Var.: a score variance of each task
[8] The perfect scores of the four tasks are 20, 50, 50, and 50 in order with 170 in sum
[9] The ICC value was calculated in the two-way mixed.
[10] CI: Confidence interval

distribution, or bell curve, in a given set of data. In all rating categories, the measures of skewness were negative near zero. This suggests that the mean scores were slightly greater than the medians, leading to the conclusion that the distribution was close to normal.

Table 14. Statistics for MARST rating criteria (N=194)

Category	Min	Median	Max	Mean	SD	Var	Skewness	Kurtosis
Accuracy	4	12	16	11.64	2.97	8.83	−.56	−.05
Fluency	5	13	16	11.51	2.80	7.82	−.69	.06
Content	4	12	16	11.61	3.06	9.37	−.57	−.34



[Accuracy] [Fluency] [Content]

Figure 13. Histograms of scores of three rating categories

4.1.1 Item analysis

Item difficulty index and item discrimination index for task scores for each rating category or dimension to be assessed – accuracy, fluency, and content – were examined. Score reliability and rater reliability were also evaluated according to classical test theory.

4.1.1.1 Item difficulty

Item difficulty (more accurately item easiness), in the context of

educational measurement and psychometrics, refers to the proportion of test takers who answer a particular item or question correctly. The higher the proportion of correct responses, the lower the item's difficulty level. This could also be interpreted as the item being easier, or that higher mean scores are inversely related to item difficulty.

According to Figure 14−a, among the mean sum scores colored in yellow, Task 1 was the easiest with the highest mean scores, followed by Tasks 2−1, 2−2, and 3 in ascending order of difficulty, with Task 3 being the most difficult. Figure 14−b showed that Task 1 appeared to be the easiest item in all rating categories as well. Task 3, on the other hand, proved the most difficult item in fluency and content although Task 2−2 the most difficult in accuracy.

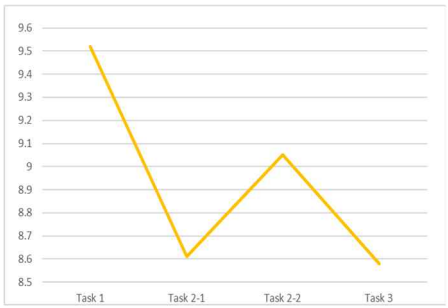


Figure 14−a. Item easiness (Sum)

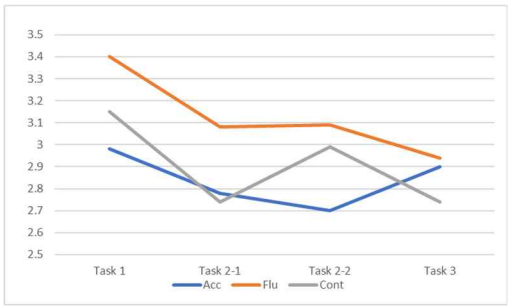


Figure 14−b. Item easiness (Categories)

It can be said that the overall pattern of item difficulty across the rating categories and sum scores is similar, with Task 1 being the easiest and Task 3 being the toughest. The only exception is the accuracy scores in Task 2−2, which proved the most difficult (2.70), and Task 3, which

became less difficult (2.90) although the score gap 0.2 was not large enough. The specific figures in table are provided in Appendix 3.

4.1.1.2 Item discrimination

Item discrimination refers to the degree to which a test item can differentiate between the test-takers who have high ability and those who have low ability. It gives an indication of how well a particular item contributes to the overall variation in test scores. In Figure 15, item-total correlations of each rating category to be assessed in four tasks are presented. Items can be identified that may not be effective at assessing the trait or ability the test is intended to measure. An item with poor discrimination might be too easy (everyone gets it right) or too difficult (everyone gets it wrong), or it might be confusing or irrelevant to the construct being measured. The specific values of item discrimination are provided in Appendix 4.

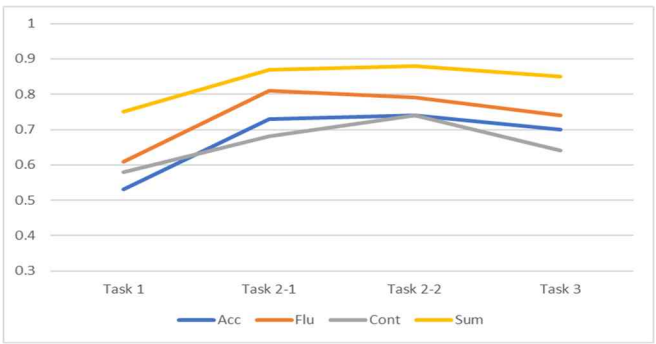


Figure 15. Item discrimination (Item-total correlation)

The Item Discrimination Index (Suppiah Shanmugam, Wong, & Rajoo, 2020) classified items with values of 0.4 and above as 'very good' and 0.3 to 0.39 as 'reasonably good' but subject to improvement. Items with values between 0.2 to 0.29 are usually subjected to revision and items below 0.19 are 'poor'. The range of discrimination index of three rating categories and sum scores in four tasks appears between 0.53 and 0.88, which means scores on each dimension and sum scores are 'fairly discriminating'.

4.1.2 Inter-rater reliability

Assessing the consistency between different raters, known as inter-rater reliability, is crucial to determine the probability of a measurement instrument incorrectly assessing a dimension. The kappa statistic, or Cohen's kappa coefficient, is a measure of inter-rater reliability, which indicates the level of agreement between two or more raters beyond what would be expected by chance.

The value of Cohen's kappa ranges from 0 to 1. The closer the coefficient is to 1, the stronger the inter-rater agreement. According to Landis and Koch (1977), a general guideline suggests that values between 0.41 and 0.60 are considered of moderate agreement, those between 0.61 and 0.80 show substantial agreement, and lastly, those above 0.81 are considered almost perfect to perfect agreement.

In Figure 16, it was discovered that the specific values of interrater

reliability in table are provided in Appendix 5. Overall, in the sum scores colored in yellow, two raters exhibited a similar degree of agreement for Tasks 2, 3, and 4, with the least similarity observed in Task 1. A similar rating pattern to the sum scores was found in the fluency category. However, there was a certain degree of disagreement between raters in the rating patterns of accuracy and content categories. This disagreement that led to barely moderate agreement between two raters according to the general guideline mentioned above was most prominent in Task 2-1 with 0.39 in the content category and Task 3 with 0.41 in the accuracy category. Further investigation into this disagreement is warranted as well.

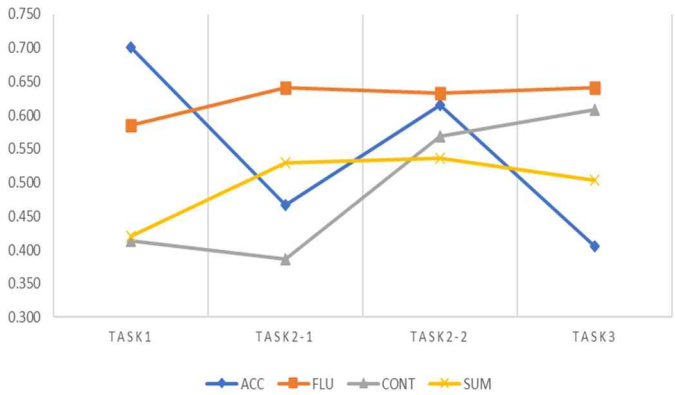


Figure 16. Measure of agreement (Cohen's Kappa coefficient, p = .000)

The Spearman-Brown prophecy formula is often used by researchers to estimate expected reliability in relation to test length. It offers a useful tool for examining the transition in reliability according to changes in the number of items so that we can find the number of items

that guarantees optimal reliability.

4.1.3 Score reliability

In classical test theory, score reliability refers to the consistency or repeatability of scores obtained from a test. If a test is reliable, it means that the same individuals should get roughly the same scores each time they take the test (assuming they are in the same mental and physical condition, and no learning or forgetting has taken place between tests).

Among some commonly used indicators of score reliability, internal consistency reliability including Cronbach's alpha indicates the extent to which items on a test measure the same concept or construct. The value of Cronbach's Alpha can range from 0 to 1, with values closer to 1 indicating higher internal consistency or reliability (See Tables 15 to 17). The specific values of score reliability for different test lengths are provided in Appendix 6. The values above 0.7 are generally considered appropriate though what constitutes a "good" value can depend on the nature of the test and the context in which it is used.

The Cronbach's alpha indices above 0.8 in Tables 15 to 17 are likely to indicate that scores given to test-takers across the rating categories seemed reliable. In reliability analysis, however, Cronbach's alpha when an item were deleted from a group of items often reports whether the internal consistency (as measured by Cronbach's alpha) would increase if a particular item were deleted. If the alpha increases when an item is deleted,

it indicates that the item is not well correlated with the other items and could be removed or modified to improve the overall reliability of the scale.

Such cases were found in the accuracy category for Task 1, where Cronbach's alpha would increase from 0.837 to 0.860 if an item were deleted. Also, in the fluency category for Task 1, Cronbach's alpha would actually increase from 0.877 to 0.886 if the same item were deleted. It is suspected that Task 1 might not have been as effective in assessing the same dimension as the other tasks. Further investigation into this suspicion is warranted.

Table 15. Item-total statistics [11] (Accuracy)

N=194	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
ACC1[12]	8.660	5.495	0.53	0.291	0.860
ACC2-1	8.853	5.298	0.73	0.582	0.768
ACC2-2	8.668	5.136	0.74	0.624	0.764
ACC3	8.729	5.021	0.70	0.518	0.780
Cronbach's alpha	.837				

Table 16. Item-total statistics (Fluency)

N=194	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
FLU1	9.111	5.253	0.61	0.390	0.886
FLU2-1	9.425	4.278	0.81	0.670	0.813
FLU2-2	9.415	4.383	0.79	0.662	0.821
FLU3	9.572	4.402	0.74	0.560	0.841
Cronbach's alpha	.877				

[11] These scores are the average of the scores graded by two raters.

[12] The number attached to the rating category refers to the task number.

Table 17. Item-total statistics (Content)

N=194	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
CONT1	8.464	6.043	0.58	0.338	0.812
CONT2-1	8.869	5.759	0.68	0.515	0.769
CONT2-2	8.626	5.579	0.74	0.579	0.743
CONT3	8.874	4.923	0.64	0.416	0.800
Cronbach's alpha	.826				

Table 18. Item-total statistics (Sum scores)

N=194	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
SUM1	61.990	221.681	0.75	1.000	0.811
SUM2-1	62.902	214.545	0.87	1.000	0.794
SUM2-2	62.464	211.895	0.88	1.000	0.789
SUM3	62.930	207.209	0.85	1.000	0.785
Cronbach's alpha	.836				

As illustrated in Figure 17, we observed an increase in reliability for tasks that measure four dimensions – including three rating categories and what is presumably represented by the sum scores – as the number of test items increases. The two curves for accuracy and sum scores were almost overlapped, which may mean that the criteria associated with linguistic accuracy had greater impact on the overall scores. The analysis also showed that the reliabilities dramatically increased, forming upward curves, until the number of tasks reached four, which is the current number of tasks for MARST in this study. In addition, the reliabilities of the rating categories tended to increase as the number of

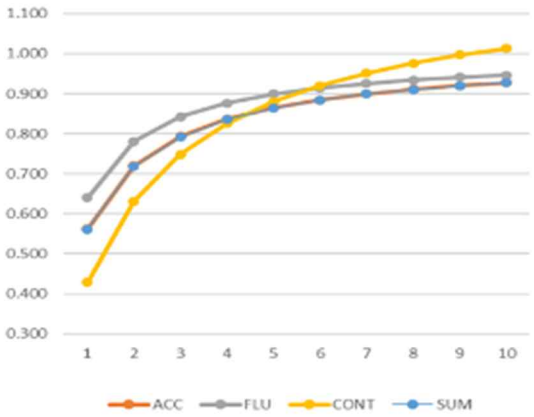


Figure 17. Predicted reliability (Spearman–Brown prophecy formula)

items increased although the degree of increase was slight after the number of items reached five. However, it appears that the reliability of the Content category continued to increase up to the maximum number of items assumed here, suggesting that this category was less directly related to language skills and remained variable. The specific figures in the table are provided in Appendix 5. The Content category, which exhibited scoring variability, necessitated additional investigation.

4.2 MTMM analysis and test comparability

To claim that a measure has construct validity, both convergent and discriminant validation must be assessed. Convergent validity refers to how closely a test is related to other tests that measure the same (or similar) constructs. On the other hand, discriminant validity refers to refers to the extent to which a test is not related to other tests that measure different constructs.

Related to these two types of validity, a multitrait-multimethod (MTMM) model assumes that correlations among the same ability with different methods (i.e., monotrait-heteromethod) which demonstrates the evidence of convergent validity would be higher than correlations among measures of different traits with the same method, or in other words, heterotrait-monomethod correlations. The latter is associated with discriminant validity.

To test this hypothesis, the data was analyzed using IBM SPSS statistics (Version 26). The size of the various types of correlations was compared to determine the extent to which the hypothesized order is observed. A matrix consisted of three types of correlations. As evidence of convergent validity, monotrait-heteromethod measures were expected to be strongly correlated, higher than heterotrait-monomethod and heterotrait-heteromethod measures. If correlations among heterotrait-monomethod measures were high, there is likely to be a strong method factor.

4.2.1 Correlation matrix

Let us begin with the monotrait-heteromethod matrices, which are called validity diagonals and colored in green in Table 20. Among the same speaking tests labelled with the letter A, the MAR-based speaking test scores(A3_M3) had moderate positive correlations with face-to-face oral interviews(A4_M4) and oral translations(A2_M2) with .605 and .573

respectively, while it had a weak correlation with indirect multiple-choice type speaking questions[13] (.347). Meanwhile, face-to- face oral interview[14] (A2_M4) and oral translations[15] (A2_M2) exhibited the highest correlation .680.

Table 19. Correlation matrix[16] (p =.001)

Method	Multiple choice (M1)				Translation (M2)		MAR (M3)	Interview (M4)
Trait	speaking (A1)	listening (B1)	reading (C1)	writing[17] (D1)	speaking (A2)	writing (D2)	speaking (A3)	speaking (A4)
M1(A1)		.720	.645	.485	.473	.490	.347	.514
M1(B1)	.720		.745	.463	.611	.556	.488	.696
M1(C1)	.645	.745		.559	.575	.605	.419	.647
M1(D1)	.485	.463	.559		.375	.419	.378	.449
M2(A2)	.473	.611	.575	.375		.656	.573	.680
M2(D2)	.490	.556	.605	.419	.656		.531	.588
M3(A3)	.347	.488	.419	.378	.573	.531		.605
M4(A4)	.514	.696	.647	.449	.680	.588	.605	

In the yellow part or the heterotrait-monomethod matrix excluding the MAR measure, it was found among multiple-choice testing. Speaking and writing translation measures, correlated with measures that speaking had positive correlations with listening, reading and writing in order of size, .720, .645, and .485, respectively. This revealed the somewhat

[13] It consists of a few items that are part of the listening section in the mocking Korean SAT, which is designed to indirectly assess speaking skills in October 2020.

[14] It is a section of the speaking test for school performance assessment, administered in Fall 2020. Teacher worked as both an interviewer and an examiner.

[15] It is a section of the speaking test for school performance assessment, administered in Spring 2020. Students were asked to translate spoken Korean sentences into their corresponding English ones.

[16] Yellow: heterotrait-monomethod
Green: monotrait-heteromethod (validity diagonals)
Orange: heterotrait-heteromethod

[17] It consists of a few of items as part of the reading section in the mocking Korean SAT that presumably are designed to indirectly assess writing skill. It was administered in October, 2020.

strong method factor effect of multiple-choice each other (.656) to a considerable extent.

In the orange part, indicating the heterotrait-heteromethod matrix, the MAR speaking measure had moderate correlations with translation writing (.531), multiple choice listening(.488), multiple-choice reading (.419), and multiple-choice type (.378). On average, the MAR speaking test shows slightly higher correlations in the monotrait-heteromethod matrix (.508) than in the heterotrait-heteromethod one (.454).

4.2.2 Factor analysis

To begin with, Table 20 summarizes the descriptive statistics of the scores of four tasks in the MARST. The analysis points to the correlations among the MARST tasks, which are briefly presented in Table 21. The poster description task (Task 2-1) and the opinion expression task (Task 2-2) showed the highest correlation (r =.83), which might be due to the same AR-based animated poster input illustrating the global greenhouse effect that the two tasks share. In addition, the tasks overall had moderate or strong correlations with one another ranging from .61 to .83, as shown in Table 21.

Table 20. Descriptive statistics of the MARST tasks

Task (N=200)	M	SD
dialogue completion	14.39	5.52
poster description	35.78	10.46
poster opinion	37.21	10.82
sequence explanation	34.90	11.85

Table 21. Correlations of the MARST tasks

	dialogue completion	poster description	opinion expression	sequence explanation
dialogue completion	1.00			
poster description	.63	1.00		
opinion expression	.64	.83	1.00	
sequence explanation	.61	.76	.77	1.00

The following factor analysis[18] finds a single factor extracted, accounting for over 78% of the total variance. In Table 22, the KMO (Kaiser-Meyer-Olkin) value of .838, an indicator of adequacy of factor analysis, indicates the variables chosen were proper for factor analysis as it was higher than .60. Bartlett's test was used to test that variances were equal for all samples. As statistically significant Bartlett's test value (p =.000) suggests the homogeneity of variances, which means multiple samples were from populations with the same variances. The Bartlett test is known to be sensitive to departure from normality. If the measure of sampling adequacy is larger than 0.5, it means that there is sufficient variance in the data that can be partitioned by using factor analysis. In Table 23 and 24, the MARST proved unidimensional with each of the four tasks substantially contributing to approximately 78% of the score variances (see Component 1 in Table 24) of test-takers' speaking ability. The scree plot in Figure 18 showed a dominant single factor with an eigenvalue of 3.1, which was sufficiently higher than 1.0, the main

[18] *Extraction method: Principal component analysis(PCA)
*Rotation method: Varimax with Kaiser normalization.
*Rotation converged in three iterations.

criterion that determined a substantial factor with all the task score variances converged into the Component 1, as illustrated in Table 24. In addition, this result upholds the fundamental assumption of unidimensionality to undertake MFRM analysis. Meanwhile, there was an attempt to extract common factors underlying various speaking measures, as illustrated in Tables 27 to 31.

Table 22. KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.838
Bartlett's Test of Sphericy	Approx. Chi-Square	551.158
	df	6
	Sig	.000

Table 23. Communalities

Task	Initial	Extraction
AR dialogue completion	1.000	.650
AR poster description	1.000	.836
AR opinion expression	1.000	.850
AR sequence explanation	1.000	.794

Table 24. Variance explained

Comp	Initial Eigenvalues			Extraction sums		
	Total	% of Var	Cum %	Total	% of Var	Cum %
1	3.130	78.241	78.241	3.130	78.241	78.241
2	.442	11.054	89.295			
3	.258	6.442	95.737			
4	.171	4.263	100.000			

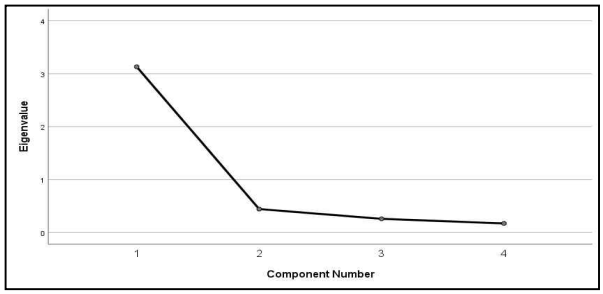


Figure 18. Scree plot (Four MARST tasks)

Table 25. Component matrix

	1
AR opinion expression	.922
AR poster description	.914
AR sequence explanation	.891
AR dialogue completion	.806

With the descriptive statistics of different speaking measures in Table 26, the factor analysis led to two components extracted with variances extracted from all the measures to a different extent, as indicated in Table 25 and in Figure 19. In Table 28, the KMO(Kaiser-Meyer-Olkin) value of .866, an indicator of adequacy of factor analysis, indicated the variables chosen were proper for factor analysis as it was higher than .60. The factor analysis proved that the score variances from all sorts of speaking measures were used, as indicated in Table 29. The two factors extracted from them substantially contributed to approximately 72% of the score variances (see Component 1 and 2 in Table 30 and the scree plot in Figure 19) of test-takers' speaking ability.

According to Table 30, the loadings of Component 1 were found in the range of .851 to .545, with particularly high loadings in AR-delivered tasks

and face-to-face interview and translation. The loadings of Component 2 varied depending on the test method ; the highest positive loadings were identified in the multiple-choice task (.683), weak positive loadings were in the two face-to-face tasks (.325, .347) and weak negative loadings were found in the four AR- delivered tasks (-.396 ~ -.172).

Table 26. Descriptive statistics of various speaking measures

(N=145)	M	SD
Oral translation	77.59	21.16
Oral interview	78.81	15.91
AR dialogue completion	15.43	4.81
AR poster description	38.02	9.00
AR opinion expression	39.59	9.54
AR sequence explanation	38.17	10.47
Multiple-choice(MC) speaking	6.42	3.97

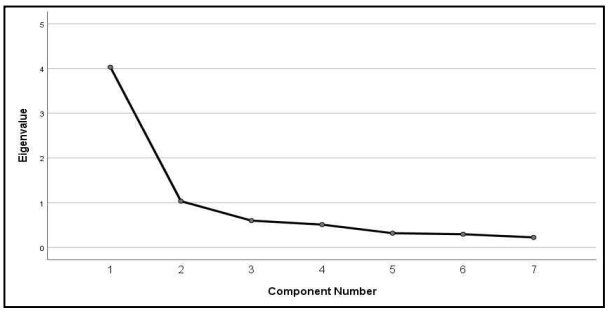


Figure 19. Scree plot (Various speaking measures)

Table 27. Correlations of various speaking measures

	Oral translation	Oral interview	AR dialogue	AR description	AR opinion	AR sequence	MC speaking
Oral translation	1.000						
Oral interview	.680	1.000					
AR dialogue	.396	.455	1.000				
AR description	.427	.480	.463	1.000			
AR opinion	.535	.550	.528	.763	1.000		
AR sequence	.534	.562	.523	.703	.717	1.000	
MC speaking	.473	.514	.269	.247	.314	.282	1.000

Table 28. KMO and Bartlett's test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.866
Bartlett's Test of Sphericy	Approx. Chi-Square	514.394
	df	21
	Sig	.000

Table 29. Communalities

	Initial	Extraction
Oral translation	1.000	.701
Oral interview	1.000	.746
AR dialogue completion	1.000	.494
AR visual description	1.000	.782
AR opinion expression	1.000	.804
AR sequence explanation	1.000	.773
Multiple choice speaking	1.000	.764

Table 30. Variance explained

Comp	Initial Eigenvalues			Extraction sums			Rotation sums[19]
	Total	% of Var	Cum %	Total	% of Var	Cum %	Total
1	4.031	57.591	57.591	4.031	57.591	57.591	3.667
2	1.032	14.746	72.337	1.032	14.746	72.337	2.698
3	.598	8.538	80.874				
4	.508	7.260	88.135				
5	.316	4.516	92.650				
6	.293	4.185	96.836				
7	.222	3.164	100.000				

Table 31. Component matrix[20]

	Component	
	1	2
AR poster description	.791	-.396
AR opinion expression	.851	-.281
AR sequence explanation	.836	-.273
AR dialogue completion	.681	-.172
Multiple choice speaking	.545	.683
Oral interview	.800	.325
Oral translation	.762	.347

[19] When components are correlated(.466), the sums of square loadings cannot be added to obtain a total variance.

[20] Extraction method : Principal component analysis(PCA)
Rotation method : Oblimin with Kaiser normalization

It seemed that the Component 1 was explained by all the speaking measures regardless of their methods, which means that it could be associated with the speaking trait, a common crucial interest to all measures. Component 2 seemed to divide the speaking measures by its test method, which was probably most strongly associated with the multiple-choice type of speaking measures (.683).

Thus, further investigation should be called for to understand whether or to what extent test method affects a newly developed speaking test by means of a sophisticated measurement model so that complicated factors involved with speaking performance assessment can be taken into account for test validation research.

4.3 MFRM analysis

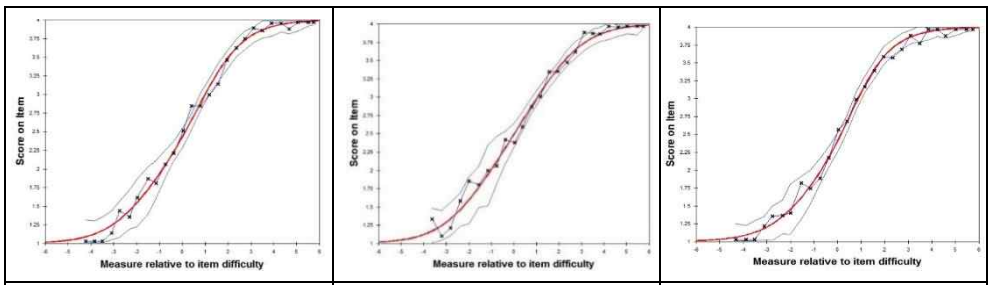
According to Fulcher (2003), MFRM (Linacre, 1989) treats items, persons and raters as "facets" of a testing situation. Each facet is given a value on the same linear scale, representing item difficulty, person ability, and rater harshness. In this study, there are four major facets: examinee, task, rater and category with region and gender as dummy variables used exclusively to figure out the interaction effect, or, in other words, perform a bias analysis.

4.3.1 Fit statistics

Approximately 53.11% of the total score variance was explained by

the Rasch model with the remaining 46.89% of residuals. Figure 20 shows that most observations lie within the confidence interval(CI). An overview of the results of the MFRM analysis is illustrated in Figure 21, which includes plotting estimates of test-takers' abilities, rater severity, and rating criteria and the scale step difficulty on the same logit scale for comparison.

The MFRM analysis viewed each rating as a function of the interaction of test-taker ability, task difficulty, criterion(and scale step) difficulty, and rater severity/ leniency (McNamara, 1996). The far-left measure acted as a ruler against which each of the four facets (test-taker, rater, task, scale criterion) as well as scale level difficulty, measured in "logit" units. A logit above zero on the ruler indicates that a test-taker is more able, a task or criterion is more difficult, and a rater is more severe, whereas a logit below zero indicates the opposite.



Accuracy	Fluency	Content

Figure 20. Item characteristic curve(ICC) of test scores and 95% CI[21]

In addition, information about the reliability of each of these estimates (e.g., standard error or SE) was provided. An SE indicates the

[21] CI: Confidence interval

uncertainty of the parameter estimate. The separation index indicates the number of levels within a given facet. Meanwhile, the reliability of separation indicates the degree to which the analysis reliably distinguishes between different levels within a given facet.

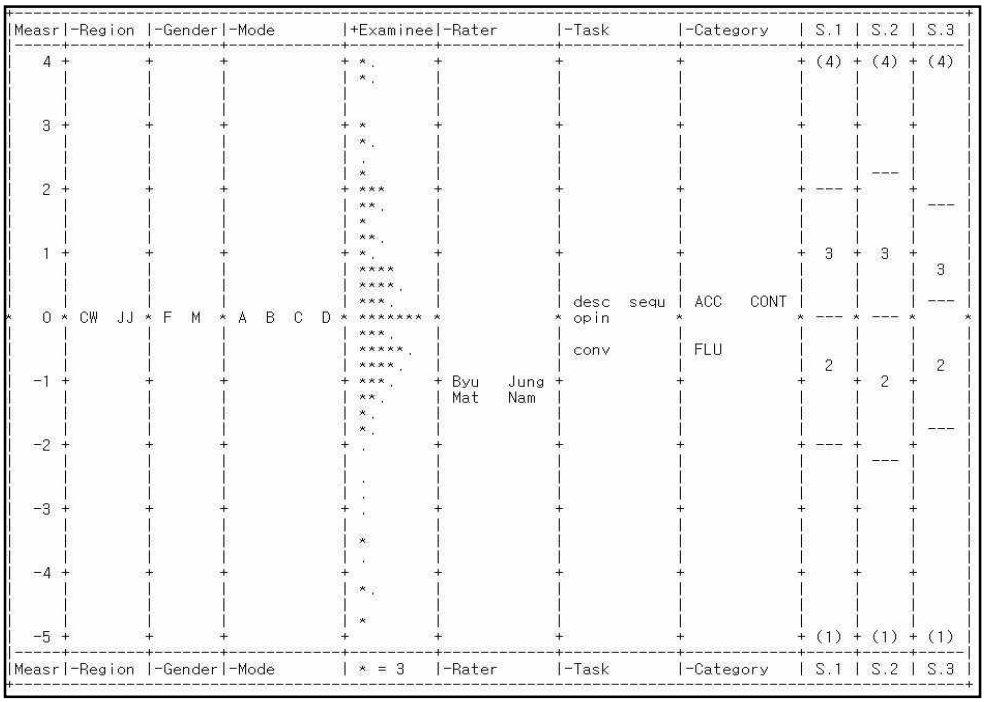


Figure 21. All facet vertical rulers[22]

In Rasch analysis, fit statistics are useful tools for judging the fit of data to the Rasch model. Infits refer to the measurement of the item's fit to the Rasch model based on the individual response patterns. It specifically examines how well an item fits within a particular range of ability levels.

[22] S.1: accuracy; S.2: fluency; S.3:content

The infit statistic considers the pattern of responses near the person's ability level and provides information about the item's performance in relation to the model's expectations. The outfit measure assesses the item's fit across the entire ability range, considering the response patterns across a broader spectrum of ability levels. Thus, they provide information about the item's performance throughout the entire measurement scale.

Infit evaluates the suitability (item fit) of an item in the Rasch model based on the response pattern of an individual test-taker. It is sensitive to the information provided by the normal values, thus responding sensitively to normal values (inlier-sensitive fit). Outfit evaluates the fit of a specific item across the range of abilities of all test-takers. It is more unstable compared to infit, as it responds sensitively to outliers (outlier-sensitive fit). Both measures range from 0.5 to 1.5, with a value close to 1 indicating good fit, meaning that the model expectations match the observed values. Values below 0.5 indicate overfitting (too predictable, redundant), while values above 1.5 are considered misfit, indicating data that does not fit the Rasch model (unmodeled noise).

For fit analysis, infit mean square(MnSq) values in the range of 0.5 to 1.5 logit are suggested as a "productive and reasonable measurement" to judge for goodness of fit statistics (Wright and Linacre, 1994; Linacre,2006; Kim, Park, & Seol, 2009). Infit MnSq values show the degree of variability in a facet relative to the amount of variability in the entire facet set. In contrast, outfit MnSq values are sensitive to outliers

(Linacre, 2004). The acceptable range for both the infit and outfit MnSq values is 0.5 to 1.5. The closer the fit statistics are to an expected value of 1, the better the assessment.

A value larger than 1.5 indicates misfit while lower than 0.5 indicates overfit. The misfit data suggests the degree of inconsistency in a score pattern. McNamara (1996) advises that any test development should aim to have misfitting students at or below an incidence of 2%. Thus, fit statistics for each test-taker provide information on the validity of the assessment (Bond and Fox, 2007) as both misfit and overfit suggest test-taker ability is not being properly measured by the test.

4.3.1.1 Test-taker (ability) facet

The reliability index was as high as .95 and the separation index is 4.32. Test takers were able to be separated into 6.09 statistically separate strata. The ability to separate them into statistically distinct strata was important for the MARST, as the result showed it could function well enough to discriminate between test takers of different ability levels. The test-takers' ability measures were widely spread from −4.75 to 4.99 logits with a fair average of 3.02 and an adjusted SD of 1.61. The mean of the fixed (all same) Chi-square test was statistically significant ($\chi^2(193) = 2794.4$, $p = .00$), which means the degree of test takers' ability was not the same. The separation and reliability indices for the difference in test-taker ability were high (4.32,

0.95). This means the MARST separated the 194 test-takers into at least four statistically distinct levels or strata in terms of the ability being measured. Table 32 summarizes the test-taker fit statistics.

Table 32. Summary of test-taker facet statistics

Test-taker Ability Estimates(N=194)	
M (Model SE)	.13(.39)
SD[23] (Model SE[24])	1.81(.27)
Min	-4.75
Max	4.99
Infit	
M	1.04
SD	.37
Outfit	
M	1.01
SD	.37
RMSE	.47
Adj (True) S.D.	1.75
Separation Statistics	
Separation	3.71
Strata	5.27
Reliability of Separation	.93
Fixed Chi-Square Statistics	2794.4
df.	193
Significance	.00

Moreover, in Table 33, fit statistics for each test-taker provide information about the validity of the assessment. Acceptable fit indices indicate a pattern of ratings that closely approximates the predicted Rasch-model rating pattern based on the test-taker ability estimate (McNamara, 1996). Usually, misfit is considered to be a more serious problem than overfit (Bond and Fox, 2007; McNamara, 1996).

Of the 194 test-takers analyzed, 12 test-takers were identified as misfitting. The percentage of misfitting test takers in the dataset was

[23] SD refers to the spread of scores between test-takers.
[24] SE refers to the spread of estimates for a test-taker.

6.1%, which did not satisfy McNamara's(1996) guideline for test development recommending that the percentage of misfitting students be at or below 2%. These cases along with 32 unexpected responses are presented in Appendices 3 and 4, both of which are required to deal with later in the qualitative analysis of spoken response data.

Table 33. Frequencies (%) of test-taker fit mean square statistics(N=194)

Fit Range	Infit MS	Outfit MS
Overfit: Fit<.50	4 (2%)	2 (1%)
Acceptable: 0.50<fit<1.50	168 (87%)	173 (89%)
Misfit: Fit>1.50	22 (11%)	19 (10%)

4.3.1.2 Item (Task) facet

Both the infit and outfit MnSq values of task facet reported in Table 34 were within the productive range of 0.5 to 1.5 for measurement, thus producing no overfit or misfit with the mean of standard errors of measurement index at 0.05, which was better as it approaches zero. No misfitting or overfitting task indicated there was little chance of tasks being poorly made or perfectly good, which might suggest no task redundancy or need for revision or removal. It can be said that each task forms part of a set of tasks that together define a single measurement trait (McNamara, 1996).

Tasks also provide unique information that the other tasks do not give since there are no overfitting tasks. The four tasks differed significantly in terms of their difficulty as indicated by the high reliability and separation indices and the fixed (all same) Chi-square statistics

($\chi^2(193) = 179.7$, $p = .00$). The tasks were separated into 6 to 7 levels of difficulty. Task 4, which asked the test-takers to explain procedural information was the most difficult (0.29 logit). Task 2-1 describing an environmental poster was the second most difficult (0.26 logit). Task 2-2, the third difficult task, was expressing an opinion about the poster(-0.08 logit) and the easiest one was Task 1: completing a conversation(-0.47 logit).

Table 34. Task measurement report

N=194	Measure	Model S E	Infit MS	Outfit MS
Task 1	-.47	.05	1.34	1.20
Task 2-1	-.08	.05	.74	.78
Task 2-2	.26	.04	.88	.93
Task 3	.29	.04	1.15	1.12
M	.00	.05	1.03	1.01
SD	.36	.00	.27	.19

RMSE .05 Adj (True) S.D. .35 Separation 7.78 Strata 10.71 Reliability .98
Fixed (all same) chi-square: 179.7 *d.f.*: 3 significance(probability): .00

The high task reliability index (0.98) indicated the degree to which tasks were replicable in terms of difficulty was sufficiently high that they could be assigned to another sample population with comparable ability levels. Combined with the high separation index, the tasks were separated into different levels of difficulty. For example, Tasks 3 and 2-2 were more difficult than Task 1 or 2-1 with another sample of test-takers.

4.3.1.3 Rater facet

The results for the rater facet are summarized in Table 35. The

random (normal) Chi-square test proved nonsignificant($\chi^2(2)=2.7$, $p=.26$), which means all raters were chosen randomly. Moreover, the degree of rater severity was significantly different according to the significant fixed (all same) Chi-square test result ($\chi^2(3)=29.4$, $p=.00$) with the reliability index of 0.86 being lower than in other facets. The separation index of 2.46 suggested that the model distinguished raters' rating performance into at least two levels.

Fit statistics on raters showed their internal consistency across test-takers, criteria, and tasks to distinguish test-taker's performance. Table 30 showed quite a good fit of the four raters, indicating that they performed with a satisfactory degree of consistency, given the infit MnSq range of 0.80 to 1.06 and the outfit MnSq range of 0.80 to 1.05, as shown in Table 35. Both infit and outfit MnSq values were all within the productive range of 0.5 to 1.5 for measurement, thus producing no overfit or misfit.

The low reliability and separation indices and the non-significant χ^2 statistic indicated that the raters were equal in severity. Reliability here means the degree to which raters are reliably separated into different levels of severity. Therefore, low reliability and low separation indices are desirable as they indicates that raters are interchangeable (McNamara, 1996; Weigle, 1999).

The largest severity difference value was between Rater 3 and Rater 4(0.36), but the severity of three out of four raters – Raters 1,2 and 4- was extremely close to one another. The average rating

difference between the severest rater (Rater 4) and the most lenient rater (Rater 3) was 0.19 of a band, no greater than 0.2.

In Table 35, the inter-rater reliability[25] indicates the gap between exact agreement expected by Rasch model and the exact agreement observed, reporting both the observed and expected percentages of the exact rater agreement. The inter-rater reliability is the same as the one explained by the Rasch model, while the + value means the observed inter-rater agreement is higher than the one explained by the Rasch model, in other words, 'overfit'. In that case, a rater does not perform ratings independently(Linacre, 1990).

Table 35. Rater measurement report[26]

	Measure	Model S E	Infit MS	Outfit MS	Corr. PtBis
Rater 1	−1.01	.05	1.06	1.05	.27
Rater 2	−1.14	.06	1.05	1.04	.26
Rater 3	−1.32	.06	.80	.80	.29
Rater 4	−.96	.03	1.05	1.03	.30
M	−1.11	.05	.99	.98	.28
SD	.14	.01	.13	.12	.02

RMSE .04 Adj (True) S.D. .05 Separation 2.89 Strata 4.19 Reliability(not inter-rater) .89
Fixed (all same) chi-square: 29.4 *d.f.*: 3 significance(probability): .00
Inter-Rater agreement opportunities: 2328 Exact agreements: 1640=70.4% Expected:1128.9=48.5%

In order of magnitude, the inter-rater reliabilities of four raters were 0.28 logit (Rater 1), 0.43 logit (Rater 4), 0.52 logit (Rater 3), and 0.56 logit (Rater 2). These all positive indices of the inter-rater reliabilities of the MARST meant that observed inter-rater agreement was consistently higher than expected by the Rasch model.

[25] Inter-rater agreement(logit) =(exact agreement-expected agreement)/(100-expected agreement)
[26] Rater 1: Jung, Rater 2: Nam, Rater 3: Matt (a native rater), Rater 4: Byun

In addition, the mean of standard errors of measurement, which was better when near to zero, was 0.05. The point-biserial correlation for each rater refers to the degree to which the rater's ratings correspond to the total ratings of all other raters of the same speaking sample.

4.3.1.4 Category (criterion) facet

The analysis allows for the estimation of rating criteria difficulty and are summarized in Tables 36 to 38. The fit statistics of the three rating criteria are within the acceptable range of 0.5 to 1.5. None of the rating criteria proved to be misfitting or overfitting. If there had been a misfit, a criterion would have not formed part of the same dimension as other criteria defined in the rating scale. In this case, it would have been assumed to measure a different trait construct.

This was an encouraging result as the assumption of unidimensionality holds for this data (Bonk and Ockey, 2003), which means that the separate analytic rating scales seem to be contributing to a common construct of 'speaking ability'. If there is overfit, on the other hand, it is highly likely that a criterion is redundant or measures the same ability as other criteria, which would affect the scores assigned to other criteria, in what is known as the halo effect (Eckes, 2005; McNamara, 1996).

Table 36. Category (Accuracy) scale statistics

Score	Observed Counts		Average Measure	Expected Measure	Outfit MS	Step Calibration	
Level	Freq.	%				Measure	SE
1	137	9	−2.08	−1.93	1.0		
2	396	26	.20	.04	1.1	−1.89	.12
3	490	32	.86	1.00	.9	.33	.07
4	529	32	2.25	2.20	.9	1.56	.07

Table 37. Category (Fluency) scale statistics

Score	Observed Counts		Average Measure	Expected Measure	Outfit MS	Step Calibration	
Level	Freq.	%				Measure	SE
1	82	5	−2.20	−1.92	.7		
2	261	17	.46	.21	1.1	−2.00	.15
3	587	39	1.20	1.38	1.0	.04	.08
4	622	38	2.73	2.62	.9	1.96	.07

Table 38. Category (Content) scale statistics

Score	Observed Counts		Average Measure	Expected Measure	Outfit MS	Step Calibration	
Level	Freq.	%				Measure	SE
1	171	11	−1.87	−1.79	.9		
2	395	26	.15	.03	1.2	−1.58	.11
3	400	26	.88	.90	1.1	.46	.07
4	586	36	2.00	2.05	1.1	1.12	.07

According to Table 39, test-takers seemed to have the most difficulty in achieving high scores on Content (0.26 logit), while Fluency was the easiest area for them to obtain high scores (−0.43 logit). Given that the three rating categories exhibited significantly different degrees of difficulty ($\chi^2(2)=169.4$, $p=.00$), the analysis showed that the rating criteria were distinguished into 7 to 9 distinct levels of difficulty with a high reliability index of 0.98.

Table 39. Rating category measurement report

	Measure	Model S E	Infit MS	Outfit MS
Fluency	−.43	.04	.95	.94
Accuracy	.17	.04	1.02	.99
Content	.26	.04	1.07	1.08
M	.00	.04	1.01	1.01
SD	.38	.00	.06	.07

RMSE .04 Adj (True) S.D. .37 Separation 9.45 Strata 12.94 Reliability(not inter−rater) .99
Fixed (all same) chi−square: 169.4 *d.f.*: 2 significance(probability): .00

The results indicated that test−takers performed significantly differently in the various aspects of speaking and the raters perceived these rating criteria differently. In addition, the mean of standard errors of measurement, which was better when nearer to zero, was 0.04.

As far as rating scale functioning is concerned, assessing the quality of the rating scale addresses how well the scale levels function in estimating the construct being measured and whether the thresholds take a hierarchical pattern to the rating scale. Figures 22 to 24 illustrate the results of scale analysis across three rating categories. The first section presents the scale levels from 1 to 4. In terms of the frequency and percentage of counts of a given score assigned across all raters and speaking samples, it was clear that the raters used scores 3 and 4 the most frequently across the three rating criteria.

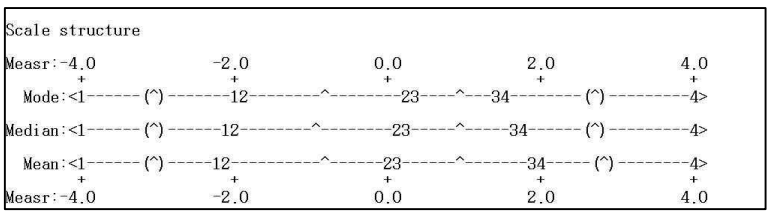


Figure 22. Category(Accuracy) scale structure

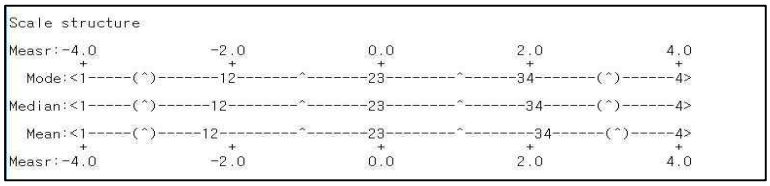


Figure 23. Category(Fluency) scale structure

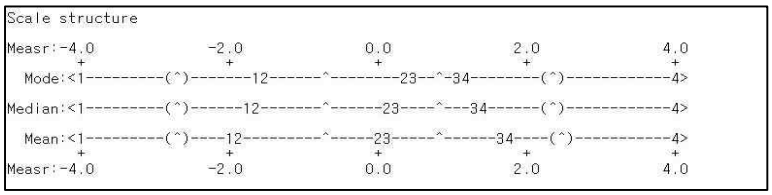


Figure 24. Category (Content) scale structure

The second section began with the average test-takers' ability measures of each score level. These measures were expected to increase monotonically in size, indicating that, on average, the higher the score was, the higher the ability of test-takers. The analysis shows the average measures increase as the score level progresses across the rating categories. The same pattern is true of the expected measure for each score level. The expected measure indicates the test-taker ability measure that the Rasch model would predict if the data were to fit the model. Therefore, it can be inferred that the scale was positively linked to progression of test-taker's speaking ability and the rating scale functions as expected.

The next analysis result is the outfit mean square(MnSq) index for each scale level with its expected value of 1.0. indicating equal observed and expected test-taker ability measures. The outfit MnSq range is known to determine whether the scale levels are reliable based on the

model value of 1.0. The outfit MnSq of this data sample is not far from 1.0. If greater than 2.0, it means that the rating in that level may not make the rating criterion is not a meaningful measurement (Linacre, 1999) as outfit MnSq indices sensitive to outlying scores tend to be higher at the end of the scale than scores in the middle.

Lastly, the step or threshold calibrations, estimates for choosing one score level over another, indicate that they should not be too close together nor too far apart on the logit scale with the range of 1.4 to 5.0 logits as a rule of thumb to make a discernable difference between scores without large gaps (Bond and Fox, 2007). In Figures 22 to 24, the scale steps of all rating categories progressed in the order as intended, with each step being progressively more difficult than the previous step on the scale. The scale structure shows thresholds where two adjacent categories intersect.

The first “Mode” scale has marks for each category placed on the logits, starting with the most observable and, placing ^ at the mode of each category. The “Median” scale places each category at a logit value of 0.5 probability, with ^ at the median of each category. The “Mean” scale places ^ at the mean of each expectation measure category.

The distance between the scores in the scale structure of each category was discernable to some extent. Overall, the rating criterion *Content* had thresholds between adjacent score levels that were a bit narrower than expected ; specifically, the distance between thresholds of 2 and 3 and thresholds of 3 and 4 were close to each other in its scale

structure according to Figure 21.

In addition, as illustrated in Figures 25 to 27, FACETS configures the probability curves for each scale category. The probability curves serve as clear indicators of the structure of the rating scale. There are two important issues involved here: 1) whether there is a distinct peak for each score level probability curve or not, and 2) whether the curves appear at an even space and are shaped like a hill. A distinct peak signifies that the scale level that belongs to a specific curve is the most probable rating that test-taker performance in a certain portion of the test-taker proficiency distribution would receive. Without a distinct hill for a scale level, the scale level would not be the most probable rating for any portion of the test-taker proficiency distribution. Thus, the level does not contribute to specifying a clear point on the scale category measured. Such steps are considered operationally not to be worthwhile.

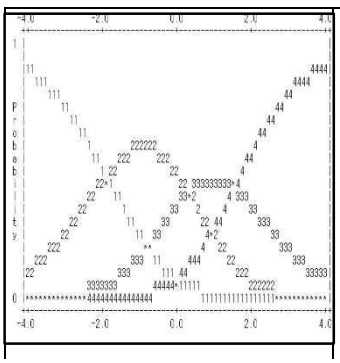


Figure 24. Probability curve (ACC)

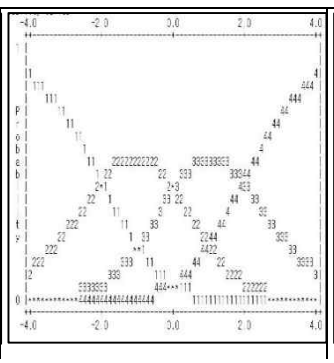


Figure 25. Probability curve (FLU)

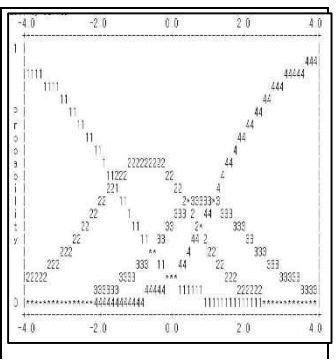


Figure 26. Probability curve (CONT)

The probability curves indicated that each rating scale found separate peaks on each score level, each of which represents the most

probable score choice for test-takers across some section of the ability being measured. Even though the overall scale functioning was not problematic, caution was warranted on the rating scale of Content, where two peaks were obviously lacking even space between Levels 3 and its adjacent levels.

4.3.2 Interaction analysis

FACETS also permits bias analysis. It is similar to differential item functioning (DIF) analysis in that it can identify any systematic patterns of interactions or bias of a facet with any other facet and estimate the effects of the interactions on test scores to address the quality of an assessment, or, more specifically, to understand the effects of assessment conditions on test scores. Investigating the interaction between two facets determines whether the bias size is statistically significant from t-test. It is important to detect and correct for bias in the Rasch model, in order to ensure that the measurement of the attribute is accurate and fair for all individuals and items being measured.

4.3.2.1 Mode interaction

(1) Gender by mode

As shown in Figure 28, there was no statistically significant difference in the perception of the MAR mode between males and females based on the significance probability range of the t-tests (p = 1.00).

Perceptions of the MAR mode did not differ between test-takers' genders.

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Gender SqN	G	measr-	Mode N	M	measr+
1555	1554.88	480	.00	.00	.07	-.01	479	.9930	.9	.9	4	2	M	.00	2	C	.00
950	949.91	336	.00	.00	.07	-.01	335	.9944	1.2	1.2	2	2	M	.00	1	D	.00
1728	1727.87	576	.00	.00	.06	-.01	575	.9935	1.1	1.1	6	2	M	.00	3	B	.00
3231	3230.77	1128	.00	.00	.05	-.01	1127	.9916	1.0	1.0	8	2	M	.00	4	A	.00
2053	2052.95	672	.00	.00	.07	.00	671	.9974	1.2	1.1	7	1	F	.00	4	A	.00
1459	1458.99	480	.00	.00	.07	.00	479	.9995	.9	.9	5	1	F	.00	3	B	.00
1230	1230.02	408	.00	.00	.07	.00	407	.9987	.9	.9	3	1	F	.00	2	C	.00
1187	1187.07	456	.00	.00	.07	.00	455	.9963	.9	.9	1	1	F	.00	1	D	.00
1674.1	1674.06	567.0	.00	.00	.07	.00			1.0	1.0	Mean (Count: 8)						
669.9	669.79	232.4	.00	.00	.01	.01			.1	.1	S.D. (Population)						
716.1	716.04	248.4	.00	.00	.01	.01			.1	.1	S.D. (Sample)						

Fixed (all = 0) chi-squared: .0 d.f.: 8 significance (probability): 1.00

Figure 28. Interaction statistics between the MAR mode and gender

(2) Region by mode

The MAR mode showed no significant difference between the two test-taker groups in the two different regions – Jinju and Changwon- based on the significance probability level of the t-tests (p = 1.00). Perceptions of the MAR mode did not differ between test-takers in the two different regions, as can be observed in Figure 29.

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Region SqN	Re	measr-	Mode N	M	measr-
2039	2038.85	624	.00	.00	.06	-.01	623	.9923	1.0	.9	3	1	JJ	.00	2	C	.00
2455	2454.82	792	.00	.00	.05	-.01	791	.9920	1.1	1.1	5	1	JJ	.00	3	B	.00
4675	4674.68	1560	.00	.00	.04	-.01	1559	.9900	1.1	1.0	7	1	JJ	.00	4	A	.00
1140	1139.90	408	.00	.00	.07	-.01	407	.9943	1.1	1.1	1	1	JJ	.00	1	D	.00
732	732.05	264	.00	.00	.09	.00	263	.9968	.8	.8	6	2	CW	.00	3	B	.00
746	746.05	264	.00	.00	.08	.00	263	.9965	.8	.8	4	2	CW	.00	2	C	.00
609	609.03	240	.00	.00	.11	.00	239	.9969	1.3	1.1	8	2	CW	.00	4	A	.00
997	997.08	384	.00	.00	.08	.01	383	.9954	.9	1.0	2	2	CW	.00	1	D	.00
1674.1	1674.06	567.0	.00	.00	.07	.00			1.0	1.0	Mean (Count: 8)						
1293.0	1292.91	416.7	.00	.00	.02	.01			.2	.1	S.D. (Population)						
1382.3	1382.18	445.5	.00	.00	.02	.01			.2	.1	S.D. (Sample)						

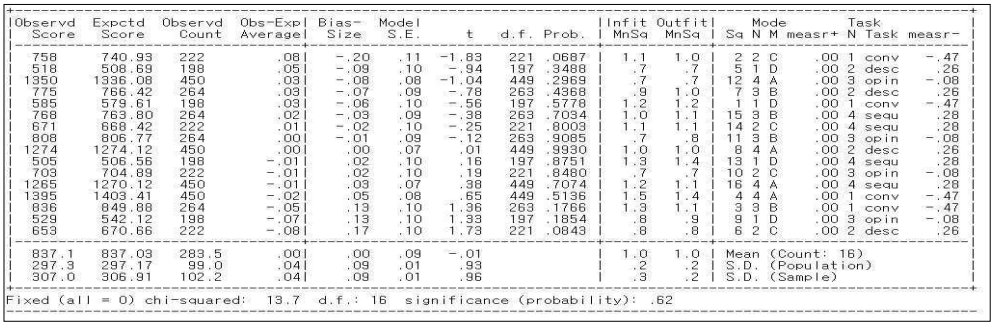
Fixed (all = 0) chi-squared: .0 d.f.: 8 significance (probability): 1.00

Figure 29. Interaction statistics between the MAR mode and region

(3) Criteria by mode

There was no statistically significant interaction between the perception of the MAR mode and the rating criteria based on the significance probability range of the t-test results, which was from 0.09

to 0.96. This might suggest no significant probability of whether test-takers' perceptions of the MAR mode affected the ratings of specific rating criteria to be assessed in the test, as illustrated in Figure 30.



Observvd Score	Expctd Score	Observvd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Mode Sq	N	M	measr+	Task N	Task	measr-
758	740.93	222	.08	-.20	.11	-1.83	221	.0687	1.1	1.0	2	2	C	.00	1	conv	-.47
518	508.69	198	.05	-.09	.10	-.94	197	.3488	.7	.7	5	1	D	.00	2	desc	.26
1350	1336.08	450	.03	-.08	.08	-1.04	449	.2969	.7	.7	12	4	A	.00	3	opin	-.08
775	766.42	264	.03	-.07	.09	-.78	263	.4368	.9	1.0	7	3	B	.00	2	desc	.26
585	579.61	198	.03	-.06	.10	-.56	197	.5776	1.2	1.2	1	1	D	.00	1	conv	-.47
768	763.80	264	.02	-.03	.09	-.38	263	.7034	1.0	1.1	15	3	B	.00	4	sequ	.28
671	668.42	222	.01	-.02	.10	-.25	221	.8003	1.1	1.1	14	2	C	.00	4	sequ	.28
808	806.77	264	.00	-.01	.09	-.12	263	.9085	.7	.8	11	3	B	.00	3	opin	-.08
1274	1274.12	450	.00	.00	.07	.01	449	.9930	1.0	1.0	8	4	A	.00	2	desc	.26
505	506.56	198	-.01	.02	.10	.16	197	.8751	1.3	1.4	13	1	D	.00	4	sequ	.28
703	704.89	222	-.01	.02	.10	.19	221	.8480	.7	.7	10	2	C	.00	3	opin	-.08
1265	1270.12	450	-.01	.03	.07	.38	449	.7074	1.2	1.1	16	4	A	.00	4	sequ	.28
1395	1403.41	450	-.02	.05	.08	.65	449	.5136	1.5	1.4	4	4	A	.00	1	conv	-.47
836	849.88	264	-.05	.13	.10	1.36	263	.1766	1.3	1.1	3	3	B	.00	1	conv	-.47
529	542.12	198	-.07	.13	.10	1.33	197	.1854	.8	.9	9	1	D	.00	3	opin	-.08
653	670.66	222	-.08	.17	.10	1.73	221	.0843	.8	.8	6	2	C	.00	2	desc	.26
837.1	837.03	283.5	.00	.00	.09	-.01			1.0	1.0	Mean (Count: 16)						
297.3	297.17	99.0	.04	.09	.01	.93			.2	.2	S.D. (Population)						
307.0	306.91	102.2	.04	.09	.01	.96			.3	.2	S.D. (Sample)						

Fixed (all = 0) chi-squared: 13.7 d.f.: 16 significance (probability): .62

Figure 30. Interaction statistics between the MAR mode and rating criteria

(4) Task by mode

According to Figure 31, there was no statistically significant interaction between the perception of the MAR mode and tasks based on the significance probability range of the t-test results (0.06 to 0.99). It can be said that the test-takers' perceptions of the MAR mode did not affect their performance in different task types.

(5) Test-takers' English proficiency by mode

Based on the significance probability range of the t-test results from0.92 to 0.99 (see Figure 32), there was no statistically significant interaction between the perception of the MAR mode and test-takers' general English proficiency level. It can be said that the test-takers' perceptions of the MAR mode did not vary according to their general English proficiency level.

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Mode Sq	N	M	measr+	Category N	Cate	measr-
988	973.75	296	.05	-.12	.09	-1.30	295	.1940	.7	.7	6	2	C	.00	2	FLU	-.35
1047	1036.51	352	.03	-.07	.08	-.83	351	.4064	1.1	1.0	3	3	B	.00	1	ACC	.17
1044	1034.13	352	.03	-.06	.08	-.78	351	.4352	1.1	1.1	11	3	B	.00	3	CONT	.18
1855	1844.74	600	.02	-.05	.07	-.68	599	.4946	.9	.9	8	4	A	.00	2	FLU	-.35
694	688.48	264	.02	-.04	.09	-.48	263	.6297	1.1	1.0	9	1	D	.00	3	CONT	.18
1722	1721.31	600	.00	.00	.06	-.04	599	.9648	1.0	1.0	4	4	A	.00	1	ACC	.17
689	690.43	264	-.01	.01	.09	.13	263	.9005	1.0	1.0	1	1	D	.00	1	ACC	.17
754	758.06	264	-.02	.03	.09	.36	263	.7180	1.0	1.2	5	1	D	.00	2	FLU	-.35
900	904.56	296	-.02	.03	.09	.39	295	.6959	1.1	1.0	10	2	C	.00	3	CONT	.18
1707	1717.67	600	-.02	.04	.06	.68	599	.4952	1.3	1.2	12	4	A	.00	3	CONT	.18
897	906.59	296	-.03	.07	.09	.82	295	.4110	.9	1.0	2	2	C	.00	1	ACC	.17
1096	1116.23	352	-.06	.14	.08	1.69	351	.0926	.7	.8	7	3	B	.00	2	FLU	-.35
1116.1	1116.04	378.0	.00	.00	.08	.00			1.0	1.0	Mean (Count: 12)						
395.3	395.10	132.0	.03	.07	.01	.82			.1	.1	S.D. (Population)						
412.9	412.67	137.9	.03	.07	.01	.85			.2	.1	S.D. (Sample)						

Fixed (all = 0) chi-squared: 8.0 d.f.: 12 significance (probability): .79

Figure 31. Interaction statistics between the MAR mode and task

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Comp Sq	N	C	measr-	Mode N	M	measr+
340	339.17	141	.01	-.01	.12	-.10	140	.9209	1.0	1.0	18	6	6	.00	3	B	.00
1035	1034.90	432	.00	.00	.07	-.01	431	.9943	.9	.9	24	6	6	.00	4	A	.00
1238	1237.92	384	.00	.00	.08	-.01	383	.9949	1.2	1.2	21	3	3	.00	4	A	.00
632	631.93	216	.00	.00	.09	-.01	215	.9950	1.0	1.0	22	4	4	.00	4	A	.00
1080	1079.93	336	.00	.00	.08	-.01	335	.9955	1.0	1.0	16	4	4	.00	3	B	.00
774	773.95	240	.00	.00	.11	-.01	239	.9956	1.1	1.1	20	2	2	.00	4	A	.00
611	610.94	240	.00	.00	.09	-.01	239	.9958	1.1	1.1	23	5	5	.00	4	A	.00
726	725.96	216	.00	.00	.11	.00	215	.9963	1.1	1.0	9	3	3	.00	2	C	.00
517	516.96	168	.00	.00	.11	.00	167	.9965	.9	.9	10	4	4	.00	2	C	.00
398	397.96	168	.00	.00	.11	.00	167	.9967	1.0	1.0	6	6	6	.00	1	D	.00
338	337.98	96	.00	.00	.17	.00	95	.9968	1.0	.9	13	1	1	.00	3	B	.00
468	467.97	168	.00	.00	.11	.00	167	.9970	1.0	1.0	11	5	5	.00	2	C	.00
472	471.97	144	.00	.00	.14	.00	143	.9971	.8	.8	15	3	3	.00	3	B	.00
994	993.97	288	.00	.00	.12	.00	287	.9973	1.4	1.1	19	1	1	.00	4	A	.00
575	574.98	168	.00	.00	.14	.00	167	.9975	.8	.8	8	2	2	.00	2	C	.00
774	773.96	288	.00	.00	.08	.00	287	.9976	1.0	1.1	4	4	4	.00	1	D	.00
467	466.98	168	.00	.00	.12	.00	167	.9979	1.1	1.2	5	5	5	.00	1	D	.00
134	133.99	48	.00	.00	.21	.00	47	.9979	.8	.8	12	6	6	.00	2	C	.00
393	392.99	120	.00	.00	.13	.00	119	.9985	.9	.9	14	2	2	.00	3	B	.00
365	364.99	120	.00	.00	.13	.00	119	.9985	.8	.8	7	1	1	.00	2	C	.00
199	198.99	72	.00	.00	.16	.00	71	.9990	.9	1.0	3	3	3	.00	1	D	.00
147	147.00	48	.00	.00	.20	.00	47	.9999	.7	.7	2	2	2	.00	1	D	.00
152	152.00	48	.00	.00	.25	.00	47	.9996	1.3	1.1	1	1	1	.00	1	D	.00
564	564.75	219	.00	.01	.10	.07	218	.9416	1.1	1.1	17	5	5	.00	3	B	.00
558.0	558.01	189.0	.00	.00	.13	.00			1.0	1.0	Mean (Count: 24)						
298.4	298.37	100.8	.00	.00	.04	.03			.2	.1	S.D. (Population)						
304.8	304.79	103.0	.00	.00	.04	.03			.2	.1	S.D. (Sample)						

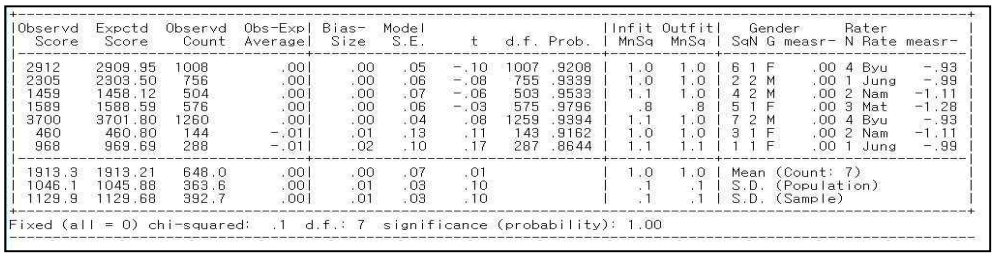
Fixed (all = 0) chi-squared: .0 d.f.: 24 significance (probability): 1.00

Figure 32. Interaction statistics between the MAR mode and test-takers' proficiency level

4.3.2.2 Rater interaction

(1) Test-taker gender by rater

According to Figure 33, there was no statistically significant interaction between the raters and test-takers' gender based on the significance probability range of the t-test results, which is 0.86 to 0.95. Thus, it followed that the behaviors of the raters were not affected by test-takers' gender.



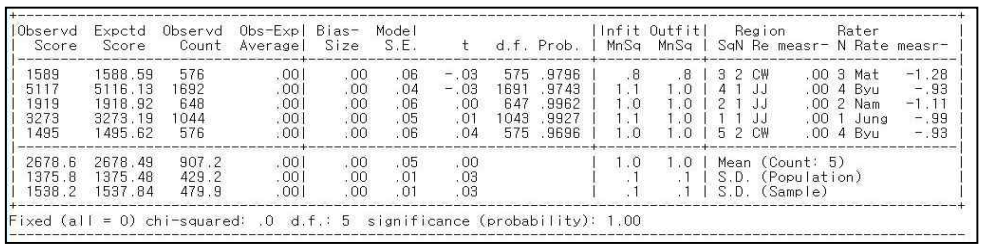
Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Gender SqN	G	measr-	Rater N	Rate	measr-
2912	2909.95	1008	.00	.00	.05	-.10	1007	.9208	1.0	1.0	6	1 F	.00	4	Byu	-.93
2305	2303.50	756	.00	.00	.06	-.08	755	.9339	1.0	1.0	2	2 M	.00	1	Jung	-.99
1459	1458.12	504	.00	.00	.07	-.06	503	.9533	1.1	1.0	4	2 M	.00	2	Nam	-1.11
1589	1588.59	576	.00	.00	.06	-.03	575	.9796	.8	.8	5	1 F	.00	3	Mat	-1.28
3700	3701.80	1260	.00	.00	.04	.08	1259	.9394	1.1	1.0	7	2 M	.00	4	Byu	-.93
460	460.80	144	-.01	.01	.13	.11	143	.9162	1.0	1.0	3	1 F	.00	2	Nam	-1.11
968	969.69	288	-.01	.02	.10	.17	287	.8644	1.1	1.1	1	1 F	.00	1	Jung	-.99
1913.3	1913.21	648.0	.00	.00	.07	.01			1.0	1.0	Mean (Count: 7)					
1046.1	1045.88	363.6	.00	.01	.03	.10			.1	.1	S.D. (Population)					
1129.9	1129.68	392.7	.00	.01	.03	.10			.1	.1	S.D. (Sample)					

Fixed (all = 0) chi-squared: .1 d.f.: 7 significance (probability): 1.00

Figure 33. Interaction statistics between rater and test-taker gender

(2) Test-taker residence by rater

It was shown in Figure 34 that there was no statistically significant interaction between the raters and test-takers' residence based on the significance probability range of the t-test results, which was 0.97 to 0.99. Therefore, the raters' behavior was not affected by test-takers' place of origin.



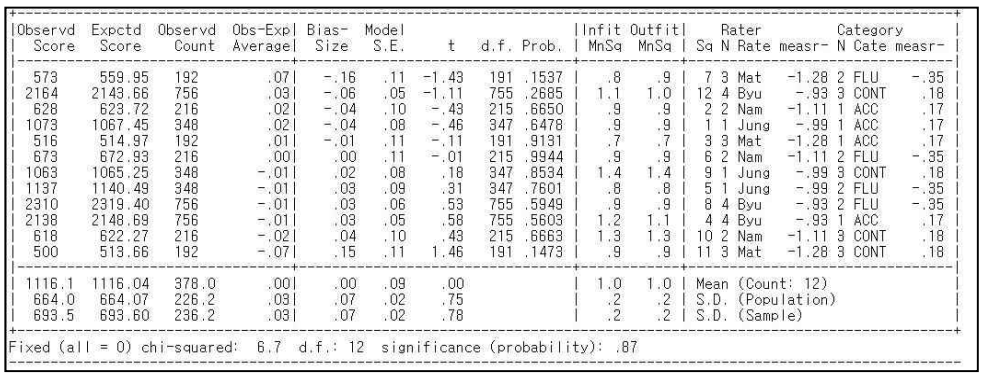
Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias- Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Region SqN	Re	measr-	Rater N	Rate	measr-
1589	1588.59	576	.00	.00	.06	-.03	575	.9796	.8	.8	3	2 CW	.00	3	Mat	-1.28
5117	5116.13	1692	.00	.00	.04	-.03	1691	.9743	1.1	1.0	4	1 JJ	.00	4	Byu	-.93
1919	1918.92	648	.00	.00	.06	.00	647	.9962	1.0	1.0	2	1 JJ	.00	2	Nam	-1.11
3273	3273.19	1044	.00	.00	.05	.01	1043	.9927	1.1	1.0	1	1 JJ	.00	1	Jung	-.99
1495	1495.62	576	.00	.00	.06	.04	575	.9696	1.0	1.0	5	2 CW	.00	4	Byu	-.93
2678.6	2678.49	907.2	.00	.00	.05	.00			1.0	1.0	Mean (Count: 5)					
1375.8	1375.48	429.2	.00	.00	.01	.03			.1	.1	S.D. (Population)					
1538.2	1537.84	479.9	.00	.00	.01	.03			.1	.1	S.D. (Sample)					

Fixed (all = 0) chi-squared: .0 d.f.: 5 significance (probability): 1.00

Figure 34. Interaction statistics between raters and test-taker's region

(3) Criteria by rater

According to Figure 35, there was no statistically significant interaction between raters and rating criteria based on the significance probability range of the t-tests (1.47 to 0.99). This finding might suggest that raters' performance was not affected by specific rating criteria.



Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Sq	Rater N	Rater	Rater measr	Category N	Cate	Category measr
573	559.95	192	.07	-.16	.11	-1.43	191	.1537	.8	.9	7	3	Mat	-1.28	2	FLU	-.35
2164	2143.66	756	.03	-.06	.05	-1.11	755	.2685	1.1	1.0	12	4	Byu	-.93	3	CONT	.18
628	623.72	216	.02	-.04	.10	-.43	215	.6650	.9	.9	2	2	Nam	-1.11	1	ACC	.17
1073	1067.45	348	.02	-.04	.08	-.46	347	.6478	.9	.9	1	1	Jung	-.99	1	ACC	.17
516	514.97	192	.01	-.01	.11	-.11	191	.9131	.7	.7	3	3	Mat	-1.28	1	ACC	.17
673	672.93	216	.00	.00	.11	-.01	215	.9944	.9	.9	6	2	Nam	-1.11	2	FLU	-.35
1063	1065.25	348	-.01	.02	.08	.18	347	.8534	1.4	1.4	9	1	Jung	-.99	3	CONT	.18
1137	1140.49	348	-.01	.03	.09	.31	347	.7601	.8	.8	5	1	Jung	-.99	2	FLU	-.35
2310	2319.40	756	-.01	.03	.06	.53	755	.5949	.9	.9	8	4	Byu	-.93	2	FLU	-.35
2138	2148.69	756	-.01	.03	.05	.58	755	.5603	1.2	1.1	4	4	Byu	-.93	1	ACC	.17
618	622.27	216	-.02	.04	.10	.43	215	.6663	1.3	1.3	10	2	Nam	-1.11	3	CONT	.18
500	513.66	192	-.07	.15	.11	1.46	191	.1473	.9	.9	11	3	Mat	-1.28	3	CONT	.18
1116.1	1116.04	378.0	.00	.00	.09	.00			1.0	1.0	Mean (Count: 12)						
664.0	664.07	226.2	.03	.07	.02	.75			.2	.2	S.D. (Population)						
693.5	693.60	236.2	.03	.07	.02	.78			.2	.2	S.D. (Sample)						

Fixed (all = 0) chi-squared: 8.7 d.f.: 12 significance (probability): .87

Figure 35. Interaction statistics between rater and rating criteria

(4) Task by rater

According to Figure 36, there were three cases of statistically significant interactions between two raters – Rater 1 Jung and Rater 3 Mat – and two tasks – Task 1 (dialogue completion) and Task 3 (explaining sequential information)– based on t–values within significance probability levels of 0.05.

To begin with the scoring behavior of Rater 3 (Mat), the following investigation reported that it did not seem to make discernable impact on the measurement of test–takers' ability to perform Task 1, dialogue completion, because this interaction did not cause any unexpected response among 32 residuals in Appendix 4. On the other hand, Task 3, explaining information, the sequence of events, with which both raters were in interaction, warrants further investigation.

According to Figure 36, both Rater 1 (Jung) and Rater 3 (Mat) made meaningful discrepancy between observed and expected scores with the bias size –.26 and .29 respectively. The negative bias value for Task 3

rated by Rater 1(Jung) might suggest that she was being consistently harsher than she rated the other tasks, while the positive bias value for the same task when rated by Rater 3 (Mat) might mean that he rated opposite, more lenient than he was to other tasks.

Observd Score	Expctd Score	Observd Count	Obs-Exp Average	Bias Size	Model S.E.	t	d.f.	Prob.	Infit MnSq	Outfit MnSq	Rater Sq	N	Rate	measr	Task N	Task	measr
450	427.25	144	.16	-.39	.13	-2.90	143	.0043	1.2	1.1	3	3	Mat	-1.28	1	conv	-.47
817	788.06	261	.11	-.26	.10	-2.72	260	.0070	1.1	1.1	13	1	Jung	-.99	4	sequ	.28
470	461.06	162	.06	-.12	.12	-1.04	161	.3000	.8	.9	6	2	Nam	-1.11	2	desc	.26
489	485.80	162	.02	-.05	.12	-.38	161	.7047	.8	.8	10	2	Nam	-1.11	3	opin	-.08
1599	1588.21	567	.02	-.04	.06	-.67	566	.5000	.9	.9	8	4	Byu	-.93	2	desc	.26
1682	1673.81	567	.01	-.03	.06	-.52	566	.6016	.8	.8	12	4	Byu	-.93	3	opin	-.08
1586	1582.69	567	.01	-.01	.06	-.21	566	.8362	1.2	1.2	16	4	Byu	-.93	4	sequ	.28
513	512.61	162	.00	-.01	.12	-.05	161	.9612	1.6	1.4	2	2	Nam	-1.11	1	conv	-.47
866	866.93	261	.00	.01	.10	.10	260	.9242	1.3	1.1	1	1	Jung	-.99	1	conv	-.47
379	380.11	144	-.01	.02	.12	.14	143	.8915	.5	.5	7	3	Mat	-1.28	2	desc	.26
401	402.55	144	-.01	.02	.12	.19	143	.8480	.6	.6	11	3	Mat	-1.28	3	opin	-.08
818	827.70	261	-.04	.09	.10	.94	260	.3456	.7	.8	9	1	Jung	-.99	3	opin	-.08
1745	1767.04	567	-.04	.10	.07	1.46	566	.1461	1.3	1.2	4	4	Byu	-.93	1	conv	-.47
772	790.50	261	-.07	.16	.09	1.74	260	.0825	1.0	1.1	5	1	Jung	-.99	2	desc	.26
447	459.46	162	-.08	.17	.12	1.45	161	.1500	1.1	1.1	14	2	Nam	-1.11	4	sequ	.28
359	378.68	144	-.14	.29	.12	2.41	143	.0173	1.0	1.0	15	3	Mat	-1.28	4	sequ	.28
837.1	837.03	283.5	.00	.00	.10	.00			1.0	1.0	Mean (Count: 16)						
498.6	498.69	169.6	.07	.16	.02	1.41			.3	.2	S.D. (Population)						
514.9	515.04	175.2	.07	.16	.03	1.46			.3	.2	S.D. (Sample)						

Fixed (all = 0) chi-squared: 31.8 d.f.: 16 significance (probability): .01

Figure 36. Interaction statistics between rater and task

In terms of interpreting the magnitude of DIF effect sizes, there is no universally agreed-upon scale or threshold for determining the magnitude of bias in MFRM. However, a commonly used guideline is to consider a DIF effect size greater than 0.5 or less than −0.5 as indicative of substantial DIF (Shealy & Stout, 1993). This threshold is based on the notion that a DIF effect size of 0.5 corresponds to a difference of one standard deviation on the log-odds or logit scale, which is considered a meaningful difference in many contexts.

According to the guideline, the bias sizes of Rater 1(0.29) and Rater 3(−0.26) were not considered substantial as the values did not exceed the threshold. In spite of the less substantial magnitude of bias size, it is worthwhile to investigate rating behaviors of Rater 3 in particular that

showed statistically significant interactions with two tasks, Task 1 and Task 3. Subsequent analysis about the testing process would be made in the following section 4.4 including not only the test users' overall perceptions on the MARST but close examination of test-takers' spoken responses to Task 3.

4.3.3 Analysis of unusual responses in MFRM

Identifying and handling unusual responses in the MFRM analysis is important to ensure the validity and reliability of the test results. These responses may indicate problems with the test items, response format, or test-takers themselves, and may require further investigation or modifications to improve the test quality. Unusual responses in the MFRM analysis can take different forms, such as unexpected response patterns, responses with interaction with, or extreme responses (e.g., selecting the highest or lowest response option for all items).

The following is an attempt to meet the need of further analysis in identifying and addressing sources of response deviations disclosed in MFRM analysis including person (ability) misfits, rater-task interaction and unexpected residuals in the previous section of 4.3.2.2. To serve these purposes, test-takers were divided into three groups according to their English proficiency levels.

4.3.3.1 Sources of person (ability) misfits

Overall, misfit cases in the MFRM can be caused by a variety of factors, and it is important to identify the sources of response deviation in order to address them appropriately and improve the validity and reliability of the test scores. It is because they may indicate that the test-taker's responses are significantly different from what is expected based on their ability level and the item difficulty.

In this dissertation, misfits were only identified in the test-takers' ability facet with none in all the other facets such as rater, rating criteria, and task. And there were 18 cases with misfits in both infit and outfit measures greater than 1.5. (See Appendix 7)

According to studies of potential sources of person misfits (Linacre, 1998; Nering & Ostini, 2010), there are a variety of factors such as guessing, carelessness, misunderstanding the item or instructions, and anxiety. In the questionnaire survey, 14 out of 18 test-takers with misfits showed positive perceptions on the use of MARST, answering 'agree and strongly agree in the Likert scales of most items.

In the interviews, however, some of them shared with the researcher troubled experiences they suffered and mistakes they accidentally made from misunderstanding the instructions so they had to stay longer on the app against their will to record their responses again. In addition, they said that during the test, unexpected ambient noises from outside or incoming calls to their mobile phones distracted their attention from speaking on the phone, which also made them take the test several

times.

Thus, the potential sources of the person misfits could probably have been related to misunderstanding the instructions delivered in the new test mode and carelessness due to lack of attention and unwilling repetition. One of solutions may be the sufficient amount of time for hands-on exercises before the actual test begins. With this exercise, test-takers can understand what are inside the app and how to use them in advance.

4.3.3.2 Sources of bias: rater x task interaction

The rater interaction with task in MFRM analysis of the previous section, 4.3.2.2 indicated that Rater 3 was harsh in rating Task 1 but lenient in Task 3. The further inspection obviously indicated that Rater 3 consistently scored test-takers in the low ability group (the average measures of 0 or under the average zero) higher than the counterpart rater, Rater 4 in at least one of the three rating criteria in 13 cases out of 20. Meanwhile, when it came to rating Task 1, Rater 3 was harsh to test-takers.

Based on the online interview with Rater 3, it can be cautiously concluded that the bias might have been due to scoring order, that is starting from the lowest scores to the highest and lack of attention. Bias from scoring order can be prevented by using randomized scoring order or balanced scoring order. Randomized scoring order involves randomly assigning the order in which responses are scored. This helps to prevent any systematic

bias from scoring in a particular order from Task 1 to Task 3.

Additionally, scoring in a consistent and unbiased manner regardless of task, rating criteria and test-taker variabilities calls for more attention to the rating process. First, monitoring the behavior of raters during the rating process to identify any potential issues, such as rater fatigue or leniency, which could affect the reliability of the ratings. Second, providing regular feedback to raters on their ratings to help them improve their consistency and accuracy. This feedback can include discussions of individual ratings, as well as more general feedback on the rating process.

4.3.3.3 Sources of unexpected responses

In the MFRM, an unexpected response is a response given by a test-taker that does not fit the expected pattern of responses based on their overall ability level and the difficulty level of the item. This means that the test-taker's response deviates significantly from what would be predicted based on their level of proficiency, and the properties of the item itself.

Unexpected responses can occur for various reasons, such as guessing, carelessness, misreading the item, or simply making an error in responding. Other sources may also include in MFRM biased item, multidimensionality. These responses can negatively affect the accuracy and reliability of the test scores, as they can distort the measurement of

the test taker's proficiency level.

Of the 32 unexpected responses (see Appendix 7) in this dissertation, the noteworthy cases were the speaking responses of four test-takers (Person 32, 129, 158 and 70). Responses of the first three test-takers were considered unexpected by two raters, and while responses of the last test-taker (Person 70) proved unexpected in some tasks (Task 1 and 4) from one rater.

Further inspection of their speaking responses were made. To begin with responses of Person 32 for Task 1, in the interview with her, it was found that during the actual test, she seemed to misread the instruction and made a mistake of failing to understand what the task intended to assess : two expressions useful to serve the two language functions in turn-taking with the simulated interlocutor : making suggestion and expressing the ability to do something. It turned out that Person 32 only demonstrated knowledge for the language function of expressing the ability to do something. Thus, unexpectedly, Person 32 obtained lower marks in Task 1 in spite of the perfect scores in the other tasks.

As for responses of Person 129 and 158 for Task 4, they were commonly characterized as lack of content, making their responses almost empty, compared to what they responded in the other tasks. Lastly, Person 70 seemed to fail to record responses in proper volume; hence, it was almost impossible for two raters to hear them. Consequently, the aforementioned analysis of unexpected responses here may suggest that carelessness occurred to those test-takers

without giving proper consideration for task performance. This may have resulted into responses that did not reflect the test-takers' true abilities or knowledge. Addressing unexpected responses can involve providing feedback to the test-takers to help them prepare for the new test by providing additional instruction or motivating exercises for training test-takers .

4.4 Analysis of the testing process

4.4.1 Perceptions of MAR mode

In the literature review, when applying a new test mode, test-takers' anxiety tends to increase because they cannot control the new test owing to the absence of a human interlocutor in the machine-delivered assessment. Thus, through questionnaires and interviews, this study aims to elucidate the effect of the modes of cognitive and emotional experiences. The results of the questionnaires given to test-takers are summarized in Table 40.

4.4.1.1 Students

When it came to the use of the device or the MAR application, which Items 1 to 3 addressed, test-takers thought it was more interesting and comfortable to speak via the MAR application than to have a face-to-face test. The most common adjectives used to describe the MAR application device wee "interesting" and "comfortable".

Table 40. Summary of test-takers' questionnaire result[27](N=199)

Category	MAR app			MAR-delivered test materials	
Item	Q1	Q2	Q3	Q4	Q5
Trait	Interest	Motivation	Comfort	authenticity	Sufficiency
Mean	3.26	2.77	3.18	3.41	2.94
Category	MAR-delivered test materials		MAR-delivered test		
Item	Q6	Q7	Q8	Q9	Q10
Trait	interactiveness	Clarity	Relevance	appropriacy	Usefulness
Mean	2.67	2.93	3.02	3.12	2.67

Regarding the test inputs in Items 4 to 7, the majority of test-takers agreed that test inputs were highly authentic (3.41), which means the tasks were representing real-life language use situations. A significant number of test-takers agreed to some extent that the test provided sufficient guidance (2.94) that made it easy for them to understand what was being asked of them and to construct their responses. The remarks most characteristic of the MAR test materials were "authentic" with "sufficient guidance".

Finally, Items 8 to 10 addressed the overall perceptions of the MARST test. The majority of test-takers agreed that the test topics were relevant for high school students (3.02), and the test tasks were appropriate to be presented in the MAR mode (3.12). Thus, it could be said that what characterized the MAR test were the terms "appropriate" and "relevant".

As for Research Question 3, a quantitative analysis was conducted to investigate whether the perceptions of the MARST differed according

[27] The overall average score is 30.00, out of a perfect score of 40.00 on a scale of 0 to 4.

to test-takers' gender and general proficiency level. The mode interaction in 4.2.3.1 found no significant difference between males and females.

After the test administration, there were interviews where ten test-takers were asked to elaborate on their numerical responses. Comments on the questionnaires, including the strengths and weaknesses of the MARST, included the following :

▶ Strengths

- Because it was not face-to-face, I was able to pronounce English words confidently because I felt less pressure to speak English.
- Often, even if you prepared hard for an exam, you could ruin it due to nervousness, but this type of test can avoid it.
- I didn't have enough chances to speak English, but this turned out to be a good opportunity. It would help me to develop speaking skills.
- The presented material was lively and fresh, so I was able to focus. I especially liked watching animated visual images as well as static ones. I realized that English tests could be fun.
- It(the environment poster) came out as a video rather than a photo, so it was easier to observe and pick up a lot of things to talk about.
- The tasks were not complicated and what I had to do was simply record, so they were appropriate for me to perform with my mobile phone.
- The tasks were related to the global environment, which was quite informative and appropriate to the level of the students since we knew what the tasks asked us was based on what we learned in class.
- It was nice to be able to hear the English pronunciations while talking with the avatar, Eco Bear, and I felt like I had a conversation partner.
- It was nice to be able to do it alone anytime, anywhere at a convenient time. Complaints about test order disappear. It was nice to be able to

access multiple people at the same time and to hear the questions multiple times because I could use the replay function within the time limit.

- This test could make it possible for teachers to implement performance assessments even in case of online classes.

▶ Weaknesses

- There was an error in implementing the AR. The video was cut off even when the marker moved slightly away from the camera. The marker seemed too sensitive.
- It was cumbersome to learn how to use the AR in the first place.
- It was inconvenient to have to hold a mobile phone to view the screen.
- Different mobile phone models may affect installing and using the application.
- I was embarrassed that my face appeared on camera.
- The letters on the screen were too small and there was no place to put up the phone. I couldn't see the letters and pictures well.
- I didn't feel like having a conversation in the dialogue completion task because the virtual interlocutor spoke slowly and the tone of her voice was like a robot.
- The screen went by so quickly that it was not easy to think of what to say in English. It was not easy to come up with an answer right away.
- I didn't know where to put my eyes when I spoke in the recording scene. I would like to see the interlocutor's face appear so that I could make eye contact in the recording scene.
- I thought that it would be difficult to catch cheating and at the same time, any meaningless actions could be mistaken for cheating.
- It was awkward to speak alone, and I couldn't confirm whether I understood what I was saying.

4.4.1.2 Teachers

The results of the questionnaires given to 17 teachers are summarized in Table 41. The teachers' responses were gathered in the teacher workshop in summer 2022, as shown in Figure 37. There was an overall tendency that teachers evaluated the MARST test higher than students did. For the use of the device or the MAR application that the Items 1 to 3 address, the test-takers thought it the most interesting and more motivating to speak via the MAR application than face-to-face. The most common comments by teachers regarding the MAR application device were that it was "interesting" and "motivating".

Regarding the MAR-delivered inputs in Items 4 to 7, akin to a large number of test-takers, the majority of teachers found the MAR-delivered test inputs to be highly authentic (3.76). Moreover, a number of test-takers agreed to some extent that the test provided a sufficient amount of input (3.35) and clear guidance (3.35). The remarks most characteristic of the MAR test materials were "authentic" with "sufficient and clear guidance". Items 8 to 10 addressed the overall perceptions of the MARST test.

Table 41. Summary of teachers' questionnaire results(N=17)[28]

Category	MAR app			MAR-delivered test materials	
Item	Q1	Q2	Q3	Q4	Q5
Trait	interest	motivation	Comfort	authenticity	sufficiency
Mean	3.88	3.47	3.29	3.76	3.35
Category	MAR-delivered test materials			MAR-delivered test	
Item	Q6	Q7	Q8	Q9	Q10
Trait	interactiveness	clarity	Relevance	appropriacy	usefulness
Mean	3.00	3.35	3.53	3.53	3.47

[28] The overall average score is 34.65 out of a perfect score of 40.00 on a scale of 0 to 4.

The majority of teachers agreed that the test tasks were relevant to the topics and skills necessary for high school students to learn (3.53), and appropriate to be presented in the MAR mode (3.53). Teachers characterized the MAR test as “relevant” and “appropriate”.

The lower scores observed in students' responses compared to those of teachers can be attributed to the psychological burden associated with the test's purpose as an achievement assessment, even if the stakes were relatively low. This emphasizes the importance of implementing MAR-based speaking assessment in the form of formative or diagnostic assessments. Such assessments are generally considered more learning-oriented, serving as valuable tools for learning and improvement, before utilizing the MAR-based assessment as a final achievement assessment.

At last, comments on the questionnaires including strengths and weaknesses of the MARST were reported as follows :

▶ Strengths

- The test was easy to understand with realistic tasks.
- Augmented reality was so interesting and fun that test-takers could concentrate on tasks.
- The exam burden will be greatly reduced. Test-takers will be less nervous.
- Students could immerse themselves in the task.
- The test tasks were appropriate to be presented via MAR and they would help test-takers’ understanding of what to do and how to do it.
- The test enables test-takers to practice and perform tasks repeatedly.
- I felt like I was actually having a conversation with a person. I would like

to practice tasks and dialogues developed with more topics other than the environment.

▶ Weaknesses

- I had difficulty getting into the application at first.
- It would be difficult for students unless there is enough prior practice.
- It was inconvenient that I had to wait for the next level until the Eco Bear(narrator) finishes its narration.
- I hope the re-recording function would be less cumbersome. I had to go all the way back to the beginning for re-recording.
- Ambient noise was recorded during the response.

Figure 37. Teacher's workshop for practicing the MARST

4.4.2 Analysis of speaking responses

The questionnaires and interview data analysis in the preceding section provided useful information regarding test-takers' reactions to test tasks, the new test mode and the overall process of test performance via the MARST. However, another qualitative method was employed in this study that more directly explored test-taking processes and the linguistic features of test-takers' responses that they revealed during the task, which was associated with Research

Question 4. It was done by analyzing transcription of authentic spoken discourse of several selected test-takers' oral performance.

This qualitative approach brought attention to spoken discourses of speaking responses in Task 3, a monologic task for dual purposes. It first attempted to investigate how the capability of technical affordances of MAR test mode – interactiveness and authenticity – supported the assessment of communicative language competence particularly associated with interactional and interpersonal features, an integral part of abilities for spoken communication. A number of samples of transcribed spoken response data are provided in Appendix 9.

In the discourse analysis of test-takers' speaking responses to Task 3, several resources were noted in the following that presumably signal the test-takers' role or the identity that they displayed in the context of the task situation. Notably, identity may be signaled by a participant's choice of specific linguistic devices (Young & Miller, 2004). The spoken data analyzed here was transcribed into text for qualitative analysis by AntConc 4.2.0 (Anthony, 2022), a freeware corpus analysis toolkit for text analysis.

The corpus analysis was conducted in reference to the American English corpus[29], which is free for research purpose. The "Corpus manager" menu in AntConc allows users to load corpora from

[29] This corpus present in AntConc 4.2.0(Anthony, 2022) is made up of 80 files with 161,469 tokens and 17,804 types that deal with topics of various areas : mathematics, natural sciences, political science, social and behavioral sciences, technology and engineering, and education etc. made in 2006.

the online corpus repository, as shown in Figures 38 and 39.

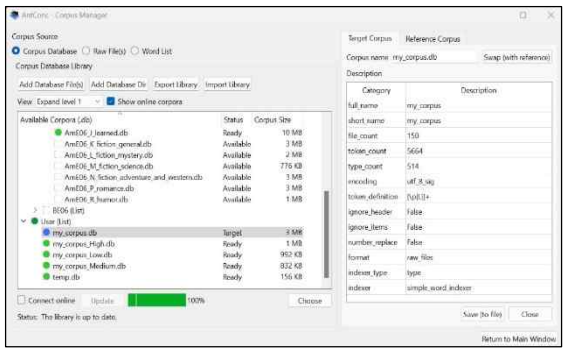


Figure 38. A screenshot of loading a target corpus in Corpus manager menu

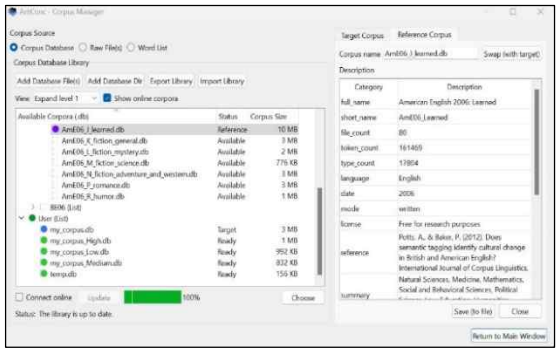


Figure 39. A screenshot of loading a reference corpus in Corpus manager menu

Additionally, the keyword analysis feature within this menu enables users to compare a target corpus with a reference corpus, helping identify words that exhibit significantly higher or lower frequencies in the target corpus.

4.4.2.1 Overview

For the qualitative analysis, the audio files of speaking responses to Task 3 were transcribed into text so that a small-size corpus was

generated as summarized in Table 42. Then, several useful ways to analyze the text provided the linguistic features communicated by the MAR test mode.

First of all, the type–token ratio (TTR) is a traditional linguistic measure of lexical diversity or richness of a text or a language sample. It calculates the ratio of unique words (types) to the total number of words (tokens) in a given text. Higher TTR values suggest greater lexical diversity, indicating that the text or sample contains a wide range of different words. Conversely, lower TTR values indicate less lexical diversity, suggesting a narrower range of vocabulary used.

Importantly TTR results should be interpreted with caution. One central issue here is that the TTR of a text sample is influenced by its length. As a text extends, the chances of the subsequent word being a repetition of a previously used word increase. Additionally, different genres or language styles can exhibit varying TTR ranges. Thus, it should be interpreted within the context of the text's length, task, topic, and genre (Tilstra & Smakman, 2018). Its relationship to language proficiency levels therefore still remains inconclusive, that is to say that the increase in TTR does not necessarily indicate the increase in proficiency levels (Espada–Gustilo, 2011; Wang, 2014).

Likewise, the TTR of Task 3 in Table 42 showed the TTR decreased as the proficiency level increased. One important thing to be taken into account here is task feature. Task 3 is characteristic of limiting opportunity for word variation as what should be included in the response

was rather context-specific. This is because when the context of a response is very specific, there is a certain range of vocabulary or expressions that are appropriate or relevant to the context of Task 3. Therefore, it is essential to consider the context of the analysis when interpreting TTR results.

Table 42. Token and type of the speaking response corpus in three proficiency groups

Category \ Proficiency	All (N=150)	High (n=50)	Medium (n=50)	Low (n=50)
Token[30]	5664	2450	1922	1292
Type[31]	514	294	344	262
TTR[32]	9	12	17.8	20.2

To identify the main topics or themes to compare with texts with different lengths, the size of normalized range[33] was calculated. The normalized range in n-grams refers to the diversity or variety of n-grams in a given text or dataset. Range refers to the number of unique words or terms used in a text by taking into account the size of the corpus. In the case of normalized range in AntConc, it assesses how frequently a particular feature appears across different texts in the corpus. A high normalized range means that a feature is consistently used across multiple texts, while a low range means the feature's use is more

[30] Token refers to the total number of words in a particular context or corpus.
[31] Type refers to the number of distinct words used in a particular context or corpus.
[32] TTR refers to type-token ratio which indicates lexical diversity.
TTR = (type / token) * 100%
[33] The concept of normalized range is used to measure lexical diversity in a text or corpus. The formula for normalized range in text linguistics is:
Normalized range = (number of different words / total number of words) * 100%

inconsistent.

4.4.2.2 Keyword analysis

According to Table 43, it can be said that as the proficiency level goes up, key words listed would be more widely used across different test-takers' responses, given the values of normalized range. The topic of test-takers' speaking responses was directly associated with the environment and as a part of the working process, these words became essential to be included when they constructed responses. In particular, keywords common across all three groups are topic-specific words such as 'cloud', 'signal', 'sound', 'ranger' and 'logging' . However, the range values of the same keywords differed by proficiency group, which that

Table 43. Top ten keywords in three proficiency groups[34]

Type	All	Type	High	Type	Medium	Type	Low
cloud	0.68	**cloud**	0.94	**cloud**	0.66	**cloud**	0.44
signal	0.58	**signal**	0.78	**signal**	0.58	forest	0.40
sound	0.53	first	0.74	**sound**	0.54	**signal**	0.38
ranger	0.47	**sound**	0.72	rainforest	0.44	rainforest	0.34
solar	0.46	go	0.70	**ranger**	0.44	**ranger**	0.32
send	0.44	sends	0.68	**logging**	0.4	tree	0.32
rainforest	0.40	**ranger**	0.66	forest	0.38	device	0.28
logging	0.39	solar	0.64	illegal	0.38	solar	0.36
forest	0.37	immediately	0.64	sends	0.38	**sound**	0.32
illegal	0.35	**logging**	0.46	microphone	0.38	**logging**	0.30

[34] Ten keywords are listed in descending order by the size of normalized range. The ranges were estimated in the log-likelihood method within the significance level of .05 (p <.05, 3.84 with Bonferroni).

speaking responses of more proficient groups used more diverse topic-specific words.

The high proficiency group exhibited the use of specific keywords, including the linking adverb 'first,' action verbs such as 'go' (0.7) and 'send' (0.68), and the temporal adverb 'immediately' (0.64). These keywords hold significance within this group. The presence of the linking adverb 'first' suggests that proficient speakers were aware of the task requirement to initiate the sequential order of events and had the ability to organize them effectively during the test. The frequent use of action verbs in the same proficiency group indicates that speakers recognized the presence of interlocutors (rainforest rangers) and were conscious of their role identity as staff in an engineering company.

The most notable characteristic observed in the response data of the high proficiency group was the frequent use of the temporal adverb "immediately." This usage may be attributed to the sense of urgency that speakers perceived due to the visually simulated presentation of the detrimental effects of illegal logging on the Amazon rainforest. These findings are likely to provide further support for the positive influence of immersive AR mode on the linguistic features observed in speaking responses.

Note that five of the top ten keywords appearing frequently at all proficiency levels, as shown in Table 43, turned out to be highly relevant to English classroom vocabulary instruction of the main text. These words were found to be the words used in a relevant unit in the English

textbook at the high school the test-takers were attending. See the underlines in Excerpt 1 from the main text below.

[Excerpt 1 from the main text (Han, Jeong, Park, Lee, Lee, & Jang (2018)]

How the Device Works

1. It all starts here! Sound of chainsaws is picked up by microphones in solar-powered cell phones.
2. Software sends a signal to cloud.
3. Real-time alert is received by a ranger on the ground nearby.
4. That enables the rangers to go to the site immediately.

White returned to Indonesia to test the device. Surprisingly, on only the second day after he installed the device, it picked up chainsaw noises. An alert message was immediately sent to White and the forest rangers. When they approached the logging spot, the illegal loggers ran away.

In other words, the distribution of these keywords can be used to evaluate the extent of each test-taker's mastery of classroom-taught material and to compare language mastery of test-takers in the accuracy rating scale. This process helps ensure that the test tasks align with their intended assessment purpose of being an achievement test.

4.4.2.3 N-gram analysis

AntConc provides several options for calculating collocation measures to analyze word associations within a text corpus. Investigating n-grams provides valuable insights into the structure, content, and meaning of a text as what we can learn from investigating n-grams are

collocations and word associations, style and tone. Thus, the n-gram analysis results provide information about syntactical information that may have been overlooked in keyword analysis. For example, while keyword analysis may reveal the usage of verbs like "send" and "go," the correct usage of their syntactic structures can be identified through n-gram results, such as "go to the site" or "send a signal to."

For Task 3 of the current study, the top ten list of 4-grams in order of normalized range was presented across proficiency levels. The selection of 4-word sequences is justified by the fact that 4-grams encompass 3-grams and also carry more tokens compared to 5-grams(Cortes, 2004, as cited in Hong, 2013). Analyses using 4-grams can be compared with findings from other research, so they are often employed in studies focusing on consecutive word sequences. (Ädel & Erman, 2012; Chen & Baker, 2010; Cortes, 2004, 2006, as cited in Hong, 2013).

Below are some common options available in AntConc. Outlined in Table 43, the use of collocations, an indicator of automaticity and fluency in speaking skills, differed across proficiency levels; for example, although the collocation phrase "send a signal to" appeared across all proficiency groups, the distribution of the phrase was higher (0.5) in the high proficiency group compared with the medium (0.2) and low (0.1) proficiency groups.

Table 44. Top ten 4-grams in three proficiency groups[35]

Type	High	Type	Medium	Type	Low
sends a signal to	0.5	a signal to cloud	0.2	a signal to cloud	0.12
a signal to cloud	0.42	sends a signal to	0.2	alert is received by	0.1
go to the site	0.42	alert is received by	0.12	go to the site	0.1
software sends a signal	0.4	go to the site	0.12	real time alert is	0.1
to the site immediately	0.4	is picked up by	0.12	sends a signal to	0.1
real time alert is	0.32	received by a ranger	0.12	to the site immediately	0.1
alert is received by	0.3	to the site immediately	0.12	by a ranger on	0.08
is picked up by	0.3	by a ranger on	0.1	is picked up by	0.08
by a ranger on	0.26	on the ground nearby	0.1	on the ground nearby	0.08
on the ground nearby	0.26	rangers to go to	0.1	received by a ranger	0.08

This result was supported by the statistically significant test of one-way analysis of variance (ANOVA) at the .05 level ($F(2) = 49.636, p = .00$). A post hoc Games-Howells test [36]indicated that the mean normalized ranges of eight 4-grams in the high proficiency group were significantly higher than that of the medium and low proficiency groups . However, there were no significant mean differences between the medium and low proficiency groups ($p = .08$). This finding may suggest that highly proficient L2 learners are more likely to use a greater number and wider variety of collocations than the other less

[35] Ten 4-grams in each proficiency group are listed in descending order of normalized range. AntConc, however, does not have a built-in function to execute a statistical test to determine the statistical significance of n-gram measures (Hong, 2013; Bal, 2010). Thus, follow-up analysis was conducted in IBM SPSS Statistics for Windows(Version 26).

[36] This test is a nonparametric approach to compare groups as the group dataset does not assume equal variances.

proficient groups, which can further explain their higher scores obtained in the rating scale of fluency. Detailed reports of the one way ANOVA test result appear in Appendix 11.

In addition, eight of the top ten 4-grams in three proficiency groups appeared frequently at all proficiency levels, as shown in Table 44. With a look at the excerpt taken from the main text (see Excerpt 2 below), it seemed likely that test-takers utilized key elements from the main text when constructing their responses. Thus, the findings from the 4-gram analysis across the three proficiency groups could potentially provide a supportive rationale for the relevance of these phrases to teaching and learning target syntactic features (i.e., the use of passives, prepositional phrases, and the syntactic structure of the verb "send") in classrooms.

[Excerpt 2 from the main text (Han et al., 2018)]

How the Device Works

1. It all starts here! Sound of chainsaws is picked up by microphones in solar-powered cell phones.
2. Software sends a signal to cloud.
3. Real-time alert is received by a ranger on the ground nearby.
4. That enables the rangers to go to the site immediately.

Put simply, the distribution of these 4-grams provides a basis for comparing test-takers' achievement and assessing the language proficiency of test-takers in the accuracy rating scale. This ensures that the test task is aligned with its intended purpose of being an achievement test.

4.4.2.4 Interpersonal/interactional resources

The interpersonal/interactional features have been incorporated into the sequential organization of speech acts. These features are especially prominent in responses to Task 3, which requires test-takers to provide directions or instructions. That is, this task demands not only a good grasp of the English language but also the ability to organize and present information logically, clearly, and sequentially.

The importance of organizational structure in Task 3 is most commonly marked by sequential adverbs that signal the beginning, continuation, or end of a particular step of the process of explaining how the target device works, forming a distinct pattern: (naming of the device)→ signaling the start of the task → explaining the sequence of how the device works and how to use it (→ ending). The sequence typically began with identifying or naming the device of interest. Next, the speaker signaled the start of the task or procedure. This was followed by detailed explanations of how the device worked and how it should be used. The sequence might optionally end with a clear conclusion or wrapping up statement.

What follows here summarizes the features present in the test-takers' responses across three different parts of this organizational structure – introduction, direction-giving, and closing. Test-takers

of different English proficiency levels commonly attempted to use sequential adverbs. However, this organizational structure was particularly noticeable in the responses from the high proficiency group, suggesting that more proficient speakers are better able to organize their speech logically and coherently by using what they had learned. Their proper use of sequential adverbs can contribute to the clarity and effectiveness of communication, which are key aspects of language proficiency.

(1) Introduction

Several test-takers, particularly those who belonged to the high proficiency group, initiated their utterances by mentioning the distinctive features or advantages of the device that connect to the purpose of developing the device.

[High proficiency group]

- This device has a sensitive microphone, so it detects illegal logging. ...(Female 1)
- The device recorded the audio of rainforest for 24 hours. ... (Female 2)
- This device is attached with solar panels so we can get electricity from the sun. And you can easily handle the device through wireless internet service. And the way the device works is seen total four steps. ...(Male 1)

[Medium proficiency group]

- RFCx has the sensitive microphone. When someone illegally cut trees...(Female 3)
- RFCx is a device that recycle cellphones and gets electricity from the sun...(Male 2)

[Low proficiency group]

- This device is intended to protect forest from illegal logging, because solar panels are attached ... (Male 3)

Although they were a minority, a few mentioned the name of the device and its source or recognition.

[High proficiency group]

• The device is called RFCx. Let me explain how it works. First, the device detects logging… (Male 4)

• The device is called rainforest connection and it works this way. … (Female 4)

[Medium proficiency group]

• People are using this device. You need to recharge RFCx… (Female 5)

[Low proficiency group]

• Rainforest connection is made by our company and it makes energy from solar energy. So,… (Male 5)

(2) Direction-giving

First, sequential adverbs are words used to describe the order in which things happen or should happen. They help to organize narratives or instructions into a logical sequence. Examples of sequential adverbs include 'first,' 'then,' 'next,' 'after that,' 'finally,' and 'last,' etc. The use of sequential adverbs varied across proficiency levels. Specifically, as proficiency levels increased, there was a greater diversity and frequency of these adverbs, and their usage tended to be more appropriate. For example, among test-takers from the medium and low proficiency groups, there were numerous instances where individuals failed to distinguish between ordinal and cardinal numbers. Ordinal numbers are used to show

order or sequence in a list (e.g., first, second, third, etc.), while cardinal numbers are used for counting (e.g., one, two, three, etc.). It is essential to be able to differentiate between the two types when learning a language.

[High proficiency group]

- First, someone uses a chainsaw to perform illegal logging. Then the smartphone catches the sound. Second, this sound goes into the cloud over RFCx. Next, the cloud sends a signal to rainforest rangers. Last, rainforest rangers quickly go to the place and stop illegal logging. (Male 5)
- Initially, the sound of a chainsaw is detected by a solar powered mobile phone microphone. ... (Female 6)

[Medium proficiency group]

- ... Second, software sends a signal to cloud. Third, a nearby forest logger get a warning call right away. Last, the forest ranger goes to the scene immediately. (Male 6)
- One, the sound of an electric saw is directored(unidentified) to a solar powered mobile microphone. Two, the software sends a signal to cloud. Three, a forest ranger at a nearby site receives real time alert. Four, that rainforest ranger can go to the scene immediately. (Male 7)

[Low proficiency group]

- The RFCx picked up a sound of chainsaw by microphone solar powered cellphone. Then, software send a signal to cloud. Real time alert is received by a rangers on the around nearby. Last, the rangers go to the site immediately. (Male 8)
- The first logger logger. And the second is the signal to RFCx. The three, signal goes to the forest guard. Finally, the forest guard is on its way. (Female7)

Next, the syntax of imperative sentences not only reflects the test-takers' understanding of the topic at hand but also their awareness of the situational context and their readiness to actively participate in interactions with virtual interlocutors. What was interesting here was the difference between the high proficiency group and the other two less proficient groups;

no instances of direct imperatives were found among the highly proficient test-takers. Interestingly, even some test-takers from the low proficiency group who used imperative structures provided instructions that strayed from the context of the task situation.

[Medium proficiency group]

• Install rainforest connection on the tree. (Female 8)
• So please keep this device from leaves. (Female 9)

[Low proficiency group]

•. Attach it around you in front… (Male 9)
• Install this device. This device listen the sound of a tree… (Male 10)
• Don't cut down too many trees. (Female 10)

The findings here seem to align with those of Roever and Al-Gahtani's (2015) research. They examined how increasing proficiency impacts the pragmatic performance of ESL learners in Australia and found that as proficiency increased, learners used a broader variety of request formats. For example, beginners used imperatives and 'want-statements', while lower-intermediate learners added 'can', upper-intermediate learners introduced 'could'. Advanced learners used more complex expressions[37].

In a similar context, an examination of the modality used in the discourse of Task 3 provides valuable insights into the intentions of the speaker, who assumes the role of a company staff member. Modality in

[37] Although requests with 'can' and 'could' were still the most common type among advanced learners, if-clauses remained popular such as 'if you could help me,' 'if you could just do me a favor,' or 'if you don't mind,'.

discourse analysis refers to the linguistic resources used to express the speaker's various levels of certainty, obligation, permission, necessity, and likelihood towards the information being conveyed. Based on the information presented in Table 45, differences in the use of modality were revealed across proficiency levels, both in terms of frequency and type.

Table 45. Frequency of modality across three proficiency groups

Group / Modality	High (n=50)	Medium (n=50)	Low (n=50)
can	24	10	6
could	2	2	·
will	9	9	3
should	·	2	·
have to	·	1	·
Number of cases	35	24[38]	9

It was revealed that as the proficiency level of the test-takers increased, they utilized a broader array of modal expressions to communicate varying degrees of permission, obligation, and probability. The choice of auxiliary 'can' or 'could', which might suggest their polite language act, was also featured in that test-takers chose to use 'should' ,'need to' and 'have to' instead for the same communicative purpose of giving instruction and guidance.

[High proficiency group]

• …When the rangers receive a message about illegal logging, they could go to the site quickly and could prevent many trees from cutting. (Male 11)

•… So the cloud sends some sign to the rainforest rangers and you can locate

[38] Note that the two modalities, 'should' and 'have to,' which only appeared at the medium level, were used by the same test-taker.

them. Then you can go to the illegal logging spot immediately. (Female 11)

• ...Second, the rainforest ranger will be notified immediately. ... (Male 12)

[Medium proficiency group]

• ...If someone cut a tree, the sound will be recognized by the machine and transmit to you.... (Male 13)

• ... After then rangers can be prompted in action of spot and catch the illegal logging. ... (Female 12)

• ... It is recharged from solar panel, so you should installed in a sunny spot, but... Then you should also checks transmitting signal... (Female 13)

[Low proficiency group]

• ...Finally, you can go the place and arrest the illegal loggers. ... (Male 14)

• ... Four, that enable forest rangers can go to the scenes immediately. (Male 15)

• ... If you cut down the tree, the phone will signal. (Female 14)

(3) Closing

They concluded their explanation by highlighting the positive impacts of the device. This often encompassed detailing the benefits the device offers, particularly with regard to its impact on the environment.

[High proficiency group]

• ... they could go to the site quickly and could prevent many trees from cutting. (Male 16)

• ... RFCx will also work under the shade of tall and big trees in rainforest. Please trust our company. (Male 17)

[Medium proficiency group]

• ... I hope this device could be helpful for protecting rainforests and keep your patrol to. (Female 15)

[Low proficiency group]

• ...So, rainforest rangers can protect the rainforest. (Female16)

Overall, there is a case for the critical role of organizational and interactional features in spoken language, especially in tasks requiring detailed and sequential explanations. The findings highlight how proficiency can influence the use of these features in the given spoken discourse.

Moreover, the way that linguistic resources were deployed to elaborate the context for a given communicative purpose might have differed by English proficiency. For example, Female 5 intended to draw the attention of rangers to the item by mentioning its popular use in the first sentence. Other male and female test-takers also mentioned the strengths of RFCx—an invention best fit for the rainforest environment and easy to handle.

The analysis of linguistic resources across the three proficiency groups yielded findings that would provide support for the importance of teaching and learning target textual features (i.e., pronouns and connectives) and pragmatic features (i.e., imperatives and modal expressions) in English classrooms. Specifically, the distribution of modalities lays the groundwork for gauging student performance and measuring the language proficiency of test-takers. This ensures the test task aligns with its primary objective of being an achievement test.

4.5 Summary

This chapter sets out to prepare the way for making validation arguments through providing both quantitative and qualitative analyses.

First, the findings of the MTMM and MFRM analyses in the quantitative section of this dissertation showed the MARST's comparability with measures of the same speaking ability than those of other abilities (i.e., reading and writing), and the statistically insignificant mode effect on multiple aspects (i.e., test-takers' gender, region of origin, proficiency as well as task, rating criteria, and rater).

However, an interaction between the rater and the task was identified, although the bias was not considered substantially significant. The mixed-method approach used to examine the bias between the rater and task, based on the analysis of the scoring data of Rater 3, a native speaker, and the online interview with the rater, indicated that this interaction might have resulted from a decrease in rating severity by Rater 3 from Task 1 to Task 3 over time.

Next, the qualitative approach to investigate the nature of the spoken discourse produced via the MAR mode suggested that the affordances of the MAR mode that served as social cues could possibly promote test-takers' sensitivity to the presence of a simulated social being – rainforest rangers – , leading to their attempts to communicate not only informative but interpersonal messages. Moreover, the discourse analysis of test-takers' spoken responses appeared to confirm that the MARST had fulfilled its intended purpose as an achievement test. This conclusion is drawn from the observation that test-takers extensively applied what they had learned when formulating their responses.

In fact, Task 1 and 3 with simulated interlocutors were designed to examine whether a semi-direct (or simulated) speaking task delivered via the MAR mode could be a viable and sound alternative to direct speaking tests administered in the face-to-face interview mode. As a result, the inclusion of the monologic Task 3, which was the most challenging task requiring test-takers to provide procedural information to concerned individuals, was unlikely to result in an underrepresentation of the speaking construct being measured. Instead, it was expected to enhance the quality of the construct being measured in the test or potentially even broaden its scope.

In the subsequent sections of validation and discussion, a comprehensive analysis of these findings was conducted to strengthen the validation argument concerning the immersive nature of the MAR mode and its effectiveness in facilitating interaction in monologic speaking tasks with constrained responses. Furthermore, the study explored the significance of integrating technology, specifically, the MAR platform, into language learning and testing. This examination aimed to offer valuable insights in the context of EFL situations, where learners often lack daily immersion in an English-speaking environment and have limited opportunities to practice their speaking and listening skills outside the classroom.

Chapter 5. Validation

In this chapter, we relied on a combination of the two major frameworks generally accepted in the validation research of language testing, as specified in 2.5.2 and 2.5.3. The validation process entailed several steps, including the discussion of the target language domain, assessment records, test interpretations, decisions regarding test use, and consideration of consequences.

5.1 Validity argument

The use of the validation framework in the current study is to investigate the degree to which scores on the MARST can be interpreted as an indicator of L2 speaking proficiency for EFL high school classroom assessment. In other words, to present an argument for interpretation of scores on the MAR-mediated EFL speaking test while taking into account how the alternative mode affects the assessment procedure. The findings from the three research questions can serve as evidence to judge the validity of the MARST score interpretation.

The validity argument here is organized on the basis of the criteria of test usefulness that Bachman and Palmer(1996) defined : reliability, construct validity, authenticity, interactiveness and practicality. Justifying such assessment use indicates the term called AUA, which articulates rationales for the test's intended use, drawn from validity

considerations, the test design, and the validation process developed through analysis of the test purpose. According to Chapelle and Lee (2021), language testing research using a validity argument framework does not conclude that the test has been validated but rather that certain warrants have been supported (or not) to a certain degree and that other unfinished validation remains. Thus, factors on the negative side offer suggestions and possibilities for future empirical research such as a qualitative study looking at the strategies used during test-taking.

Following the process of building justification by Wang and colleagues (2012), which articulates AUA for the use of the newly developed MARST here comprises three parts : (a) claims and warrants that are desired for the intended test use; (b) supporting evidence that has been identified from both quantitative and qualitative analysis; and (c) potential rebuttals for further research.

Claims and warrants in this AUA are supported by backing or evidence gathered from the MTMM and factor analysis for test comparability, and MFRM for investigating significant factors that will likely affect the test construct, and inevitably the test validation. In the first place, to make a judgement about a test use, test users and stakeholders need to examine the evidence provided by the test developer about the test's intended use.

▶Intended test use: The MARST is primarily designed for relatively low-stakes speaking assessment, and its results are included in the

school grades at the end of a semester. It intends to provide high school students score reports that can be used to make decisions about their achievement of spoken communicative skills and the knowledge necessary for practical and academic settings in high school level in the Korean EFL context.

▶**Length and administration:** Each task time varies from 4 to 8 minutes, including preparation time with a maximum of 20 minutes. The test is administered via test-takers' mobile phones at the time and place of their convenience. In the absence of a human proctor or examiner, test-takers have their facial image of producing spoken responses video-recorded in real time during the test.

▶**Scores and scoring procedures:** All the constructed spoken responses, audio-recorded simultaneously with facial images recorded in real time for security checks, are assigned one of four grades in each rating criterion by three human raters, including one native speaker. Individual test-takers can check their scores on the screen of the mobile application to evaluate their strengths and weaknesses in terms of the rating criteria – fluency, accuracy and content – and make their own decisions about in what areas further improvement and practice are needed. The spoken responses are scored via the rating scale model (RSM; Andrich, 1978) and the test scores are analyzed by the MFRM, which calibrates the difficulty parameters of each item with different

facets (i.e., task, rater, three rating criteria) taken into account.

▶**Brief description:** The MARST is a mobile-based English language test consisting of four speaking tasks: dialogue completion, description of visual information, expressing one's opinions, and explaining sequential information. The mobile application, by which the test is self-administered, can be downloaded from the application store of a phone with an Android or an iOS system.

5.2 Analysis of target domain

The validity argument includes the "domain definition", meaning that the test tasks are developed based on an adequate domain analysis to obtain relevant observations of performance.

▶**Claim 1**: Observation of performance on the MARST reflects the TLU domain of general English.

[Warrant 1] Observation of performance on the MARST reveals relevant language functions and knowledge that students are required to obtain in a Korean EFL high school classroom.

[Evidence 1a] According to the national curriculum for high school English subjects in Korea, required language functions and skills are clearly stipulated as part of achievement standards of English speaking that high school students need to obtain for various spoken

communicative needs in both daily and academic settings. Thus, in the test design step, the test tasks were developed based on analysis of relevant achievement standards and instructional goals in the national curriculum and textbooks that are published in compliance with the specifications of the national curriculum. The featured language functions identified in this analysis include “making suggestions”, “expressing one’s abilities” in Task 1, “describing visual/factual information”, “expressing personal opinions” in Task 2−1 and 2−2 respectively, and “explaining procedural information” in Task 3. These are important skills that represent multiple and interrelated dimensions that constitute the L2 speaking proficiency needed for general and academic settings in EFL high school courses.

[Evidence1b] The characteristics of each individual task were reported in the Methodology test design section. The characteristics of the tasks that test−takers will perform in the test are clearly specified. Linguistic and discourse features of test−taker output are related to key assessment features specified in the analytical rating scales: lexico−grammatical range and accuracy, fluency, and content.

5.3. Assessment records: Evaluation and generalization

A claim concerning assessment records is supported by warrants about consistency across tasks, test forms, occasions, and raters (Bachman & Palmer, 2010). Consistency of assessment records means

the extent to which test-takers' performances on different assessments of the same construct yield similar assessment records. Test-takers' performance and their scores are always affected to some extent by irrelevant or unexpected factors. Thus it is necessary to minimize the effects of those potential sources of inconsistency that are controllable to some extent.

In Kane's interpretation/use argument, a claim about assessment records corresponds to the second and third links- "evaluation" and "generalization" inferences. An evaluation inference concerns the quality of the scoring procedures for accurately summarizing the relevant aspects of performance(Chapelle & Lee, 2021). A generalization inference which links the observed scores to expected scores indicates whether observed scores are reliable estimates of expected scores over relevant tasks and consistent with and across raters.

▶Claim 2: Test-takers' performance on the MARST is evaluated adequately to yield observed scores reflective of speaking ability levels.

[Warrant 2-1] Rating procedures of the MARST are appropriate for raters to assess targeted speaking abilities.

[Evidence 2-1] Rubric development and rater training were judged to support the warrant for the evaluation inference although some limitations and challenges derived from the findings in the statistical analysis later called for future improvement. Unlike conventional speaking tests, mobile-based tests are characteristic of immediacy and

autonomy which means they are learner-oriented. To serve this purpose, analytic scales are useful. The scale criteria reflect multiple and interrelated aspects of the targeted speaking construct based on the target domain analysis in Claim 1, and the criteria supply diagnostic descriptors so that test-takers can identify individual strengths and weaknesses.

In multiple training sessions, raters spent most of their time discussing the features that differentiated adjacent levels. To avoid rater bias, raters compared each other's or the researcher's rating outcome to analyze what caused score gaps between raters. Prior to rater training, several responses were sampled that corresponded to each descriptor level. After that, rater training sessions resumed, with the majority of time spent discussing the features that differentiated adjacent levels and letting raters try scoring audio samples appropriately. To avoid rater bias, raters compared their ratings with others or the researcher's rating outcomes to analyze what caused the score gaps between raters. Easy access to the rating rubric in the mobile application during the rating process also helps human raters to rate without bias.

[Warrant 2-2] Test administration conditions of the MARST are appropriate for providing evidence of targeted speaking abilities.

[Evidence 2-2a] The mean score of the questionnaire Item 7 on clarity (2.93 out of 4) indicates that a majority of test-takers experienced little difficulty understanding the instructions or time pressure when performing the tasks through the MAR device. In addition, the mean score

of Item 8 on relevance (3.02 out of 4) indicates that they understood the English speaking skills that the MARST tasks intend to assess. Teachers' mean scores on the corresponding items were higher than that of test-takers: 3.35 and 3.53 respectively.

[Evidence 2-2b] In the interviews, raters reported that they used the rubric to indicate different speaking ability levels, – demonstrating positive attitudes of the clarity of what is stated in each descriptor of the rubric.

[Potential rebuttal 2-2] In the interviews, some test-takers expressed concerns about the difficulty of ascertaining if a test-taker is cheating or if meaningless actions could be mistaken for cheating. This highlights the need for more advanced and sophisticated mobile technology when administering the MARST for a higher-stakes purpose.

▶Claim 3 : Observed scores of the MARST are reliable estimates of expected scores, that is the scores test-takers would expect to obtain across different assessment tasks, coupled with different aspects of the rating procedure, with consistency across different groups of test-takers.

[Warrant 3-1] Test, task, and rating specifications are well-defined so that parallel tasks and test forms can be created.

[Evidence 3-1] To collect evidence for evaluating reliability, one of the qualities of usefulness, an analysis of the characteristics of each task was conducted, as mentioned in 2.4. in the Literature Review section and in 3.1.4. in the Methodology section. Detailed tasks specifications are

explicitly described, including the characteristics of the setting, test rubrics, the input, expected responses and the relationship between inputs and responses. The rating rubrics with detailed descriptors are available in Appendix 1.

[Warrant 3-2] Scoring of test-takers' performance in several test tasks is consistent within and across raters.

[Evidence 3-2a] The reliability of the MARST and the questionnaire was .94 and that of the questionnaire .93, respectively. In the test design, holding a couple of rater training sessions before scoring should be regarded as an essential measure to ensure a reliable test. With this measure, it is possible to control potential sources of inconsistency.

[Evidence 3-2b] In the MFRM analysis, the fit statistics of the rater facet indicate that raters used the rating scale consistently and maintained severity across test-takers, tasks and criteria. In Table 35 in the result section, the infit and outfit MnSq range of 0.80 to 1.06 and 0.80 to 1.05, respectively are both within the productive range of 0.5 to 1.5 for measurement, thereby producing no overfit or misfit. In other words, the four raters did not exhibit more variation in their ratings than expected so that their ordering of test-takers by ability is consistent with the estimated ability measures of those test-takers. Therefore, it can be said that they are able to use the rating scale consistently across tasks and test-takers.

[Evidence 3-2c] According to the fit statistics of the task facet in the MFRM analysis, the high task reliability index (0.98) indicates the degree

to which tasks are replicable in terms of difficulty is sufficiently high to assign the same task to another sample population with comparable ability levels.

[Evidence 3-2d] According to the fit statistics of the test-taker facet in the MFRM analysis, the separation and reliability indices for the difference in test-taker ability is high (4.32 and 0.95 respectively), which means the MARST separates 194 test-takers into at least four statistically distinct levels or strata in terms of the ability being measured. The high reliability suggests that if test-takers took another speaking test, the ordering of test-takers by ability would likely be the same. Such a result means we can have great confidence in the consistency of score-based inferences (Bond and Fox, 2007).

[Potential rebuttal 3-2] There are several potential sources of test-taker ability misfits and bias detected from statistically significant rater and task interaction (Rater 3 in Task 1 and 3). It was suspected from interviews and the inspection of response data that these unusual responses were due to construct-irrelevant factors, such as inattention, misunderstanding the test instruction, or the failure to maintain test-taker's motivation. The rater and task interaction was highly likely to result from the rating behavior of Rater 3, affected by the rating order from Task 1 to Task 3 as his rating severity went weaker over time.

[Backing] However statistically significant, the bias values are not generally considered substantial.

5.4 Test interpretations: Explanation and extrapolation

The meaningfulness, impartiality, relevance and sufficiency warrants all pertain to the construct that the test is intended to assess. The meaningfulness and generalizability warrants refer to the authenticity of task and performance. An "explanation" inference is included in the validity argument to link the score and the language ability construct.

5.4.1 Meaningfulness

Test result is an indicator of the construct, that is, some aspects of language ability to be measured. Thus, when someone gives a language assessment, (s)he intends to interpret the performance on this assessment as an indicator of some aspect of the individual's language ability. Interpretations about the ability to be assessed are meaningful with respect to a learning syllabus, an analysis of the abilities to perform tasks in TLU domains, and the general theory of language ability. Moreover, the concept of meaningfulness also implies test developers have a responsibility to ensure that the labels used to describe the ability to be assessed are understandable.

▶Claim 4: Test scores are meaningful indicators of student' achievement in the course.

[Warrant 4-1] The procedures for administrating the test enable test-takers to perform at their highest level and to demonstrate their English

proficiency for communicating in secondary school-level EFL academic settings.

[Evidence 4-1a] The topic, language forms and functions, and expressions to be assessed in the test are exclusively elicited from what is covered in test-takers' English classes.

[Evidence 4-1b] Test-takers' feedback collected from the questionnaire and interviews suggests that the authentic and engaging features of the MARST test, test administration at test-takers' own discretion, and recording their responses more than once enables them to perform at their best with low test anxiety. Further, a sufficient amount of input and clear guidance give them ideas on how to perform the tasks. On top of this, they find test tasks relevant and necessary for high school students to learn. Lastly, many of them agree the test tasks are appropriate to be presented in the MAR mode.

[Evidence 4-1c] The qualitative analysis shows that test-takers' speaking responses highlight the use of targeted language features across all proficiency groups although the degree of their mastery varies at different proficiency levels.

[Warrant 4-2] The separate analytic rating scales of the MARST contribute to a common construct of "speaking ability" to be assessed.

[Evidence 4-2] According to the fit statistics of the rating criteria facet in the MFRM analysis, the three rating criteria fits are within the acceptable range of 0.5 to 1.5. None of the rating criteria were misfitting or overfitting. With misfits, a criterion would not form part of the same

dimension as other criteria of the rating scale. Meeting the assumption of unidimensionality, the ratings on one criterion correspond well to ratings on other criteria, that is to say, the ratings on each criterion converge into a single pattern of proficiency across all criteria. With overfit, on the other hand, a criterion is considered redundant or measuring the same ability as other criteria or significantly affected by the scores assigned to other criteria, which means there is a halo effect (Eckes, 2005; McNamara, 1996).

[Potential rebuttal 4-2] According to the fit statistics of test-taker's ability facet in the MFRM analysis, of the 194 test-takers analyzed, 12(6.1%) were identified as misfitting, which does not satisfy McNamara's(1996) guideline of below 2%.

[Backing] One source of test-taker misfit, however, lies in the few observations per test-taker. Particularly, Rasch model, which treats rating criteria as "different items", treats different scores assigned on each rating criteria as departures from expected patterns. Therefore, according to what Bonk and Ockey argue, test-taker misfit is, in fact, not as serious of a problem as one might expect.

[Warrant 4-3] There is no task that is redundant or in need of revision and deletion.

[Evidence 4-3] According to the fit statistics of the task facet in the MFRM analysis, Table 34 in 4.3.1.2 reports both infit and outfit MnSq values are within the productive range of 0.5 to 1.5 for measurement, producing no overfit or misfit. No misfitting or overfitting task indicates

little chance of tasks being either poorly or perfectly good. Each task forms part of a set of tasks that together define a single measurement trait (McNamara, 1996), thus providing information that the other tasks do not since there are no overfitting tasks.

[Warrant 4-4] Test-takers perform significantly differently in the various aspects of speaking.

[Evidence 4-4a] According to the fit statistics of the rating criteria facet in the MFRM analysis, the three rating criteria exhibit significantly different degrees of difficulty ($\chi^2(2)=169.4$, $p=.00$). The analysis shows that the rating criteria are distinguished into 7 to 9 distinct levels of difficulty with a high reliability index of 0.98.

[Evidence 4-4b] In Figures 22 to 24 in 4.3.1.4. of the Results, the average ability measures across the rating categories increase as the score level progresses. The same is true of the expected measure for each score level which refers to the ability measure that the Rasch model would predict if the data were to fit the model. Thus, as expected, it can be concluded that the scale is positively linked to progression of test-taker speaking ability and the rating scale functions.

[Evidence 4-4c] In Figure 22 to 24 in 4.3.1.4. of the Results, the outfit MnSq range determines whether the scale levels are reliable based on the expected model value of 1.0, indicating the equal observed and expected test-taker ability measures. The outfit MnSq range of this data sample is 0.7 to 1.2, which is not far from 1.0. If it is greater than 2.0, the rating criterion may not be a meaningful measurement (Linacre, 1999).

Overall, the distance between the scores is discernable, even though the *Content* criterion has thresholds between adjacent score levels that are a bit narrower than expected.

[Evidence 4-4d] The probability curves for the rating scales in Figures 25 to 27 in 4.3.1.4. of the Results, indicate that each rating scale has separate peaks on each score level; these refer to the most probable score choice for test-takers across the aspects of the ability being measured.

[Warrant 4-5] Score reports are user-friendly in terms of accessibility and language.

[Evidence 4-5] Test-takers and test users can see the score reports on the app screen once human raters assign ratings to individual test-takers after listening to audio files on the app, which records spoken responses. The score report is designed in a simple format where rating categories are listed with essential words signifying four different competency levels. Therefore, test-takers can easily ascertain which areas they are successful.

5.4.2 Impartiality

In test validation, interpretations about the ability to be assessed are impartial to groups of test-takers.

▶ **Claim 5** : All aspects of the administration of the assessment are free

from bias that would favor or disfavor some test-taker groups.

[Warrant 5-1] Test tasks are based on the course content and test-takers are notified about the test and what it will cover in advance.

[Evidence 5-1] Comparing the course syllabus with the test specs and with the actual test itself shows which test tasks are based on which parts of the course content, and with the actual test itself. Moreover, prior to test administration, a workshop was held for students. Teacher notes and handouts provided to students during the workshop informed them as to what the test would cover and how the learning goals and contents were represented as test tasks by using the mobile application.

[Warrants 5-2] Test-takers have equal access to the test, in terms of location, and familiarity with conditions and equipment.

[Evidence 5-2] The mobile-based test makes it easy for students with disabilities to access the input of an assessment task or perform the task. Unequal test access that may result from travel expenses and lack of unfamiliarity with the test equipment is addressed by the MARST because test-takers are allowed to use their own mobile phones and choose the test location at their convenience.

[Potential rebuttal 5-2] Any technical malfunction that occurs on test-takers' mobile devices might affect their performance during test administration.

[Backing] Malfunctioning devices can be replaced with alternative ones available in school; in fact, there were a couple of cases that the app did not work on some of students' mobile phones. At that time, their mobile

phones could be replaced with tablets provided at school which had upgraded models or a different operating system. Also, internet disconnection could be easily addressed because the MARST did not have to be administered at the same location and time.

[Warrant 5-3] Interpretations of the test construct are consistently meaningful across different groups of test-takers.

[Evidence 5-3] The bias analysis in Section 4.2.3. of the Results suggests that test-takers' gender and location did not affect rater behaviors. That is to say, raters' rating behaviors did not vary across different gender and regions of residence.

5.4.3 Generalizability

Generalizability refers to the degree of correspondence between a given language assessment task and a TLU task in their task characteristics, which is linked to authenticity. When interactions between test-taker and test task is of interest, however, generalizability pertains to interactiveness.

The validity argument is associated with the "extrapolation" inference of the score a test-taker would obtain in the domain of interest. In other words, an extrapolation inference is about constructs defined as performance on particular task (Chapelle and Lee, 2021). The extrapolation inference requires empirical evidence that test scores are highly correlated with scores on criterion measures, which can be either

test measures or non-test measures (Xi and Sawaki, 2008).

▶ **Claim 6**: Interpretations of the ability to be assessed are generalizable to the TLU domain.

[Warrant 6-1] The task characteristics of the MARST correspond to those in diverse real-life English-medium settings for both academic and daily communication purposes.

[Evidence 6-1a] The section 3.1.2. of the Methodology indicates that the task characteristics are supposed to reflect those of the TLU domains. To promote authenticity, which is comparable to generalizability, the four tasks commonly deal with the most concerning global issue of environmental crisis which people across the world are bound to experience in both casual and school class settings. A simulated addressee to whom a test-taker speaks in each task – a friend, a polar bear and a rainforest ranger further motivates test-takers to speak.

[Potential rebuttal 6-1a] No evidence was found in the results concerning the correspondence of the MARST scores and any speaking non-test performance across English and disciplinary courses at school.

[Backing 6-1a] The corpus analysis of test-takers' response data to Task 3 was conducted in reference to American English corpus present in AntConc 4.2.0(Anthony, 2022) by dividing it into three proficiency groups. The linguistic features of the target corpus were goal-oriented and spoken in a formal speaking style, which was considered appropriate for an IT company staff to explain the sequence of how an invented

device works for protecting rainforests from illegal logging to potential users of rainforest rangers. These might include the appropriate use of grammatical knowledge (i.e., lexical diversity and syntactic complexity) to language functions and pragmatic knowledge to effectively communicate the sequence of events and serve its communicative purposes such as direction and persuasion. And the evidences were able to discriminate test-takers according to proficiency. It can be mentioned that the MARST did not underrepresent the English speaking constructs that were integral to measure the communicative competence in the EFL contexts.

[Evidence 6-1b] The features of the MAR in this test include mixed- or multi-modal presentation of information (integrating both virtual and real-life or both visual and verbal), socialization via simulated human-like embodiment, and animation processing functions supporting the increase of test authenticity by conveying the test input which offers various simulated situations that feel as close to real-life as possible during test performance. According to Section 4.3, which explores perceptions of the MAR mode, a majority of respondents, including test-takers and teachers, consider the test mode of the MARST has positive potential in terms of task authenticity and interactiveness. Relative to other conventional speaking test, a high degree of correspondence of the features of the MARST tasks to those of a TLU task help score interpretations generalize beyond the test to language use in the TLU domain.

[Potential rebuttal 6-1b] Even though the researcher has put great effort into creating an authentic environment, there remains room for improvement in developing test tasks close to the simulated TLU situations due to the technological limitations associated with the affordances of the device software that creates and delivers the lifelike simulations. Some comments from test-takers and teachers point to the unnatural and monotonous voice tone of the virtual interlocutors on the app. They even cannot respond to test-takers in an impromptu manner.

[Backing 6-1b] There is a theoretical argument that one of the unique features of the AR mode – semantic congruity or proximity – enhances transfer as it takes place when the learning and application environment are similar. To put it another way, the MAR mode makes it easier for students to put the knowledge and skills that they have learned in the classroom setting into actual practice or real life.

Empirical evidence from the survey and interview also found the use of the AR mode effective for test-takers to better perceive similarities between the test context created by AR and real-life situations. Therefore, it can be said that transfer will be more likely to occur with the use of AR, which strengthens the assumption of extrapolation inference to some extent.

Moreover, the fact that the MAR app allows test-takers to revise their responses as they want within a specific period of time makes sure that students are given 'control over the assessment' or 'autonomy' with their learning being further reinforced during the assessment. Thus, it

can be said that the new mode helps the speaking test fulfill its purpose as achievement assessment or what is called 'learning-oriented assessment' that evaluates the mastery of skills and knowledge that are necessary for students to learn in the curriculum.

[Warrant 6-2] The test tasks engage not only test-taker's areas of language knowledge but also affective schemata and topical knowledge.

[Evidence 6-2a] The theme of "global environment in crisis" is a test topic highly close to test-takers' real-life experiences it is one of the most frequently covered news topics globally and many closely feel the devastating consequences of the climate change in everyday life. Thus, it can cause emotional responses that facilitate language use, stimulating the so-called "affective schemata" of test-takers. Through the test tasks can they express how strongly they feel and what they know and think about the issue.

[Evidence 6-2b] The socialization feature of the MAR mode helps to lower test anxiety that test-takers may feel from face-to-face interaction with examiners in an oral interview format. This alternative mode can serve as a measure to prevent limiting their language ability due to negative affective responses that may inhibit test-takers' speaking performance.

[Evidence 6-2c] MARST task development is based on a topic-centered design, which means test-takers are required to employ some topical knowledge as well as language knowledge in their test performance. According to the test specifications, the AR-medicated

warm-up activity before Task 1 activates test-takers' real-world knowledge about recycling in daily life, on which their responses are based on. Task 2 expects test-takers to employ their topical knowledge about the causes and effects of global warming with the AR-mediated animation activating test-takers' prior knowledge. Task 3 requires test-takers to retrieve the meanings of some technical terms or subject-specific vocabulary from what they have learned in class and use them.

Although the topic of environment is not a scoring criterion in the MARST and the degree of individual interest in the topic may vary, efforts were made in the English class to minimize the difference in familiarity with the topic; for example sufficient time was spent in discussing the topic with reading relevant materials and communicative activities such as a role-play.

[Warrant 6-3] The test results are comparable to those of other direct speaking test scores to some degree.

[Evidence 6-3] In the MTMM analysis in Section 4.1.1. of the Results, the correlation matrix indicates that the MARST scores have positively moderate correlations with other speaking measures such as oral interviews (.605) and oral translations (.573).

5.4.4 Relevance and sufficiency

Relevance refers to the degree to which the interpretation provides the sufficient information for the decision-maker to decide. Sufficiency

refers to the degree to which the interpretation provides enough information for the decision-maker to decide (Yang, 2021).

▶ Claim 7 : Interpretations about the ability to be assessed are relevant to the decisions made at educational institutions.

[Warrant 7] The characteristics of the MARST tasks highly reflect instructional tasks in EFL speaking courses.

[Evidence 7] The MARST is designed as an achievement test and part of the end-of-semester grades. Thus, it is natural that the test tasks are similar to tasks already dealt with in the classroom. The contents in the textbook provide teachers and students important language functions, skills and knowledge that reflect the achievement standards of the national curriculum of high school English subjects.

▶ Claim 8 : Score interpretations about the ability to be assessed are sufficient for the decision to be made.

[Warrant 8] The MARST can provide sufficient testing contexts to collect the evidence need to infer test-takers' speaking ability.

[Evidence 8] The multimodal feature of the AR mode makes it possible to integrate virtual and real-world materials. Such semantic proximity of the AR mode allows for more sensitive and accurate assessment of learner knowledge by increasing authenticity in multi-modal materials used for various test contexts and by motivating test-takers to engage in deeper processing. In short, the adoption of the AR mode in the

MARST assesses a test-taker's language ability in more various simulated contexts, generated by integrating the virtual and real-world environments. Therefore, the number of tasks that the MARST covers in a single administration and the time it takes is considered to be not less than other conventional speaking tests that measure different aspects of speaking proficiency in different task types.

5.5 Decisions and test use: Utilization

▶ Claim 9 : The test scores are useful for meeting the test purpose and making decisions about test-takers' English speaking competence.

[Warrant 9-1] The test scores are useful for determining the extent to which test-takers have mastered English speaking skills and knowledge as part of the school academic achievement.

[Evidence 9-1] The test scores, which are based on test-takers' performance on the given tasks, inform test-takers on achieving English spoken language standards required for high school students in the Korean EFL setting, based on test taker's performance on the given tasks. Good performance on the MARST implies that the test-takers are ready to accomplish the authentic tasks that require them to make suggestions to friends, describe visual information, express personal opinions about environment issues and explain the working procedure of a device.

[Warrant 9-2] The MARST is sensitive to and takes into consideration local educational and societal values.

[Evidence 9-2] Developing the MARST was an ambitious attempt to overcome some constraints of the EFL language learning and teaching, namely – decontextualized test settings, rating bias, reluctance to assess speaking skills in EFL language assessment, and low practicality.

[Potential rebuttal 9-2] Existing educational and societal values may still view the MARST's administration method as inappropriate in Korean society. First of all, assessment via the MARST is student-centered, allowing them to determine the test location and time. Moreover, unlike most typical assessments, which are practiced in a single trial by one or joint examiners in one designated place with testing equipment supplied, the MARST gives a second chance to take the test within the specified time limit of 20 minutes. Thus, it would be hard to reach a consensus among community members to administer the MARST for high-stakes test purposes.

▶ Claim 10 : Decisions on the basis of the score interpretations are equitable for stakeholders.

[Warrant 10-1] The decisions are not affected by the personal attributes of the assessor (raters) such as ethnicity, gender, age or socioeconomic status.

[Evidence 10-1] Given that the testing conditions do not include direct interactions between test-takers and a human examiner or interlocuter, there would be no bias derived from raters' personal characteristics.

[Warrant 10-2] Test takers have equal opportunities to learn or acquire

the ability to speak in their EFL speaking courses.

[Evidence 10-2] Before the test, test-takers received instruction and information on what the test covered and sufficient preparation for their test performance.

5.6 Consequences

Significantly less evidence has been presented for the claims regarding decisions and consequences. One reason may be that arguing and supporting the claims about assessment records and score interpretations is viewed as the most typical responsibility of test developers.

The purpose of the MARST is to decide the extent to which high school students have achieved English speaking knowledge, skills, and abilities at the high school level in the Korean EFL setting where English is not a direct means of communication in everyday life.

▶ Claim 11 : The consequences of using the MARST and decisions based on MARST scores are beneficial to test-takers and the stakeholders (i.e., teachers and parents).

[Warrant 11-1] The impact on test takers (i.e., the amount or type of test preparation, experience of taking the test, or perceptions of feedback), instruction, educational systems and society is promising.

[Evidence 11-1a] In the MARST test survey and interviews, there were

comments from both test-takers and teachers about the alternative format to indicate a positive test experience. It was reported that the MARST tasks were useful for creating communicative contexts close to real-life and encouraging test-takers to use their speaking language skills.

[Evidence 11-1b] Raters' perceptions toward and their use of the rating scale during the rating process were explored. Overall, they perceived the scale to be clear and comfortable to use.

[Potential rebuttal 11-1b] The raters are not homogenous in their levels of experience. The result of the rater facet analysis in the MFRM also found that the difference between the most severe and least severe rates in the fair averages was 0.19. The fixed Chi-square statistic ($\chi^2 = 29.4$, $df = 3$, $p = .00$) and the separation index of 2.46 with a reliability of .86 suggest that the four raters exercised approximately two and a half statistically distinct levels of severity. Further, significant interactions in several cases were detected between two raters and two tasks. Thus, the rating severity across raters is neither equivalent, nor is within-rater rating consistency guaranteed. All in all, the assumption of rating consistency is not sufficiently supported.

[Warrant 11-2] The MARST is practical as designed, developed and used within the limits of existing resources.

[Evidence 11-2] The MARST developed for this study is expected to save material resources such as space for test development and administration, equipment, and time for development and tasks. That said, the AR app is

expensive to develop. Moreover, online-based language assessment as discrete occasions entails a lot of stakes, given that it involves a system where the database of test tasks and test-takers are accumulated and analyzed and fed back to test-takers. However, it is effective in helping make a valid inference about test-takers' ability and growth.

[Warrant 11-3] The MARST may hold the potential for positive washback by allowing English language learners to regularly access to technology.

[Evidence 11-3a] In the post-test interview, test-takers reported greater satisfaction with the option available in the mobile app that they could improve their answers by being allowed to review and revise their answers within the specific period of time.

[Evidence 11-3b] Not only can test-takers revisit previous tasks, but teachers can also access to test-takers' records at their convenience, which may offer insights about second language learners' learning growth.

[Evidence 11-3c] Easy accessibility to online materials presented in hand-held devices such as a mobile application will develop digital literacy in the target language, which relates to one of the general objectives of the 2022 revised national curriculum.

5.7 Summary of the validity argument

Drawing on the validity framework, the current study collects and integrates evidence to evaluate the usefulness of a newly-developed

speaking test and explores the possibility of integrating a new technology of MAR into L2 assessment. The validity argument for the MARST is summarized in Table 46 with 11 claims, 24 warrants, and 33 pieces of evidence in total, although 9 potential rebuttals and 4 counterclaims to rebuttals(backing) are not included.

In this framework, one finding of high interest is that the MAR technology intersects with various inferences, possibly affecting interpretation and test use. In particular, functions of the MAR mode serve to enhance a range of inferences in various test usefulness criteria, including assessment records, test interpretation, decision/use, and consequences.

With respect to assessment records, Warrant 2-2a is supported by the evidence collected from the qualitative analysis that shows that owing to the easy access to the rating rubrics from the mobile application, test users are well aware of the rating criteria and their descriptors in the rubric. Thus, a technology (mobile)-related feature can contribute to making administration conditions of the MARST appropriate for providing evidence of targeted speaking abilities.

Another example can be taken from the test interpretation. One inference states that interpretations about the ability to be assessed are generalizable to the TLU domain. For this, it should be warranted that the task characteristics correspond to those in diverse real-life settings for both academic and daily communication and the tasks engage not only test-taker's areas of language knowledge but also affective schemata

and topical knowledge. Here the features of the MAR, including mixed-or multi-modal presentation of information, socialization via simulated human-like embodiment, and animation processing functions increase test authenticity by offering various simulated situations close to real-life during test performance.

Moreover, the theme of "global environment in crisis", one of the most frequently covered news in the world, draws emotional responses that stimulate test-takers' language use and affective schemata. A simulated addressee to whom a speaker talks to in each task – a friend, a polar bear and a rainforest ranger further motivates test-takers to speak, although they may affect test-takers' performance to a different degree. Indirect interactions with examiners, which is a feature of the MAR mode, not only help to lower the anxiety that test-takers may feel from face-to-face interactions in oral interviews, but also ensure that the administration of the assessment is free from bias that may favor or disfavor certain test-taker groups.

The positive test experiences reported by test users according to the surveys and interviews suggest that using the MARST and decisions based on MARST scores can benefit both test-takers and stakeholders. Potential test use for improving EFL teaching and learning practices was also reported among test users.

On the other hand, although the strengths of the MAR mode outweigh its weaknesses there are some concerns as to the test validity of the MAR mode. Some technical issues are involved in assessment

records, test interpretation and test use ; for example, a failure to run the applications due to incompatible devices, unnatural test materials presented via the MAR channel, and the absence of preventive measures against cheating, which may be resolved by further progress in the field of mobile-based AR technology.

All in all, the integration of MAR technology in L2 speaking assessment may not substantially affect test scores and the internal structure of the test score variance. However, with the use of technology the validity argument of speaking assessment can be further strengthened by evidence that would have been less sustainable otherwise. The MARST had a positive impact on test-takers' cognitive and affective aspects and created sufficient and meaningful opportunities or testing contexts to judge the extent of the test-takers' mastery of the target speaking construct.

Table 46. Summary of articulating the validity argument of the MARST test use

Claim	Warrant	Evidence	Judgement of degree of support
1. Observation of performance on the MARST reflects the TLU domain of general English.	**1.** Observed performance reveals relevant language functions and knowledge in the Korean EFL high school classroom.	**1.** In the test design step, the test tasks were developed based on an analysis of relevant achievement standards and instructional goals in the national curriculum and textbooks, which are published in compliance with the specifications of the national curriculum.	Supported
2. Test-takers' performance on the MARST is evaluated adequately to yield observed scores reflective of speaking ability	**2-1.** MARST rating procedures are appropriate for raters to assess the targeted speaking abilities. **2-2.** The MARST	**2-1.** In multiple training sessions, raters spent most of their time discussing the features that differentiated adjacent levels. To avoid rater bias, raters compared each other's or the researcher's rating outcome to analyze what caused score gaps between raters. Easy access to the rating rubric in the mobile application during the rating	Partially

levels.	test administration conditions are appropriate for providing evidence of targeted speaking abilities.	process helps human raters to rate without bias. **2−2.** The characteristics of each individual task are reported in the test design section of the Methodology.	supported
3. Observed scores of the MARST are reliable estimates of expected scores.	**3−1.** Test, task and rating specifications are well−defined. **3−2.** Scoring of test−takers' proficiency is consistent within and across raters, as is rater severity.	**3−1.** Detailed tasks specifications are explicitly described including the characteristics of the setting, test rubrics, the input, expected responses and the relationship between input and responses. The rating rubrics with detailed descriptors are available in Appendix 1. **3−2.** In the MFRM analysis, the fit statistics of the rater facet indicates that raters used the rating scale consistently and maintained severity across test−takers, tasks and criteria. The infit and outfit MnSq range of 0.80 to 1.06 and 0.80 to 1.05 respectively, are both within the productive range of 0.5 to 1.5 for measurement, producing no overfit or misfit. According to the fit statistics of the task facet in the MFRM analysis, the high task reliability index (0.98) indicates the degree to which tasks are replicable in terms of difficulty is high. In the MFRM analysis on the test−taker facet, the separation and reliability indices for the difference in test−taker ability are high at 4.32 and 0.95 respectively. In the analysis of unusual responses in MFRM, it was evidenced that potential sources bias might have included the lack of attention and misunderstanding of the task instruction. Also, the follow−up interview after bias analysis indicated the effect of rating order upon Rater 3 who rated Task 1 and Task 3.	Partially supported
4. Test scores are meaningful indicators of students' achievement in the course.	**4−1.** The administration procedure enables test−takers to perform to the best of their ability to demonstrate English proficiency in EFL settings. **4−2.** The separate analytic rating	**4−1.** The topic, language forms and functions, and expressions in the test are exclusively elicited from test−takers' English classes. Test−takers' feedback from the questionnaires and interviews suggests that the authentic and engaging features of the MARST test, the test administration at test−takers' own discretion, and recording their responses more than once allow them to perform at their best with low test anxiety. Additionally, the sufficient input and clear guidance give them ideas on how to perform the tasks. They find test tasks relevant and necessary for high school students to learn. Many find the test tasks appropriate to be presented in the	Supported

	scales contribute to a common target construct without redundant criteria.	MAR mode. The qualitative analysis of test-takers' responses reveals the presence of specific language forms and functions across all proficiency levels, despite varying degrees of mastery.	
4. Test scores are meaningful indicators of students' achievement in the course.	**4-3.** There is no task redundancy or need for revision or deletion. **4-4.** Test-takers perform significantly differently in the various aspects of speaking. **4-5.** Score reports are user-friendly in terms of accessibility and language.	**4-2.** In the MFRM analysis, the three rating criteria fits are within the acceptable range of 0.5 to 1.5. None of the rating criteria were misfitting or overfitting. **4-3.** In the MFRM analysis, both infit and outfit MnSq values in the task facet were all within the productive range of 0.5 to 1.5, producing no overfit or misfit. No misfitting or overfitting task indicates little chance of tasks being poorly made or perfectly good. **4-4.** In the MFRM analysis, each rating scale has separate peaks on each score level. The average ability measures across the rating categories increase as the score level progresses. As expected, the scale is positively linked to progression of test-taker speaking ability and the rating scale functions as expected. **4-5.** Test-takers and users can see the score reports on the app screen once human raters score individual test-takers' audio files on the app, which records spoken responses.	
5. All aspects of the administration of the assessment are free from bias that may favor or disfavor certain test-taker groups.	**5-1.** Tasks are based on the course content and test-takers are notified about the test in advance. **5-2.** Test-takers have equal access to the test. **5-3.** Interpretations of the test construct are consistent across different groups of test-takers.	**5-1.** Prior to test administration, a workshop was held that guided users on how to use the mobile app. Teacher's notes and handouts explained what the test would cover and how the learning goals and contents were represented in the test tasks. **5-2.** The mobile-based test makes it easy for students with disabilities to access the test input and addresses the issue of unequal test access, which may result from travel expenses. Lack of familiarity with the test equipment can be addressed because test-takers in the MARST use their own mobile phones and decide test time and location at their convenience within a specified period. **5-3.** The bias (interaction) analysis of the MFRM analysis indicates that raters' rating behaviors did not vary across different gender and location of residence.	Supported
		6-1. The four tasks commonly deal with the most concerning issue of global warming, which people across the world experience. However, due to technological limitations, some	

6. Interpretations of the ability to be assessed are generalizable to the TLU domain.	**6-1.** The task characteristics correspond to those in diverse real-life settings for both academic and daily communication. **6-2.** The tasks engage not only test-taker's areas of language knowledge but also affective schemata and topical knowledge. **6-3.** The test results are comparable to those of other direct speaking test scores to some degree.	comments from test-takers and teachers point to the unnatural and monotonous voice tone of the virtual interlocutors on the app. The features of the MAR including mixed- or multi-modal presentation of information, socialization via simulated human-like embodiment, and animation processing functions support test authenticity by offering various simulated situations close to real-life during test performance. The corpus analysis of speaking response data of Task 3 supported that linguistic features of the task were fit for what the task intended to measure. It can be generalizable that the MAR-mediated task did not underrepresent the speaking constructs intended in Task 3. The responses were also able to discriminate test-takers according to proficiency. **6-2.** The theme of "global environment in crisis", one of the most frequently covered news topics in the world, can generate emotional responses that stimulate test-takers' language use and affective schemata of test-takers. The socialization feature of the MAR mode also helps to lower the anxiety that test-takers may feel from face-to-face interaction in oral interviews. A simulated addressee to whom a speaker speaks in each task – a friend, a polar bear and a rainforest ranger further motivates test-takers to speak. **6-3.** The MTMM analysis indicates that the test scores have positively moderate correlations with other speaking measures such as oral interviews (.605) and oral translations (.573).	Supported
7. Interpretations about the ability to be assessed are relevant to the decisions made at educational institutions.	**7.** The characteristics of the tasks highly reflect instructional tasks in the EFL speaking course.	**7.** The MARST was designed as an achievement test, which is part of the end-of-semester grades. The test tasks are similar to tasks already dealt with in the classroom. The contents of the textbook provide language functions, skills and knowledge reflecting the achievement standards of the national curriculum.	Supported
8. Score interpretations concerning the ability to be assessed are sufficient for the decision to be made.	**8.** The test provides sufficient testing contexts to collect evidence for inferring test-takers' speaking ability.	**8.** The multimodal features of AR in the MARST assesses a test-taker's language ability in various simulated contexts, generated by integrating virtual and real-world environments.	Supported
	9-1. The test	**9-1.** Good performance on the	

9. The test scores are useful for meeting the test purpose and making decisions about test-takers' English speaking competence.	scores are useful to determine the extent of the mastery of English speaking skills and knowledge as part of school academic achievement. 9-2. The MARST is sensitive to and considers local educational and societal values.	MARST implies that test-takers are ready to accomplish authentic tasks associated with making suggestions, describing visual information, expressing personal opinions and explaining the sequence of the working process of a device. 9-2. Developing the MARST was an ambitious attempt to overcome some constraints of the EFL language learning and teaching, namely – decontextualized test settings and reluctance of assessing speaking skills in EFL language assessment. However, it is doubtful whether it will earn larger recognition due to low credibility for self-assessment in our society.	Partially supported
10. Decisions on the basis of the score interpretations are equitable for stakeholders.	10-1. The decisions are not affected by raters' or interlocutors' personal attributes. 10-2. Test-takers have equal opportunities to learn or acquire the speaking abilities in the EFL speaking course.	10-1. The unique feature of the MAR mode provides a simulated interlocuter and indirect interaction between test-takers and raters. 10-2. Test-takers received instruction and information on what the test covered as well as the test preparation.	Supported
11. The consequences of using the MARST and decisions based on the MARST scores are beneficial to test-takers and stakeholders.	11-1. The impacts on test-takers, instruction, educational systems and society are promising. 11-2. The MARST is practical, as it is designed, developed and used within existing resources. 11-3. The MARST may hold the potential for positive washback by getting English language learners to regularly access to technology.	11-1. Test users' interview and survey responses on positive test experiences and potential test use for improving EFL teaching and learning practices were reported. 11-2. The unique features of the MAR mode – mobility, autonomy and immediacy– made the test practical. 11-3. The post-test interview reported test-takers' satisfaction with the option available in the mobile app that they could improve their answers by being allowed to review and revise their answers within the specific period of time. Raters could also access to test-takers' records at their convenience, which might offer insights about their learning growth. Easy accessibility to online materials would contribute to fostering digital literacy in the target language, which relates to one of the general objectives of the 2022 revised national curriculum.	Supported

Chapter 6. Discussion

This study presented an empirical attempt to investigate the effects of new technology integrated into existing testing procedures by developing an MAR-mediated speaking assessment for Korean EFL high school learners. And it went on to articulate a validation argument by examining whether the newly developed test could serve as a suitable test platform for assessing oral proficiency in an EFL setting.

The subsequent section begins by providing a summary of the results pertaining to each research question introduced in Chapter 1, followed by a presentation of the findings derived from the validity argument. Next, the section proceeds to advocate for the contextualization of MAR technology within L2 assessment. This is accomplished by addressing various validation issues associated with the integration of MAR technology in language assessment.

6.1 Summary of results for research questions

RQ1. To what extent are the test scores and the test's underlying factor structure comparable to those of other measures of the same speaking trait (i.e., oral translation and oral interview) and those of other traits (i.e., listening, reading, and writing)?

According to the hypothesis of the MTMM matrix, correlations

among measures of the same ability(monotrait correlations) would be higher than correlations among measures of different traits using different methods (heterotrait-heteromethod correlations). The MTMM analysis conducted in this study revealed that, on the whole, the scores obtained from the MARST exhibited slightly higher correlations with measures found in monotrait-heteromethod matrices. These measures include oral translations and oral interviews. In comparison, the correlations with measures in heterotrait-heteromethod matrices, such as multiple-choice listening, reading and writing, and writing translation, were slightly lower.

The following factor analysis among these speaking measures extracted two factors: one that corresponds to the speaking trait with higher factor loadings among all the measures, and the other factor that corresponds to different methods of the measures (i.e., MAR, multiple-choice, and face-to-face). Multiple-choice speaking test scores had the highest method factor loading, which means such test scores were more explained by the test method than the target construct to be assessed. On the other hand, among all the measures, the MAR test scores had the highest factor loading on the trait factor but the lowest factor loading on the method factor.

The four MARST task scores converged into one factor. Based on these findings, it can be concluded that the test method effect on the MARST was found to be negligible. The variance in MARST scores was predominantly accounted for by the trait factor, which represents the

target construct being assessed within the MARST.

The positive correlations observed between the MARST scores and other speaking measures, along with the unidimensional internal factor structure of the MARST, provide empirical evidence that supports the validation argument for meaningfulness, generalizability, and extrapolation inference. These findings indicate that the MARST test scores contribute to a common construct of speaking ability that is being assessed. Furthermore, the interpretations made about this assessed ability can be generalized to the broader domain of Target Language Use (TLU). In other words, the results suggest that the MARST effectively measures and captures the speaking ability in a way that can be meaningfully interpreted and applied to real-world language use situations.

Weak correlations among the four MAR test tasks indicate that each task played an independent role in measuring the target ability. The subsequent factor analysis extracted four factors that correspond to each task variables. Therefore, it can be inferred that the structure of the underlying dimension of the MARST has four separate sub-dimensions, each of which constitutes different aspects of the intended speaking proficiency in different tasks.

RQ2. To what extent do the assessment settings (e.g., test-taker, rater, task, and rating category) affect test scores?

The MFRM analysis sees each rating as a function of the interaction of test-taker ability, task difficulty, rating category (and scale step) difficulty, and rater severity. The current study sets up four facets – test-taker, task, rating category and rater.

To be more specific, the MFRM analysis could provide reliable estimation of expected scores based on observed scores of the MARST as the scoring of test-takers' proficiency was proven to be consistent within and across raters. The fit statistics of the rater facet were all within the productive range of 0.5 to 1.5, producing no overfit or misfit.

According to the fit statistics of the task facet, the high task reliability index (0.98) indicates that the degree to which tasks are replicable in terms of difficulty is high. The task fit statistics are all within the productive range of 0.5 to 1.5, producing no overfit or misfit. No misfitting or overfitting task indicates there is little chance of tasks being poorly made or perfectly good.

The average ability measures across the rating categories increased as the score level progressed, and each rating scale showed separate peaks for different score levels. The scale was positively linked to the progression of test-taker speaking ability and the rating scale functioned as expected.

On the test-taker facet, the separation and reliability indices for the difference in test-taker ability were high with 4.32 and 0.95 respectively. The MARST demonstrated high reliability and the ability to differentiate between test-takers of different ability levels. The test effectively

measured a wide range of abilities and showed statistically significant differences among test-takers. The presence of misfitting cases and unexpected responses, however, formed the need for subsequent qualitative analysis.

The bias (interaction) analysis indicates that raters' rating behaviors did not vary across different genders, test-takers region of residence, and rating criteria. More importantly, regarding the mode effect, there was no significant differences across gender, test-takers' general English proficiency level, region of residence, rating criteria and task type. The statistically significant interaction between rater and task, however, underscores the importance of careful rater training, the involvement in monitoring rater behaviors in the rating process after training, and maintaining examinee's motivation to participate in the test.

Overall, the analysis results evidence the validation arguments for meaningfulness and impartiality or, in other words, evaluation and generalizability inferences. The observed MARST scores are reliable estimates of expected scores, and the separate analytic rating scales contribute to the target construct. There is no task redundancy or need for revision and deletion. Test-takers perform significantly differently in the various aspects of speaking.

RQ3. What are the perceptions of test takers toward the use of the MARST and whether they will differ in individual characteristics such as gender and general English proficiency?

Regarding the mode effect, the interaction analysis in Section 4.3.2.1 revealed that test-takers' perception of the MAR mode did not differ across the six levels of test-takers' general English proficiency, nor did differ across gender, either. Thus, it can be concluded that the mode effect on the MARST test scores proves to be nonsignificant. The findings from this study evidence the validation claims regarding impartiality and the absence of biased elements in the assessment administration that could unfairly advantage or disadvantage certain groups of test-takers. Thus, interpretations of the test construct are consistent across different groups of test-takers.

The result of test-takers' questionnaires showed that compared with face-to-face speaking tests, the majority felt testing via the MAR device was more comfortable and interesting. Overall, they also thought that the test input presented by the MAR mode was highly authentic and provided sufficient guidance on what to do and how to construct their responses. They also gave high scores to the items asking whether the test tasks were appropriate enough to be presented in the MAR mode and relevant to what they had learned in classroom. The results of the teachers' questionnaires turned out to be more positive than those of the test-takers. All in all, test users believe the MARST to be useful as an alternative mode of L2 speaking assessment in EFL contexts.

RQ4. What are the linguistic features of MAR-mediated communication

and how do they inform the MAR-mediated test validation?

The qualitative analysis revealed that the use of MAR in language assessment enhanced test-takers' sensitivity to the presence of a simulated social being and promoted communication of both informative and interpersonal messages. It suggests that the MAR mode can allow for establishing interactions in speaking tasks where responses are limited to monologue or "talk alone". Thus, the construct being elicited in MAR-mediated monologic tasks may be operationalized accordingly. As a result, the type of monologic Task 3, which asks test-takers to explain sequential information to an non-appearing but significant interlocutor, is unlikely to underrepresent the speaking construct to be measured, but rather improves its quality or expands it.

The analyses of test-takers' speaking response corpora and the organizational structure of their spoken discourses revealed that some interpersonal and interactional elements required in the given task situation across all of the three proficiency groups – the use of sequential adverbs, modal expressions and imperative syntactic structures. Regarding this, the subsequent analysis found out that their responses reflected what they had learned in class to a large degree, which also leads to supporting the intended purpose of the MARST test as achievement assessment. In particular, using such language features, the high proficiency group produced consistently higher results. It may be suggested here that not only their pure language proficiency but also their

highly achieving attitudes towards academics should be taken into consideration as potential contributing factors on this output.

Tasks 1 and 3, which involved simulated interlocutors, were designed to assess the viability and soundness of using semi-direct (or simulated) speaking tasks delivered via the MAR mode as an alternative to direct speaking tests delivered via face-to-face interviews.

6.2 Validation issues

6.2.1 Integrating MAR technology in L2 assessment

Technology has made significant contributions to language testing. Computer-based language testing (CBT) has been widely used as an alternative to traditional paper-and-pencil tests. In recent years, MAR has emerged as a potential alternative to computer-based testing. Based on what have been found in this dissertation, Table 47 summarizes the potential benefits of MAR-mediated language testing are specified in comparison with computer -based testing in terms of devices, language test purpose, test constructs, test tasks, and testing conditions. This would be considered a useful way to contextualize MAR technology integrated in L2 assessment.

MAR has a unique advantage in its ability to provide an immersive and interactive testing environment. The interactive nature of MAR can be particularly useful for testing language skills specific to oral communication revealing interactive and dynamic interactions, as MAR

can simulate real-life communication scenarios more effectively than CBT.

Table 47. Integrating MAR technology into developing language assessment

Test category \ Mode	Computer	MAR
Device	desktop, laptop	smartphones, tablets
Purpose	• standardized large-scale, high-stakes • assessment for general language proficiency	• small scale classroom-based in local context, low-stakes • assessment for ESP
Construct	overall language proficiency including four skills	listening and speaking abilities in specific target language domains
Task	• requiring low level of interactivity (reading/ writing), multiple-choice, fill-in-the-blank, etc. • more general topic, input/output	• simulating real-life communication, performance-based • topics of specific fields, input/output tied to specific use and sociocultural contexts
Testing condition	controlled	less controlled
Advantage	• standardized • efficient and cost-effective particularly with automatic scoring system	• immersive and engaging authentic, contextualized • immediate feedback
Disadvantage	• limited interactivity and authenticity • limited feedback	• difficulty in standardization • costly

It has been a proven fact over a couple of decades that the powerful use of CBT is manifest for large-scale standardized assessment of general language proficiency, centering on evaluating the entire four skills at once. Meanwhile, MAR seems to be fit for local contexts such as classroom assessment which intends to promote or reinforce language learning and practice.

As one of the most significant advantages of MAR-mediated language testing is its flexible and accessible device. MAR can be easily used on smartphones and tablets with one hand, while CBT is limited to desktop or laptop computers. Owing to this flexibility, can MAR-mediated language testing reach a wider range of learners and provide a more convenient and accessible testing environment.

When it comes to test constructs, MAR-mediated language testing can be effective in assessing constructs related to communication and social interaction, such as sociolinguistic competence, discourse competence, and pragmatic competence. This is because MAR allows for more realistic simulations of social situations. On the other hand, computer-based testing may be more suitable for assessing written communication or linguistic knowledge, including grammar and vocabulary.

The types of tasks used in language testing can also vary between MAR-mediated and CBT. MAR allows for more interactive and immersive task types, such as role-plays and simulations. Computer-based testing, however, is more suited for tasks that require a significant amount of text or visual input, such as reading comprehension and writing tasks.

Unlike CBT, which offers efficient and cost-effective testing conditions for large-scale assessments, MAR testing allows test-takers to have discretion or autonomy in deciding when and where to take the test. However, this autonomy can potentially pose threats to test

reliability and validity. Suggestions on how to minimize these potential threats will be addressed in 6.2.5.

Overall, both MAR-mediated and computer-based language testing have their advantages and disadvantages, and the choice of test mode depends on the specific context and purpose of the test. Ultimately, language educators and testers must carefully consider the advantages and limitations of each test mode to select the most appropriate testing approach for their needs. A summary of contextualizing MAR technology in developing language assessment in reference to CBT is outlined in Table 43.

Revisiting the modeling of MAR-mediated speaking test performance in Figure 5, the accessibility and interactivity of MAR test mode are now known to influence test-taker, task (and construct), and performance. Therefore, test mode must be included as a major component in modeling MAR-mediated speaking test performance, as argued in 2.3.4 where MAR test mode is called construct-relevant.

6.2.2 Mode effect on test construct

The second issue concerns whether the use of MAR may underrepresent the oral communication construct by not requiring examinees to invoke some of the pragmatic competences. In the literature about the test mode effect on language production of test-takers, O' Loughlin (2001) which examined the linguistic features in

responses to a tape-based speaking test in comparison with the live version, characterized the response data to be of monologic ability with more lexical density, while the responses of the live version were associated with monologic ability. Likewise, the more recent qualitative study conducted by Nakatsuhara and colleagues (2017) on the video-conferenced speaking test revealed that VC communication tended to involve more explicit language for negotiating meaning. This indicated that VC communication might not always allow for subtle ways of establishing understanding and taking turns. Consequently, when assessing the speaking construct in VC tests, it was crucial to operationalize it in a manner that encompasses explicit negotiation of meaning and effective turn management, capturing these aspects as part of the construct.

In the current dissertation, meanwhile, the qualitative analysis of 150 sampled speaking responses to Task 3 indicated that the MAR-based monologic task requiring test-takers to explain the sequence of events to a simulated interlocutor in a specific field of occupation elicited not only monologic (i.e., describing a procedure) but also interactive (i.e., negotiating or engaging in conversation with a simulated interlocutor) features of communicative utterance. It seems that they were attributed to the immersive and interactive nature of the MAR environment, which possibly facilitated the cognitive process of transferring information during the speaking task according to the animation principle, one of the cognitive principles of MAR technology, as explicated in 2.2.3. There

needs to be more follow-up research on investigating what occurs in the cognitive aspects of test-takers in performing MAR-based tasks.

Both monologic and interactive abilities are important aspects of speaking proficiency in a second language, and they require different language skills and strategies. Assessing both monologic and interactive abilities can provide a more comprehensive and accurate picture of a learner's overall speaking proficiency. At this point, it can be proposed that this immersive and interactive MAR capabilities may hold a promising potential to bridging gaps between the two conflicting perspectives on L2 ability : Communicative Competence (Canale & Swain, 1986 ; Bachman & Palmer, 1996) and Interactive Competence (Kramsch, 1986; Hall, 1993, 1995; Young, 2013). The two positions are different in that communicative competence captures innate traits that reside within individual language learners in a given social and interactive testing context. Thus, factors that might affect the target abilities have been labelled as 'construct-irrelevant', while the interactive competence focuses on social interactions co-constructed and shared among participants.

The two camps deserve criticism ; for the former, on one hand, is said to take a limited perspective on language learning, disregarding the complexity of communication, and on the other hand, the latter does not fully serve assessment purposes in reality particularly in the EFL context, where L2 learners are highly restricted to interactive and social contexts to use the target language.

In Table 48, competence measured by MAR-mediated speaking assessment is assumed to reside in a kind of a neutral zone where the two extremes can be flexibly adjusted, with the two positions keeping in balance and canceling out negative effects. Although it cannot perfectly replicate real-life communicative situations, when aligned with advancements of MAR technology, the featured MAR strengths of authenticity and interactivity that contribute to generating immersive environment for language use will probably uphold the dynamic features

Table 48. MAR-mediated competence in connection to communicative competence and interactive competence for L2 speaking assessment

L2 ability / Test category	Communicative competence (Canale & Swain,1980, 1981; Bachman & Palmer, 1996)	MAR-mediated competence	Interactive competence (Hall, 1993, 1995; Kramsch, 1996; Young, 2013)
Focus	individual's communication in a social context	individual's communication in online interaction with simulated participants[39]	jointly participation by all individual participants
Construct	a trait or bundle of traits of an individual independent of interactive practice	individual's competence depending on simulated interactive practice	interactional competence co-constructed by all participants in interactive practice
Practice	general	local	local
Mode	F2F, P&B, CBT[40]	MAR	F2F[41]
Condition	controlled	less controlled	less controlled

[39] for example, 3D animated avatars resembling physical and affective features of human beings
[40] F2F, P&B, and CBT refer to face-to-face (interview), pencil and paper, and computer-based test modes respectively.
[41] F2F here refers to not only face-to-face pair interview but also group discussion tests as well.

of mutual communication in interactive test practices, as what the camp supporting interactive communication has argued.

6.2.3 Mode effect on test task

Regarding the mode effect on task, the n-gram analysis of test-takers' responses to Task 3, mentioned in 4.4.2.3, can offer insights into a strand of research that explores the types of tasks that MARST can best serve: tasks that involve greater complexity. And it echoes what York (2019) argued: the virtual environment might be particularly beneficial for learners when dealing with complex tasks.

N-grams, which indicate collocations, are multiword expressions composed of a defined number (represented as 'n') of words within a reference corpus (Saito and Liu, 2022). Researchers have examined meaningful correlations between n-gram frequency and L2 speaking proficiency assessment (Kyle and Crossley, 2015; Eguchi and Kyle, 2020); for example, according to Kyle and Crossley (2015), trigram frequency explained the largest amount of variance related to holistic proficiency scorings on TOEFL iBT Speaking tasks (r = .59).

In addition, collocation effects are considered to be clear when tasks are well-structured with known content ; particularly in rating linguistic accuracy and fluency of the picture description task rather than the interview task. It may be due to the fact that raters already know the story, paying more attention to the linguistic characteristics than

semantic content/details of their speech (Saito and Liu, 2022).

From pedagogical point of view, mastery of collocations helps learners improve communication fluency and produce a more authentic and fluent language by gaining access to ready-made phrases and expressions (i.e., some prepositional phrases and passive structures required to perform Task 3) that convey meaning more accurately and effectively. Due to their familiarity with the test content, which presented the procedure of how the device works in an animated MAR mode, learners were able to concentrate on using learned collocations accurately and fluently. The findings support the assumption that the more proficient group would display higher n-gram frequencies compared to the less proficient groups. This suggests that the MAR mode has the potential to alleviate the cognitive burden that test-takers may experience when addressing Task 3.

Learning collocations offers another advantage in terms of contextual appropriateness, which relates to the specific contexts or situations where language is used. By acquiring collocations, learners gain an understanding of how to appropriately use language in various contexts — a crucial aspect that is often seriously lacking in EFL classroom settings.

For this, MAR technology will likely be of great use for learners to practice using collocations in tasks that present different simulated situations. Applied to the current study, for instance, test-takers had learned some prepositional phrases and adverbs tightly associated with

addressing movement, locations and directions, as well as the appropriate use of passive syntactic structure that focuses on what is discussed over the agent that discusses it.

It can be via the MAR app that learners practice collocations, used in the given situation where rainforest rangers keep alert to stop illegal logging deep in the rainforest areas, such as 'sends a signal' , 'go to the site immediately' , 'is picked up by' or 'on the ground nearby'. Through interactive tasks and challenges presented in the MAR environment, learners can apply collocations in contextually appropriate ways. This situational practice strengthens their ability to use collocations effectively and accurately.

MAR creates opportunities for learners to engage in authentic language use contexts. They can participate in virtual conversations or simulations where they must use collocations in a realistic and contextually appropriate manner. This helps learners develop their proficiency in using collocations in authentic communication settings.

To summarize, Task 3 may offer an example of the positive technical features of MAR test mode, characterized as authenticity and interactivity, stimulating test-takers' individual cognitive aspects. This would, in turn, lead to improving test-takers' performance including automaticity, associated with communication fluency. Consequently, the quality of the construct or language ability to be measured in the test is enhanced.

6.2.4 Mode effect on test-takers

Technology can create a learning and assessment platform where learner's affective and cognitive processes intersect(Lajoie, 2014). Thus, some cognitive principles and concepts of AR are adopted to prepare test-takers for dealing with task complexity and assessment difficulties. In 2.2, some principles and concepts related to AR technology were presented, which can be utilized to aid test-takers in handling the cognitive intricacies associated with complex tasks and challenging assessments.

The extent to which individual characteristics contribute to the differences in test-takers' emotional reactions to the immersive virtual environment needs to be more researched. The earlier studies of Wang et al. (2009), Liou (2011) and Chen and Kent (2020) reported that learners experienced less anxiety in virtual environment, building confidence, boosting motivation, and empowering them via avatar anonymity. On the other hand, Ockey et al.(2017) found out contrary results because the learner's role was controlled by the machine and they could not get support when they ran into difficulties.

In this dissertation, the post-test questionnaire surveys to test-takers and teachers, who are potential users, and interviews indicated that a majority of participants seemed to be positively aware of the features of AR technology integrated in the mobile app, which served as a test mode. Some notable responses about the new mode included that

it was authentic, interesting, reduced the burden and pressure to speak, and was helpful for practicing speaking. It was also found to be convenient, as it allowed multiple users – both test-takers and raters – to access it at once. Moreover, users appreciated the flexibility to choose their own test time and place. This design could potentially lessen test-takers' feelings of pressure and burden.

Attention should be drawn to the questionnaire survey results concerning the effect of the MAR on test-takers' motivation. This scored an average of 2.77 on a scale of 1 to 4, which was lower than that of teachers, who scored an average of 3.47 on the same scale. Follow-up interviews with some test-takers revealed concerns about how their speaking performances would impact their official school grades, even if it was only a small portion.

In relation to this, previous research (Putwain & Daly, 2013) found out that students demonstrating either low test anxiety with high academic resilience or medium test anxiety with high academic resilience achieved the highest academic performance while, in contrast, students who experienced high test anxiety coupled with low academic resilience displayed the lowest performance. This could imply that anxiety isn't always detrimental to a test-taker's performance, indicating that further investigation is needed on how test anxiety influences individuals in the MAR assessment environment.

Another consideration derived from the post interviews may be that if a test incorporates a new mode, it should ideally be designed for

formative or diagnostic purposes. All in all, more empirical evidence needs to be gathered on the physical testing environments and conditions that determine whether pressure and burden will have opposing psychological effects, whether positive or negative.

6.2.5 Control of variabilities of test conditions

Keeping variabilities of test conditions under control when the MARST is administered can be challenging, requiring careful attention to the technology used, task design, test-taker instructions, and rater training. The qualitative analysis in this dissertation highlights the importance of several careful considerations when applying MAR technology to language assessment. One such consideration is the standardization of technology, which involves ensuring that all test-takers use the same type of mobile device and have the same version of the app.

The analysis also underscores the importance of thorough preparation in terms of providing clear and detailed instructions to test-takers prior to the assessment. This includes communicating the purpose of the assessment, outlining the task requirements, and specifying the expectations for performance. To ensure consistency and address potential variabilities in the task or technology, several alternatives can be considered. One effective approach is to conduct a pilot test of the assessment, which allows for the identification of any issues or

challenges that may arise. The pilot test provides an opportunity to refine the assessment and make any necessary adjustments before administering it to the actual test-takers.

As reported in 4.3.3.2, where the sources of interaction between Rater 3 and Task 1 and 3 were discussed, the potential rating order effect should not be overlooked. To minimize this effect, it is crucial not only to provide the training session prior to rating but also to actively monitor the rating process. For instance, it is recommended to provide appropriate oral or written feedback tailored to the rating in specific contexts. This approach ensures consistent scoring of speaking responses across all test-takers and tasks.

In conclusion, it is crucial to continue efforts in controlling variabilities of test conditions when administering the MARST. By doing so, the MARST can become a reliable and valid assessment of language proficiency. Therefore, further research including new and follow-up studies should explore and address any remaining challenges in need of improvement in order to enhance the reliability and validity of the MARST as an assessment tool.

Chapter 7. Conclusion

7.1 Technological implications of MARST

In the attempt to make technological innovations in language testing, new technologies are assumed to serve a range of functions in many contexts (Chapelle and Lee, 2021). Research on the MARST adds to the growing number of studies shedding light on the importance of situating technology issues within a validity framework.

In the same vein, the current study may introduce the potential use of the MARST, which will serve to narrow the gap between assessment and learning in the classroom context by changing perceptions about assessment. The MARST, which is not considered a high-stakes test, can offer test-takers a second opportunity to demonstrate their academic or instructional knowledge including linguistic and topical one. The MARST pays attention to test-takers' affective schemata, and they do not need any testing equipment other than their own mobile and AR markers. Moreover, they are free to determine the test location and time.

The study's findings suggest potential applications of MAR in English language testing for specific purposes. These applications include providing diverse, concrete, and sensitive contexts closely related to the TLU domain. These factors are primary concerns in test development and administration, and they inevitably impact the test validation process. The MAR mode enables the development of a speaking test which

includes tasks requiring test-takers to incorporate input materials derived from sources that are both cognitively and affectively engaging. As a result, test-takers construct their responses based on these sources, rather than relying solely on their personal experiences and opinions.

To enhance the quality of MARST, several factors need to be considered. Primarily, the quality of MARST relies on the technological sophistication of the device. It is crucial to ensure that the hardware used for delivering the assessment is current and capable of supporting high-quality AR experiences. The introduction of more advanced face recognition technology could also bolster test security. If implemented, it could monitor potential cheating in real-time on behalf of proctors. This would not only strengthen test security but also increase practicality, thereby contributing to the reliability of the test.

Integrating voice recognition technology could also prove beneficial, providing immediate feedback on pronunciation and intonation. The combination of games and MAR could make the MARST more effective in facilitating learning-oriented assessments. Indeed, the inclusion of user-friendly interfaces from interactive language games, improved visual effects, and 3D avatars in MARST can support multiple test-takers' participation in the same real-life simulated scenarios.

Last but not least, use of the MARST raises ethical considerations. These include data privacy and security, fairness, and accessibility. It is crucial to ensure that these factors must be diligently addressed during

the development and implementation of the assessment.

7.2 Pedagogical implications of MARST

Regarding pedagogical implications, the MARST can contribute significantly to achieving the standards of high school English subjects outlined in the 2022 revised national curriculum, from various perspectives. Firstly, the MARST can offer a more interactive and engaging language use experience. It achieves this by providing opportunities to interact with virtual characters, test inputs, or other learners in a more natural and immersive way.

It can provide a personalized language learning experience by catering to individual needs and preferences of learners. It enables practice of language skills at any time and place, and provides immediate feedback on areas requiring improvement. Consequently, the MARST can foster a sense of responsibility for their language learning in learners, aligning with the curriculum's emphasis on self-directed learning.

As an innovative and cutting-edge technology, the MAR can inspire learners to approach language learning with greater enthusiasm and motivation. By leveraging the benefits of MAR, educators can design innovative, engaging tasks for language learning and assessment. These tasks support the development of language proficiency and communicative competence, in line with the curriculum's emphasis on fostering a positive attitude towards second language learning.

7.3 Limitations and suggestions for future research

The validation process is portrayed as an ongoing activity without a clear endpoint, as noted by Chapelle, Jamieson, and Hegelheimer (2003). Indeed, the validation discussed here was not merely meant to confirm and defend an initial design and its operationalization. Instead, it was open to change and refinement, reflecting the iterative nature of test development and validation. Thus, the validity argument for the MARST is not considered complete but rather an ongoing process. Other researchers or language testers may employ MARST, and evaluate necessary steps for test interpretation and use.

This study offers several practical suggestions for future research. First, although rater and test-taker opinions were collected via questionnaires and interviews, future studies on rating behaviors and task performances could employ screen-capturing software like Camtasia, or more advanced technology such as eye-tracking (Yang, 2021). For instance, efforts could be made to gather more evidence that enhances rater consistency, thereby further supporting evaluation and generalization inferences. One potential source of evidence could be to examine probable causes of variations in rater severity and within-rater differences across tasks, from a cognitive perspective, using think-aloud protocols or eye-tracking (Choi, 2021), regardless of how minor the interaction. Moreover, it would be beneficial to explore methods for

achieving and maintaining homogeneity in rating different MARST tasks.

Second, the extrapolation inference could be further reinforced with non-test evidence from English and other disciplinary courses that align with the test-takers' MARST scores. This approach would ensure that the speaking construct assessed by the MARST reflects the speaking skills required in real-life settings. The inference from the MARST scores is only partially supported by its similarities with other speaking measures based on shared evaluation criteria, such as grammatical accuracy, fluency, and topic development. This is because the correlations between scores on these measures were not high enough.

Third, there might be high demand for more preparation or warm-up activities before the main tasks, which encompass various topics (such as recycling, environmental posters, and invented devices for saving rainforests) and rhetorical functions (e.g., exchanging information, making suggestions, describing, expressing personal opinions, and explaining the working steps of a device). Extra measures should have been introduced to reduce the cognitive load on test-takers during the tasks within the given time limit.

Moreover, qualitative analyses examining the strategies involved in the speaking process for successful task completion, and investigations into the discourse features of test-takers' responses, could support the explanation inference in the validity argument. This implies that the expected scores are attributed to the target construct of speaking ability.

As a continuation of the current research, the security function in the

MARST could be upgraded with advanced technologies, and the system could be customized. This would allow teachers to design and develop speaking assessments that align with their instructional interests.

Recently, there has been an increasing number of attempts to research virtual environments for assessing language for specific purposes. As a result, L2 researchers, language testers, and professionals in technology-mediated language assessment should collaborate to design and develop more authentic learning and testing contexts for language learners.

Bibliography

Anthony, L. (2022). AntConc (4.2.0) [Computer Software]. Tokyo, Japan: Waseda University.

Bachman, L. F. (2004). *Statistical analysis for language assessment*. Cambridge University Press.

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

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

Bal, B. (2010). Analysis of Four-word Lexical Bundles in Published Research Articles Written by Turkish Scholars." Thesis, Georgia State University, 2010. https://doi.org/10.57709/1665591

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 Research Papers. IELTS Partners: British Council, Cambridge Assessment English and IDP: IELTS Australia. Retrieved from https://www.ielts.org/teaching-and-research/research-reports.

Blagg, D. (2009). Augmented reality technology brings learning to life. Retrieved from http://www.gse.harvard.edu/blog/uk/2009/09/augmented-reality-technology-brings-learning-to-life.html.

Bond, W. J., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental*

measurement in the human sciences(2nd ed.). Mahwah. NJ: Lawrence Erlbaum.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*(1), 89-110.

Bonk, W. J., & Van Moere, A. (2004). *L2 group oral testing: The influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores*. Paper presented at the Language Testing Research Colloquium, Temecula, California.

Byun, J.-H. (2020). Exploring the potential of integrating mobile augmented reality into speaking performance assessment. *STEM Journal, 21*(1), 65-93.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105

Campbell, A., & Main, S. J. (2014). Performance, assessment and communication in one app: Mobile tablet assessment is here to stay. *eCulture, 7*(1), 1-3.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Canale, M., & Swain, M. (1983). A theoretical framework for communicative competence. In A. S. Palmer, P. J. M. Groot & G. A. Trosper (eds.), The construct validation of tests of communicative

competence, Proceedings of a colloquium at TESOL '79 (pp. 31-36). Washington, D.C. : Teachers of English to Speakers of Other Languages.

Chalhoub-Deville, M. (1999). *Issues in computer-adaptive testing of reading proficiency*. Cambridge, UK: Cambridge University Press.

Chao, K. H., Chang, K. E., Lan, C. H., Kinshuk, & Sung, Y. T. (2016). Integration of mobile AR technology in performance assessment. *Educational Technology & Society, 19*(4), 239-251.

Chapelle, C. A., Jamieson, J., & Hegelheimer, V. (2003). Validation of a web-based ESL test. *Language Testing*, *20*(4), 409-43

Chapelle, C., A., & Douglas, D. (2006). *Assessing Language through Computer Technology*. UK: Cambridge University Press.

Chapelle, C. A., & Lee, H.(2021). Understanding Argument-Based Validity in Language Testing. In C. A. Chapelle and E. Voss (Eds.), *Validity Argument in Language Testing: Case Studies of Validation Research* (pp.19-44). Amsterdam: Cambridge University Press.

Chen, C. H. (2010). The implementation and evaluation of a mobile self- and peer-assessment system. *Computers & Education, 55*(1), 29-236.

Chen, J. C., & Kent, S. (2020). Task engagement, learner motivation and avatar identities of struggling English language learners in the 3D virtual world, *System, 88.* https://doi.org/10.1016/j.system.2019.102168.

Chen, W., Gu, X., & Wong, L. (2017). To click or not to click:

Effectiveness of rating classroom behaviors on academic achievement with tablets. *British Journal of Educational Technology, 50*(1), 440-455.

Chiang, T. H. C., Yang, S. J. H., & Hwang, G.-J. (2014). Students' online interactive patterns in augmented reality-based inquiry activities. *Computers & Education, 78*, 97-108.

Choi, J. B. (2019). *Phono Sapien*. Sam and Parkers.

Choi, Y. (2021). Generalization Inference for a Computer-Mediated Graphic-Prompt Writing Test for ESL placement. In C. A. Chapelle and E. Voss (Eds.), *Validity Argument in Language Testing: Case Studies of Validation Research* (pp.120-153). Amsterdam: Cambridge University Press.

Cuendet, S., Bonnard, Q., Do-Lenh, S., & Dillenbourg, P. (2013). Designing augmented reality for the classroom. *Computers & Education, 68*, 557-569.

Cumming, A. (1997). The testing of writing in a second language. In C. Clapham & D. Corson(Eds.), *Encyclopedia of Language and Education* (Language testing and assessment, Vol. 7). Dordrecht; Kluwer Academic.

Dede, C., Salzman, M., Loftin, R. B., & Ash, K. (2000). The design of immersive virtual learning environments: Fostering deep understandings of complex scientific knowledge. In J. M. Jacobson, & R. B. Kozma (Eds.), *Innovations in science and mathematics education: Advanced designs for technologies of learning* (pp. 361-

413). Mahwah, NJ: Lawrence Erlbaum Associates.

Dunkel, P. A. (1999). Considerations in developing or using second/foreign language proficiency computer-adaptive tests. *Language Learning & Technology, 2*(2), 77-93.

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessment: A many-facet Rasch analysis. *Language Assessment Quarterly, 2*, 197-221.

Eguchi, M., & Kyle, K. (2020). Continuing to explore the multi-dimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal 104(2),* 381-400.

Ellis, R. (2003). *Task-based Language Learning and Teaching*. Oxford University Press, New York.

Espada-Gustilo, L. (2011). Linguistic features that impact essay scores: A corpus linguistic analysis of ESL writing in three proficiency levels. *The Southeast Asian Journal of English Language Studies*, *17*(1), 55-64.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

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

Fulcher, G. (2015). *Re-examining Language Testing: a philosophical and social inquiry*. Oxon and New York, Routledge.

Gan, Z. (2010). Interaction in group oral assessment: A case study of

higher-and lower-scoring students. *Language Testing, 27*(4), 585-602.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2009). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics, 30*(3), 315-334.

Hall, J. K. (1993). The role of oral practices in the accomplishment of our everyday lives: The sociocultural dimension of interaction with implications for the learning of another language. *Applied Linguistics, 14*, 145-166.

Hall, J. K. (1995). (Re)creating our worlds with words: A sociohistorical perspective of face-to-face interaction. *Applied Linguistics, 16*, 206-232.

Han, S. Jeong, E., Park, S., Lee, B., Lee, H., & Jang, E. (2018). *High school English*, YBM Holdings.

He, L. & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*(3), 370-401.

Henning, G. H. (1987). *A guide to language testing: Development, evaluation, and research*. Rowley, MA: Newbury House.

Henning, G. H. (1992). Dimensionality and construct validity of language tests. *Language Testing*, *9*(1), 1-11.

Hewlett Packard. (2018). HP Reveal [Mobile application]. Retrieved from https: //studio. hpreveal.com /landing

Hidsdon, J. (1991). The group oral exam: Advantages and limitations. In

J.C. Alderson & B. North (Eds.), *Language testing in the 1990's* (pp.189-197). London: Modern English Publications and the British Council.

Hong, S. (2013). *An N-gram Analysis of Korean English Learners' Writing. Korean Journal of English Language and Linguistics, 13*(2), 313-336.

Hsu, T. C. (2017). Learning English with augmented reality: Do learning styles matter? *Computers & Education, 106*, 137-149.

Hwang, G. J., & Chang, H. F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education, 56*(4), 1023-1031.

IBM Corp. (2019). IBM SPSS Statistics for Windows (Version 26.0) [Computer software]. IBM Corp.

Jeon, M., & Choe, Y. (2019). A Coh-Metrix analysis of Korean EFL learners' summary writings in the English argumentative and expository texts. *Korean Journal of English Language and Linguistics, 19* (3), 539-559. DOI: 10.15738/kjell.19.3.201909.539.

Johnson, D. W., & Johnson, R. T. (1994). *Learning Together and Alone: Cooperative, Competitive, and Individualistic Learning.* Allyn and Bacon, Boston, MA.

Johnson, M. (2001). *The art of non-conversation: A reexamination of the validity of the oral proficiency interview*. New Haven, CT: Yale University Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 38*, 319-342.

Kane, M. T. (2006). Validation. In R. Brennen (Ed.), Educational measurement (4th ed., pp.17-64). Westport, CT: Greenwood Publishing.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73.

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.

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

Kim, S. S., Park, C. O. & Seol, H. S. (2009). Validation of The Leadership Style Rating Scale for Young Children Using Rasch Measurement Model. *Journal of Korea Open Association for Early Childhood Education, 14*(3), 517-556.

Kim, S.-Y. (2010). The effects of virtual reality based CMC on English language learning. *English Language Teaching*, *22*(4), 53-74.

Kim, Y.-M. (2016). *Development of edutainment contents for children's English education based on virtual reality - centered on ABC House*. [Master's thesis, Namseoul University].

Kopriva, R. (2008). *Improving testing for English language learners*. New

York: Routledge.

Kramsch, C. (1986). From language proficiency to interactional competence. *The Modern Language Journal, 70,* 366-372.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly 49(4)*, 757–786.

Lai, C. L., & Hwang, G. J. (2015). An interactive peer-assessment criteria development approach to improving students' art design performance using handheld devices. *Computers & Education, 85*, 149-159.

Lajoie, S. P. (2014). Multimedia learning of cognitive processes. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning, second edition* (pp. 623–646). New York: Cambridge University Press.

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*, 159–174, DOI: 10.2307/2529310

Lazaraton, A. (2002). *A qualitative approach to the validation of oral tests*. Cambridge: Cambridge University Press.

Leaper, D. A. (2014). *Consistency in Performance in the Group Oral Discussion Test: An Interactional Competence Perspective* (Published doctoral dissertation). Macquarie University: Canada.

Lee, G.-Y. (2017). *The effect of learning activities using virtual reality on vocabulary learning and learning attitudes of elementary English*

learners. [Master's thesis, Cyber Hankuk University of Foreign Studies].

Lee, S.-H. (2010). *The effect of using 3D VR avatar chat on university students' English learning : focus on vocabulary recognition and English composition*. [Doctoral dissertation, Chung-Ang University].

Lee, S.-H. (2018). *The effect of English speaking lessons using virtual reality on learners' speaking ability and affective domain*. [Master's thesis, Cyber Hankuk University of Foreign Studies].

Linacre, J. M. (1989). Many-facet Rasch measurement. Chicago, IL: MESA Press.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*(3), 266-283.

Linacre, J. M. (2006). A user's guide to FACETS Rasch model computer program. Available on line at: www.winsteps.com

Liu, T.-Y. (2009). A Context-aware ubiquitous learning environment for language listening and speaking. *Journal of Computer Assisted Learning, 25*, 515-527.

Liou, H.-C. (2011). The roles of Second Life in a college computer-assisted language learning (CALL) course in Taiwan, ROC. *Computer Assisted Language Learning*, *25*(4), 365–382.

Low, R., & Sweller, J. (2005). The modality principle in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning, first edition* (pp.147-158). New York:

Cambridge University Press.

Lowe, R. K., & Schnotz, W. (2014). *Animation principles in multimedia learning.* In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning, second edition* (pp. 513-546). New York: Cambridge University Press.

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

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

McNamara, T. F. (2002). Discourse and Assessment. *Annual Review of Applied Linguistics, 22,* 221-242.

Mayer, R. E. (2006). Ten research-based principles of multimedia learning. In H. F. O'Neil & R. S. Perez (Eds.), *Web-based learning: Theory, research, and practice*. Mahwah, NJ: Lawrence Erlbaum.

Mayer, R. E. (2014a). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning, second edition* (pp.43-71). New York: Cambridge University Press.

Mayer, R. E. (2014b). Principles based on social cues in multimedia learning: personalization, voice, image, and embodiment principles. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning, second edition* (pp.345-368). New York: Cambridge University Press.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement: Issues and Practice, 17*, 6-12.

Mislevy, R., Almond, R., & Lucas, J. (2003). A brief introduction to evidence-centered design. (RR-03-16). Educational Testing Service, Princeton, USA.

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language Testing, 28*(4), 483-508.

Nakatsuhara, F., Inoue, C., Berry, V., & Galaczi, E. (2017). Exploring performance across two delivery modes for the IELTS Speaking Test: face-to-face and video-conferencing delivery (Phase 2). IELTS Partnership Research Papers, 3. IELTS Partners: British Council, Cambridge English Language Assessment and IDP: IELTS Australia. Retrieved from https://www.ielts.org/teaching -and-research/research-reports

Nering, M. L., & Ostini, R. (2010). *Handbook of polytomous item response theory models*. Routledge.

Nikou, S. A., & Economides, A. A. (2018). Mobile-based assessment: A literature review of publications in major referred journals from 2009 to 2018. *Computers & Education, 125*, 101-119.

Norris, J. (2001). Concerns with computerized adaptive oral proficiency assessment. *Language Learning & Technology, 5*(2), 99-105.

Norris, J. (2002). Interpretations, Intended Uses ad Designs in Task-based Language Assessment. *Language Testing, 19*(4), 337-346. https://doi.org/10.1191/0265532202lt234ed

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*.

London, UK: Cambridge University Press.

Ockey, G. J., Gu, L. & Keehner, M. (2017) Web-Based Virtual Environments for Facilitating Assessment of L2 Oral Communication Ability, *Language Assessment Quarterly, 14*(4), 346-359, DOI: 10.1080/15434303.2017.1400036

Palmer, A. S., Groot, P. J., & Trosper, G. A. (Eds.). (1981). *The construct validation of tests of communicative competence.* Washington, DC: TESOL.

Pang, H.-S. (2008). *Developing e-learning contents based on VR.* [Master's thesis, Graduate school of Techno design, Kookmin University].

Putwain, D. W., & Daly, A. L. (2013). Do clusters of test anxiety and academic buoyancy differentially predict academic performance?, *Learning and Individual Differences, 27*, 157-162.

Redondo, B., Cózar-Gutiérrez, R., González-Calero, J.A., & Ruiz, R. S. (2020). Integration of Augmented Reality in the Teaching of English as a Foreign Language in Early Childhood Education. *Early Childhood education Journal 48*, 147–155, DOI: 10.1007/s10643-0 19-00999-5

Roever, C., & Al-Gahtani, S. (2015). The development of ESL proficiency and pragmatic performance, *ELT Journal*, *69*,(4), 395–404, https://doi.org/10.1093/elt/ccv032

Saito, K., & Liu, Y. (2022). Roles of collocation in L2 oral proficiency revisited: Different tasks, L1 vs. L2 raters, and cross-sectional vs.

longitudinal analyses. *Second Language Research, 38*(3), 531-554

Santos, M. P., Hernádez-Leo, D., Pèrez-San Agustín, M., & Blat, J. (2012). Space-Aware Design Factors for Located Learning Activities Supported with Smart Phones. In *Proceedings of the 6th International Conference of Multimedia Ubiquitous Engineering (MUE 2012)*, Leganés, Spain 792-798.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Shealy, R. T., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*(2), 159-194.

Shohamy, E. (2004). The validity of direct versus semi-direct oral tests. *Lang Tests, 11*(2), 99-123.

Shohamy, E., Or, I. G., & May, S. (Eds.). (2017). *Language testing and assessment* (pp. 441-454). Springer.

Sung, Y. T., Chang, K. E., & Liu, T. C. (2016). The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis. *Computers & Education, 94*, 252-275.

Suppiah Shanmugam, S. K., Wong, V., & Rajoo, M. (2020). Examining the quality of English test items using psychometric and linguistic characteristics among grade six pupils. *Malaysian Journal of Learning and Instruction, 17*(2), 63-101.

Swanson, J. A. (2020). Assessing the effectiveness of the use of mobile technology in a collegiate course: A case study in M-learning. *Technology, Knowledge and Learning*, *25*, 389-408.

The Ministry of Education. (2015). 2015 revision of the national English curriculum for secondary schools. Retrieved from http://ncic.re.kr/mobile.kri.org4.inventoryList.do.

Tilstra, K., & Smakman, D. (2018). The Spoken Academic English of Dutch University Lecturers, *English Studies, 99* (5), 566-579, DOI: 10.1080/0013838X.2018.1483620

Traxler, J. (2010). Distance education and mobile learning: Catching up, taking stock. *Distance Education, 31*(2), 129-138.

Triantafillou, E., Georgiadou, E., & Economides, A. A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. *Computers & Education, 50*(4), 1319-1330.

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

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

Wang, A. I. (2015). The wear out effect of a game-based student response system. *Computers & Education, 82*, 217-227.

Wang, H., Choi, I., Schmidgall, J., & Bachman, L. F. (2012). Review of Pearson Test of English Academic: Building an assessment use argument. *Language Testing 29*(4). 603-619.

Wang, C. X., Song, H., Xia, F., & Yan, Q. (2009). Integrating second life into an EFL program: students' perspectives. *Journal of Educational Technology Development and Exchange (JETDE)*, *2*(1), 1~16. DOI: 10.18785/jetde.0201.01

Wang, X. (2014). The relationship between lexical diversity and EFL writing proficiency. *University of Sydney Papers in TESOL, 9*, 65-88.

Wang, Y.-H. (2017). Exploring the effectiveness of integrating augmented reality-based materials to support writing activities. *Computers & Education*, *113*, 162-176.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. *Assessing Writing, 6*, 145-178.

Weng, C., Otanga, S., Christianto, S. M., & Chu, R. J. C (2020). Enhancing students' biology learning by using augmented reality as a learning supplement. *Journal of Educational Computing Research, 58*(4), 747-770. https://doi.org/10.1177/0735633119884213

West, M., & Vosloo, S. E. (2013). *UNESCO policy guidelines for mobile learning*. Paris: UNESCO.

Whiteside, A. J. (2002). Beyond interactivity: Immersive web-based learning experiences. *The e-Learning Developers' Journal, 283*, 1-10.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A

qualitative investigation. *TESOL Quarterly, 47*(4), 762-789. DOI:10.1002/tesq.73.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling, 3*, 3-24.

Wright, B. D., & Linacre, J. M. (1994). *Reasonable mean-square fit values*, available at: www.rasch.org.

Xi, X., & Sawaki, Y. (2017). Methods of validation. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language Testing and assessment, Third edition* (pp.193-209). Cham: Springer.

Yang, H. (2021). Support for the Evaluation Inference : Investigating conditions for rating responses on a Test of Academic Oral Language. In C. A. Chapelle and E. Voss (Eds.), *Validity Argument in Language Testing: Case Studies of Validation Research* (pp.96-119). Amsterdam: Cambridge University Press.

York, J. (2019). *Language learning in complex virtual worlds: Effects of modality and task complexity on oral performance between virtual world and face-to-face tasks*. (Published) [Doctoral dissertation, Meiji University] ResearchGate.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199-225. DOI:10.1177/0265532212456965.

Young, R. F.(2011). Interactional competence in language learning, teaching, and testing. In E. Hinkel, (Ed.). *Handbook of research in second language teaching and learning* (Vol.2), (pp.426-443). New

York, NY: Routledge.

Young, R. F. (2013). Learning to talk the talk and walk the walk: International competence in academic spoken English. *Ibérica 25*, 15-38.

Young, R. F., & Miller, E. R. (2004). Learning as changing participation: Negotiating discourse roles in the ESL writing conference. *Modern Language Journal, 88*(4), 519-535.

Zhou, Y. (2015). Computer-delivered or face-to-face: effects of delivery mode on the testing of second language testing. *Language Testing in Asia 5*(2). DOI: 10.1186/s40468-014-0012-y

국문 초록

한국 고등학교 영어 학습자를 위한 모바일 증강현실 기반 말하기 평가 개발과 타당화 연구

변정희
영어영문학과(응용언어학, 제2언어 평가 전공)
서울대학교

이 논문은 모바일 기반 증강현실(MAR) 기술을 이용한 제2외국어 말하기 평가의 실행 가능성을 탐구할 목적으로 한국 고등학생들을 대상으로 MAR 기반 영어 말하기 시험을 개발하고 검증하는 데 기울인 노력에 대해 자세히 설명하였다.

MAR 기반 영어 말하기 시험(이하 MARST로 표기)을 개발하기 위해 "Eco English Test"라는 모바일 애플리케이션을 제작하여 고1 남학생 110명과 여학생 90명을 포함한 약 200명의 한국 고등학생들에게 말하기 평가를 시행했다. 이 말하기 시험은 교실 수업에서 평가 이전에 학습한 언어 기능과 스킬의 숙달도를 평가하는 성취평가의 목적을 띠고 한 학기 수행평가로 치러졌다. 글로벌 환경과 관련된 단원 학습 후 해당 주제에 대한 네 가지 간접(semi-direct) 말하기 과제를 제시했다.

언어평가의 타당도 프레임워크인 Assessment Use Argument (AUA; Bachman and Palmer, 2010)와 Interpretation/Use Argument (I/UA; Kane, 1992, 2006, 2013)를 사용하여 MARST의 타당도 검증을 수행하고, 특히 이러한 혁신적인 평가 모드가 제2 언어 평가의 검증 과정에서 어떻게 맥락화(contextualize)될 수 있는지에 중점을 두었다.

다음 네 가지 연구 질문을 다루었다; (1) MAR에 의한 말하기 평가의 기저 구조(underlying structure)가 동일한 말하기 능력 및 다른 능력(읽기, 쓰기, 듣기)을 측정하는 평가들과 어느 정도 비교가능한가? (2) 평가의 여러 국면들(예: 채점자, 과제, 측정기준)은 얼마나 MARST 점수에 영향을 미치는가? (3) 시험 사용자들의 MARST 사용에 대한 인식은 어떠하고, 인식의 정도가 성별과 일반 영어 능력 등의 개인 특성에 따라 다른가? (4) MAR을 매개로 산출된 발화의 언어적 특징은 무엇이며 MAR 기반 평가의 타당도 검증에 어떤 도움을 주는가?

데이터 분석을 위해 말하기 능력 및 다른 스킬의 여러 측정치 시험 점수, 설문조사와 면접 응답을 수집했고, Multi-Trait Multi-Method (MTMM) 및 Many-Facet Rasch Model (MFRM)과 같은 심리측정학적 접근법 뿐만 아니라 응시자의 말하기 응답에 대한 코퍼스 및 담화 분석을 포함하는 혼합 방법(mixed method)을 사용하여 분석했다. MTMM 분석의 결과는 MARST 점수와 다른 말하기 측정치 간의 긍정적인 상관관계를 보여주는 것뿐만 아니라, MARST의 일원적인 내부 요인 구조를 밝혀냈다. MFRM 분석은 다음과 같은 유효성 주장 – (1) MARST의 관찰된 점수는 예상 점수의 신뢰할 수 있는 추정치이다 (2) 별도의 분석 등급 척도가 목표로 하는 구인 (targeted construct)측정에 기여한다 (3) 과제의 중복이 없으며 수정이나 삭제할 필요가 없다 (4) 시험 응시자들을 다양한 수준으로 변별할 수 있다 (5) 테스트 구조의 해석은 테스트 응시자 그룹 간에 일관성이 있다 – 을 지원하는 실증적 증거를 제공했다.

또한, MAR 모드 효과는 다양한 측면에서 통계적으로 유의하지 않았다. 다만, 편향분석(bias analysis)에서 발견된 두 명의 등급자의 점수 부여 행동과 과제 1(대화 완성) 및 과제 3(사건의 순서 설명) 사이에 유의미한 상호 작용은 문헌에서 제안된 기준과 후속 면접에 기반하여, 시험 응시자의 수행 능력

측정에 실질적인 영향을 미치지는 않은 것으로 나타났다. 한편, 과제 3에 대한 150개 응답 샘플을 코퍼스 및 담화 분석한 결과, 상호작용적(interpersonal /interactional) 의사소통의 특징들이 드러났다. 이는 과제 수행 과정에서 응시자들이 가상의 대화자에 대한 존재 인식 및 발화자로서 자신의 역할에 대한 인식, 그리고 주어진 상황에 대한 감수성의 증가에 기인한 것으로 판단되었다. MAR 모드의 고유한 특징인 몰입 효과가 가상의 대화 상대와의 상호 작용을 촉진함으로써, 직접적인 말하기 테스트가 제한적일 수 밖에 없는 EFL 상황에서 영어 말하기 평가의 대안으로서 잠재력을 보여주었다. 결론적으로, 위에서 밝혀진 분석 결과들은 MARST가 과제 3에서 의도한 말하기 평가 구인(construct)을 제한하지 않았음을 시사했다.

이후, 타당도 주장에서 세 가지 핵심 이슈 – (1) 제 2언어 말하기 평가에서 MAR 기술의 맥락화 (contextualize), (2) 평가 구인 (construct), 평가 과제 (task), 그리고 시험 응시자 (test-taker)에 끼치는 MAR 모드의 영향, 그리고 (3) 시험 조건 (test conditions)의 가변성 (variability) 통제- 에 대해 이론적 분석과 통찰을 제시했다. MARST에 대한 기술적 및 교육적 시사점 중의 하나로 말하기 평가 과정에서 일어나는 응시자의 과제 수행 행동과 전략 연구의 필요성을 언급했다. 끝으로 언어 평가자들과 기술 전문가들이 협력하여 언어 학습자들을 위한 더욱 실제적인 언어 학습 및 평가 상황을 설계하고 개발하는 것을 후속 연구로 제안한다.

Appendices

Appendix 1. Test design

[Situation] 두 친구가 세계환경의 날을 기념하기 위한 활동으로 재활용품 만들기를 제안하고 주어진 사례를 이용하여 만들 수 있는 재활용품을 소개하는 대화를 완성한다(문항1). 가상의 북극곰이 들려주는 온난화로 인한 생존의 위협을 알게 되고, 지구의 미래는 인간의 선택에 달려있음을 보여주는 환경포스터를 감상한다(문항2). 이후 지구 온난화를 막는 숲의 중요성을 인식하고 불법 벌목을 감시할 목적으로 폐휴대폰을 재활용하여 만든 장치인 "RFCx"의 작동법을 익혀 열대우림 레인저들에게 이를 소개한다. (문항3)

(종이화면#1)

Let's celebrate the World Environment Day on June 5!

Please open the AR app on your mobile phone and bring the camera to the marker to find out what we can do to celebrate it.

an old jean

styrofoam wastes

plastic bottle wastes

(AR화면 #1) 에코베어 등장, 나레이션

Hi, I am Ecobear. Do you know what day is celebrated on June 5? It's World environment day. I would like to discuss recycling as a way to celebrate the day this year with your friend. What can we make from recycling old and waste materials? Let's find it out by bringing your camera to each item.

(AR 화면#2) 종이화면의 old jeans와 연결

a tote bag

a wall-hanging organizer

(AR 화면#3) 종이화면 styrofoam waste와 연결

a picture frame

a flowerpot

(AR 화면#4) 종이화면 plastic bottle waste와 연결

a pencil case

a chair

(AR화면#5) 에코베어 등장, 나레이션 : In the following conversation, you are going to suggest to your friend making a recycled item on the world environment day, using the examples you have seen.

(종이화면 #2)

TASK 1. Open the app on your smartphone and bring the camera on the marker to complete the conversation on the AR screen.

[Preparation : 3mins / Recording : 1mins]

단원	Inventions for rainforests	
문항번호	1	
문항분류	영역 듣기,**말하기**,읽기,쓰기	행동 **지식,**이해,**적용,**분석,종합,창의
관련성취기준	친숙한 일반적 주제(환경)에 대한 의견 표현하기	
평가요소	제안하기와 가능성 표현하기 (making a suggestion and expressing one's ability),	
예상소요시간	4mins	

(AR화면#6) 수험자의 발화를 담은 오디오 음성이 구현

친구 A의 발화는 기계음의 대화 음성 자료로 처리, 수험자에게 발화를 시작하라는 알림음 포함.

예 시 답	A: Do you have any idea what we can do to celebrate the World Environment Day? B: Hmm. Why don't you(we) try making a recycled item? / How about recycling old ones to make new ones./ How about making new items recycling old ones?

<table>
<tr><td>안</td><td>A: Sounds like fun. I wonder what to make.
B : Lots of things. I can make a tote bag from my old jean / I can make a chair out of plastic bottles/ I can make a wall hanger recycling old jeans.
A: Maybe I'll try that.</td></tr>
<tr><td>배점 및 채점기준</td><td>*배점: 20점
<table>
<tr><th></th><th>Language use
(8)</th><th>Delivery
(6)</th><th>Content / Topic development (6)</th></tr>
<tr><td>탁월
★★
★★</td><td>제안과 능력 표현에 사용되는 문법과 문맥 에 적절한 어휘를 정확 하게 사용하여 소통함 (8)</td><td>흐름에 막힘이 거의 없이 잘 준비해서 말한 느낌. 발음과 억양이 영어권 화자의 발화에 가까워 자연스러움 (6)</td><td>대화의 흐름에 맞게 첫 빈칸에는 기념일에 할 일이, 두번째 칸에는 만들 것이 알맞게 들어가고 전체 대화가 자연스럽게 연결됨. (6)</td></tr>
<tr><td>만족
★★
★☆</td><td>제안과 능력 표현에 사용되는 문법과 어휘를 구사하였으나 소통에 방해 되지 않는 오류 있음 (6)</td><td>대체적으로 막힘 없이 흘러감. 의사소통에 방해 되지 않으나 발음 과 억양에 한국어 발음 의 흔적으로 부자연 스러운 부분이 있음 (4)</td><td>두개의 빈칸 문장이 각각 이어지는 말과 연결이 되나 대화 전체의 흐름에는 어색함이 있음. (4)</td></tr>
<tr><td>보완
★★
☆☆</td><td>제안과 능력 표현에 사용되는 문법과 어휘 사용에서 문장수준의 소통에 방해되는 오류 있음 (4)</td><td>머뭇거림, 끊김이 단어 수준에서 더러 있어서 듣는 이의 인내심이 요구 (2)</td><td>두개의 빈칸 문장 중 어느 하나가 이어지는 말과 연결되지 않고 비약이 있어 대화 전체의 일관성을 낮춤. (2)</td></tr>
<tr><td>미흡
★☆
☆☆</td><td>전적으로 단어, 구 수준에서 의사소통이 이루어짐. (2)</td><td>끊김이 많고 단어 수준의 발화조차 일관되게 힘들어서 평가 근거가 희박 (0)</td><td>답안의 내용이 적거나 두개 빈칸 문장 모두가 이어지는 말과 연결이 되지 않고 비약이 있어 대화전체의 흐름을 찾기 힘듦. (0)</td></tr>
</table></td></tr>
<tr><td>난이도</td><td>상, 중, 하</td></tr>
<tr><td>출제근거</td><td>고1 영어 교과서,YBM홀딩스 4단원 78-79쪽, 한상호 외 저, 2018
https://www.sheknows.com/living/articles/1062108/crafts-for-your-old-jeans/
http://www.newdaily.co.kr/site/data/html/2010/05/12/2010051200046.html
https://m.cafe.daum.net/2846ajy/Lb3L/227
http://www.greenmax-recycling.com/bulky-waste-styrofoam-would-be-a-big-problem-for-city-eps-recycling-industry/
https://slowalk.com/1474
https://cen.acs.org/policy/trade/Existing-treaty-help-manage-global/96/i26
https://youtu.be/Rt0TjOQfWtk
https://images.app.goo.gl/5fMr52KUwNVvWucP8
손으로-그려진-북극곰-국경 png from pngtree.com
전화 png from pngtree.com</td></tr>
<tr><td>유의사항</td><td>A는 기계음으로 B의 대답과 일정한 간격을 두어 재생되도록 한다.</td></tr>
</table>

단원	Inventions for rainforests
문항번호	2

문항분류	영역 듣기,**말하기**,읽기,쓰기	행동 지식,**이해,**적용,**분석,종합,창의**
관련성취기준	대상 묘사 및 설명하기(환경포스터 설명하기), 친숙한 일반적 주제(환경)에 대한 주관적 감정 및 생각 표현하기	
평가요소	대상에 대한 설명하기 (explaining a familiar topic), 주관적인 생각이나 감정 표현하기(expressing one's feeling)	
예상시간	8mins	

(종이화면 #3) With the app on, bring your smartphone camera to our planet Earth below and you will see an environment poster.

(AR화면# 7) 애니메이션 효과, 캐릭터의 설명이 포함된 환경포스터

*AR 애니메이션 효과 삽입 : 위, 아래에 슬로건 'Choose one for our future. Start from now.' 등장, 빨간 쪽은 지구에 불길이 일어나서 Co_2가 방출되는 모습, 녹색 쪽은 열대우림에서 나는 소리와 나무에서 O_2가 방출되는 모습

(종이화면 #4) Please, bring your phone camera to the marker to learn more about the poster and what Escobar asks you to do.

(AR화면#8) 에코베어 등장, 나레이션

Hi, I'm Ecobear living in the north pole, the coolest place in the world. But, I feel sad about my home disappearing due to the global warming. Look at the environmental poster. There are two choices you can make for our planet. Describe **EACH** choice including its color, sound, and appearance. Then, tell me **which choice** will help living creatures on earth like me to survive and **what message** the poster tries to tell humans including you.

(종이화면 #5)

TASK2. Please reply to Ecobear by creating your own AR on the marker.

[Preparation : 5mins / Recording : 3mins]

_____

_____

(AR화면#9) 수험자의 발화를 담은 오디오 음성이 구현

<table>
<tr><td>예시
답안</td><td>Left side has a blue color with green nature. Also it has a smile face and trees with zero carbon dioxide. It looks like a eco-friendly face with fresh air and no environmental pollution. On the other hand, the right side looks so hot because of its red color. And its blood-shot eyes show that it is in a terrible situation. It is also surrounded by strong flames and the ground is melting so it seems that a disaster is about to break out. Thus, we, human beings, have to choose the left side.
This poster shows the future is in our hands, informing us of the current situation we are facing, which is that we are now at crossroad to choose between the left and right sides of the earth. It is up to us to decide whether to live in clean nature or suffer from terrible disasters. At this crossroads, we have to choose the left side and start from now. We must try to realize it right now.</td></tr>
<tr><td>배
점

및

채
점
기
준</td><td>*배점: 50점
<table>
<tr><th></th><th>Language use (10)</th><th>Delivery (20)</th><th>Content/Topic development(20)</th></tr>
<tr><td>탁월
★★
★★</td><td>다양한 문법, 전문 용어를 포함한 어휘의 선택과 사용이 충분하고 약간의 사소한 문법적 오류를 제외하고 정확하게 구사 (10)</td><td>흐름에 막힘이 거의 없고 충분한 연습으로 잘 준비해서 말한 느낌의 발음과 억양이 특징. 듣기에 편함 (20)</td><td>문제에서 요구하는 모든 사항(객관적묘사, 주관적 의견)에 대해 대상을 주의 깊게 관찰했고, 관련된 내용을 충분히 제시함. (20)</td></tr>
<tr><td>만족
★★
★☆</td><td>기본적인 문법, 전문 용어를 포함한 어휘 선택과 사용이 어느 정도 드러나지만, 소통에 방해 되지 않는 오류가 더러 있음 (8)</td><td>대부분의 발화에서 흐름이 느껴짐. 의사소통에 방해되지 않으나 발음상의 부정확성과 한국어의 영향이 더러 있음. (15)</td><td>문제에서 요구하는 객관적 묘사와 주관적 의견 표현 모두에서 필수적인 내용을 제시. (15)</td></tr>
<tr><td>보완
★★
☆☆</td><td>단순한 구조와 좁은 범위의 어휘 사용으로 최소한의 문장단위 발화가능. 문법과 어휘에서 소통에 방해가 되는 오류가 눈에 띔 (5)</td><td>흐름이 자주 끊기고 한국어의 영향으로 단조로운 억양과 부정확한 발음이 강하여 듣는 이가 듣기에 불편하여 반복 요청을 하게 함. (10)</td><td>문제에서 요구하는 객관적 묘사와 주관적 의견 표현 가운데 어느 한 가지에서 필수적인 내용이 결여되어 준비가 미흡했다는 인상을 줌. (10)</td></tr>
<tr><td>미흡
★☆
☆☆</td><td>문장단위가 아닌 분리된 몇 개의 단어 나열로 의사소통이 심각하게 한정됨 (3)</td><td>단어수준의 발화로 인해 흐름이 느껴지지 않아서 평가 근거가 희박 (5)</td><td>두가지 사항 모두에서 필수적인 내용이 결여되어 전혀 준비되지 않은 인상을 줌(5)</td></tr>
</table></td></tr>
<tr><td>난이도</td><td>상, 중, 하</td></tr>
<tr><td>출
제
근
거</td><td>https://blog.naver.com/PostView.nhn?blogId=mckko&logNo=221496425411&parentCategoryNo=7&categoryNo=22&viewDate=&isShowPopularPosts=false&from=postView
https://www.facebook.com/106224341135349/posts/saturdaythoughts-our-future-depends-upon-usso-start-from-now-save-nature-save-yo/125053629252420/</td></tr>
</table>

단원	Inventions for rainforests
문항번호	3

문항분류	영역 듣기,말하기,읽기,쓰기	행동 지식,**이해**,적용,**분석**,**종합**,창의
관련 성취기준		
평가목표	제시된 업사이클링 혹은 리사이클링 발명품에 대한 자료를 이해하여 제작법, 작동원리, 혹은 사용법을 단계별로 말하기	
평가요소	논리적 순서, 인과관계 말하기(그림보고 연결하여 말하기)	
예상소요 시간	8mins	

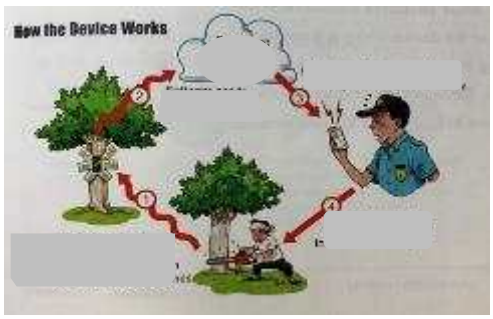
(종이화면 #6)
*지시문: With the app on, bring your smartphone camera to the marker below and find out how useful Ecobear thinks a recycled device is to save rainforests.

(AR 화면 #10) 에코베어 등장, 나레이션이 사진과 함께 제시

Topher White, a software engineer, invented a small device to save the rainforests, recycling old smartphones. It is called RainForest Connection or RFCx. The device has a sensitive microphone to catch illegal logging. Solar pannels attached to the device allows it to get electricity from the sun. And, wireless internet service makes it easy for rainforest rangers to use the device.

(종이화면 #7) With the app on, bring your smartphone camera on the marker to figure out how the device works. Next, you are supposed to explain it to rainforest rangers who work with the device every day to protect the rainforest.

(AR화면 #11) : 에코베어 등장, 나레이션이 애니메이션 자료와 함께 제시될 예정.
(일러스트 예시: 교과서 그림에서 발췌)



(종이화면 #8)
TASK 3. Please reply to Ecobear by creating your own AR on the marker.
[Preparation : 5mins / Recording : 3mins]
[Memo]

_____
_____
_____

(AR 화면 #12) 학생 응답 데이터 : 단계적으로 기계의 작동법을 설명한 오디오 자료파일

<table>
<tr><td>예시
답안</td><td>First, when a guy tries to cut down a tree with a chainsaw, the sound of the chainsaw is picked up by microphones attached to the solar-powered RFCx, located high up on a tree. Second, the detected noise is transmitted into the cloud. Third, you, a ranger nearby, will receive the signal from the cloud. Finally, you, a ranger nearby, go to the spot immediately to catch illegal logging. Don't worry. The RFCx will also work under the shade of trees as it is solar-powered. Trust our company.</td></tr>
<tr><td>배
점

및

채
점
기
준

배
점

및

채
점
기
준</td><td>*배점: 50점
<table>
<tr><th></th><th>Language use (10)</th><th>Delivery (20)</th><th>Content/
Topic development (20)</th></tr>
<tr><td>탁월
★★
★★</td><td>다양한 어법과 연결 어 및 단어반복을 활용해 단계 간 연결이 잘 됨. 전문 용어를 포함한 다양 한 어휘 선택과 사용이 문맥에서 잘 드러남. (10)</td><td>흐름에 막힘이 거의 없이 잘 준비해서 말한 느낌. 발음과 억양이 자연스러워 듣기에 매우 좋음 (20)</td><td>문제에서 요구하는 4단계의 절차적 정보 모두에 대해 순차적 으로 필요한 정보를 제공하여 순서 정보 가 명확히 드러남. 설명을 듣는 이를 의식하며 말하는 인상을 줌. (20)</td></tr>
<tr><td>만족
★★
★☆</td><td>기본적인 문법, 연결어 및 전문 용어 를 포함한 필수 어휘를 구사하여 순서 정보가 드러남. 소통 에 방해되지 않는 오류가 더러 있음 (8)</td><td>대체적으로 흐름이 느껴 짐. 의사소통에 방해 되지 않으나 한국어의 영향으로 단조로운 억양과 부정확한 발음이 더러 있음. (15)</td><td>문제에서 요구하는 4단계에 대해 빠짐 없이 순차적으로 제시하였으나 단계간 연결성은 약함. 설명을 듣는 이를 의식하며 말하는 인상을 줌. (15)</td></tr>
<tr><td>보완
★★
☆☆</td><td>문법 및 전문용어를 포함한 어휘사용 이 단순하거나, 흔적이 부족함. 문장간 연결 어를 거의 사용 하지 않아 순서 정보 가 명확히 드러나지 않음. 소통에 방해 되는 오류 있음 (5)</td><td>흐름이 자주 끊기고 한국어의 영향으로 단조로운 억양과 부정확한 발음이 강 하여 듣는 이가 듣기 에 불편하여 반복 요청을 하게 함. (10)</td><td>문제에서 요구하는 4단계 가운데 한 단계 이상에서 정보가 누락되거나 관련 성이 부족한 내용이 있음. (10)</td></tr>
<tr><td>미흡
★☆
☆☆</td><td>몇 개의 단어 나열로 순서정보가 드러나지 않고 제한된 문법 어휘로 의사 소통이 심각하게 제한됨 (3)</td><td>단어수준의 발화로 인해 흐름이 느껴 지지 않아서 평가 근거가 희박 (5)</td><td>두 단계 이상에 대한 설명이 누락되어 작동방식과 절차를 이해하기 힘듦. (5)</td></tr>
</table></td></tr>
<tr><td>난이도</td><td>상, 중, 하</td></tr>
<tr><td>출제
근거</td><td>고1 영어 교과서,YBM홀딩즈 4단원, 한상호 외 저
*동영상1. https://youtu.be/JtCk1Obg02s (Rainforest Connection 제공 동영상)
*동영상2. https://youtu.be/qEHH9VSWYTI (Rainforest Connection 제공 동영상)</td></tr>
</table>

Appendix 2. Questionnaires

*1: highly disagree, 5: highly agree

No.	Statement	0	1	2	3	4
1	I find the MAR test experience interesting and innovative.					
2	The MAR test experience encourages me to speak English.					
3	The MAR test experience is less stressful and burdensome in comparison with face-to-face speaking test.					
4	Materials presented in the MAR test look realistic and authentic.					
5	Materials presented in the MAR test offers sufficient clues for me to construct responses.					
6	I spoke, being aware of the presence of simulated interlocutors such as a friend, Ecobear, and rangers.					
7	I had no difficulty in understanding task instructions and performing tasks through the MAR device.					
8	I understood the English speaking ability that the MAR-based tasks test.					
9	I understood the English speaking ability that the MAR-based tasks test.					
10	I am willing to take the MAR-based speaking test again.					

Appendix 3. Mean scores of four dimensions(item easiness)

	Task 1	Task 2-1	Task 2-2	Task 3
Accuracy	2.98	2.78	2.70	2.90
Fluency	3.40	3.08	3.09	2.94
Content	3.15	2.74	2.99	2.74
Sum	9.52	8.61	9.05	8.58

Appendix 4. Item-total correlation (item discrimination)

	Task 1	Task 2-1	Task 2-2	Task 3
Accuracy	0.53	0.73	0.74	0.70
Fluency	0.61	0.81	0.79	0.74
Content	0.58	0.68	0.74	0.64
Sum	0.75	0.87	0.88	0.85

Appendix 5. Measure of agreement (Cohen's Kappa, p = .000)

	ACC	FLU	CONT	SUM
TASK1	0.70	0.59	0.41	0.42
TASK2-1	0.47	0.64	0.39	0.53
TASK2-2	0.62	0.63	0.57	0.54
TASK3	0.41	0.64	0.61	0.50

Appendix 6. Predicted reliability for different test lengths (Spearman-Brown Prophecy formula)

N	1	2	3	4	5	6	7	8	9	10
ACC	0.562	0.720	0.794	0.837	0.865	0.885	0.900	0.911	0.920	0.928
FLU	0.641	0.781	0.842	0.877	0.899	0.914	0.926	0.934	0.941	0.947
CONT	0.429	0.631	0.749	0.826	0.880	0.921	0.952	0.977	0.997	1.014
SUM	0.560	0.718	0.793	0.836	0.864	0.884	0.899	0.911	0.920	0.927

Appendix 7. Unexpected responses (32 residuals) in MFRM analysis

Cat	Score	Exp.	Resd	StRes	N	Re	N	G	N	M	Num	Exa	N	Rate	N	Task	N	Cate	Sequence
3	3	4.0	-1.0	-5.5	1	JJ	1	F	4	A	4	004	1	Jung	3	opin	3	CONT	81
3	3	1.2	1.8	4.8	2	CW	1	F	1	D	158	158	4	Byu	4	sequ	2	FLU	3719
2	2	3.8	-1.8	-4.4	1	JJ	2	M	4	A	129	129	2	Nam	4	sequ	3	CONT	3084
3	3	4.0	-1.0	-4.4	1	JJ	2	M	4	A	135	135	4	Byu	3	opin	1	ACC	3235
3	3	3.9	-.9	-4.2	1	JJ	1	F	3	B	24	024	4	Byu	4	sequ	2	FLU	575
2	2	3.8	-1.8	-4.1	1	JJ	1	F	4	A	32	032	4	Byu	1	conv	3	CONT	759
2	2	3.8	-1.8	-4.0	1	JJ	2	M	4	A	129	129	4	Byu	4	sequ	3	CONT	3096
3	3	1.2	1.8	4.0	2	CW	1	F	1	D	158	158	3	Mat	4	sequ	2	FLU	3707
2	2	3.8	-1.8	-3.9	1	JJ	1	F	4	A	110	110	2	Nam	2	desc	3	CONT	2622
2	2	3.8	-1.8	-3.8	1	JJ	1	F	4	A	32	032	1	Jung	1	conv	1	ACC	745
2	2	3.7	-1.7	-3.8	1	JJ	1	F	4	A	130	130	2	Nam	4	sequ	2	FLU	3107
3	3	1.3	1.7	3.8	1	JJ	2	M	4	A	70	070	1	Jung	4	sequ	1	ACC	1666
1	1	3.4	-2.4	-3.8	2	CW	1	F	1	D	200	200	4	Byu	1	conv	2	FLU	4646
3	3	3.9	-.9	-3.7	1	JJ	1	F	2	C	122	122	4	Byu	2	desc	1	ACC	2920
2	2	3.7	-1.7	-3.7	1	JJ	1	F	4	A	32	032	4	Byu	1	conv	1	ACC	757
2	2	3.7	-1.7	-3.5	1	JJ	1	F	4	A	130	130	4	Byu	4	sequ	2	FLU	3119
2	2	3.7	-1.7	-3.5	1	JJ	2	M	3	B	117	117	2	Nam	2	desc	3	CONT	2790
1	1	3.3	-2.3	-3.4	1	JJ	2	M	3	B	19	019	4	Byu	3	opin	1	ACC	451
3	3	1.3	1.7	3.4	2	CW	1	F	4	A	149	149	4	Byu	1	conv	2	FLU	3542
3	3	1.3	1.7	3.4	2	CW	1	F	4	A	199	199	4	Byu	1	conv	2	FLU	4622
2	2	3.7	-1.7	-3.3	1	JJ	1	F	3	B	53	053	1	Jung	2	desc	3	CONT	1254
2	2	3.7	-1.7	-3.3	1	JJ	1	F	4	A	80	080	1	Jung	2	desc	3	CONT	1902
1	1	3.3	-2.3	-3.3	1	JJ	2	M	3	B	69	069	1	Jung	1	conv	1	ACC	1633
3	3	1.4	1.6	3.2	1	JJ	2	M	4	A	70	070	1	Jung	3	opin	1	ACC	1663
2	2	3.7	-1.7	-3.2	2	CW	1	F	1	D	157	157	3	Mat	3	opin	2	FLU	3680
4	4	1.9	2.1	3.1	1	JJ	2	M	1	D	14	014	1	Jung	4	sequ	3	CONT	324
2	2	3.7	-1.7	-3.1	1	JJ	2	M	2	C	58	058	1	Jung	1	conv	3	CONT	1371
3	3	3.9	-.9	-3.1	1	JJ	2	M	4	A	18	018	1	Jung	2	desc	3	CONT	414
1	1	3.2	-2.2	-3.0	1	JJ	1	F	2	C	3	003	4	Byu	4	sequ	1	ACC	70
3	3	3.9	-.9	-3.0	1	JJ	1	F	4	A	46	046	1	Jung	4	sequ	3	CONT	1092
2	2	3.6	-1.6	-3.0	1	JJ	2	M	1	D	115	115	2	Nam	3	opin	2	FLU	2744
2	2	3.6	-1.6	-3.0	1	JJ	2	M	3	B	140	140	2	Nam	4	sequ	2	FLU	3347
Cat	Score	Exp.	Resd	StRes	N	Re	N	G	N	M	Num	Exa	N	Rate	N	Task	N	Cate	Sequence

Appendix 8. Misfit cases of test-takers' ability measures in MFRM analysis

Case No.	Test-taker No.	Measure	Infit MS	Outfit MS
1	5	-4.32	1.54	1.00
2	12	.50	1.83	1.85
3	17	-.05	1.58	1.59
4	26	-.12	1.61	1.57
5	27	.85	1.56	1.74
6	32	1.55	2.12	2.47
7	33	-.27	1.78	1.67
8	69	.25	1.83	1.85
9	70	-3.55	1.66	1.78
10	111	.61	1.64	1.72
11	112	-.41	1.55	1.57
12	127	.04	1.56	1.65
13	129	2.27	1.94	1.64
14	130	1.82	1.54	1.46
15	145	-1.01	1.71	1.67
16	149	-4.75	1.74	.98
17	150	.52	1.67	1.75
18	157	.92	1.58	1.77
19	158	-4.75	2.13	1.80
20	173	-.64	1.80	1.82
21	199	-4.75	1.74	.98

Appendix 9. Sample transcripts of spoken responses to Task 3

Speaker A

This device has a sensitive microphone, so it detects illegal logging. When logging is detected, information about it is sent to the cloud. So the cloud sends some sign to the rainforest rangers and you can locate them. Then you can go to the illegal logging spot immediately.

Sample B

First, sound of chainsaws is picked up by microphones and solar powered cellphones. Second, software sends signal to cloud. Third, real time alert is received by a ranger on the ground nearby. Fourth, that enables the rangers to go to the site immediately.

Speaker C

Installing forest connection on the tree, then when rainforest connection does illegal logging, the microphone sensitively catches the sound of the electric saw, find the location and receives the location from the forest guard through the cloud to help the forest guard move quickly.

Speaker D

First, if the illegal logging is happening, and I can detect the noise of the logging with the sensitive microphone. Then it will send the signal with this wireless internet service to rainforest ranger.

Speaker E

Put this device on the tree and wait. If someone cuts a tree with an electric saw, the device will recognize the sound and send a message to protect forest future.

Speaker F

If noise occurs from the machines of those who illegally damaged the mountain, the machine, first, detects noise through sensor and passed into the forest security through the cloud and the forest ranger and follow the signal to protect the rainforest.

Speaker G

The deforestation method is a method in the which a microphone attached to a tree detects the noise of a chainsaw that is illegally logging and store the noise in the cloud and then send a signal to the rainforest ranger so that the security guard locates the seat and dispatches it.

Speaker H

The RFCX is made by recycling old smartphones. First, it will change a sound while charging solar battery. The device has sensitive microphone to catch illegal logging sound. If it get the sound, sends a signal to cloud to call rainforest ranger.

Speaker I

If illegal logger cut the tree using the electric saw, RFCX detect a noise using their sensitive microphone. Software sends a signal to cloud. Real time alert is received by a ranger on the ground nearby. That enables the rangers to go to the site immediately.

Speaker J

The rainforest connection detected the noise of chainsaw and send a message to rainforest rangers through the cloud. So, rainforest rangers catch the bad guy.

Speaker K

Device detects illegal logging. It sends information to the cloud. Cloud then sends information to the rainforest ranger, and the rainforest ranger is dispatched to the place where illegal logging take place.

Speaker L

When sound of chainsaws is picked up by microphones in solar powered cellphone, software send a signal to cloud and real time alert is received by a ranger on the ground nearby. Last, that enables the ranger to go to the site immediately.

Speaker M

The first is illegal logging with chainsaws. Second, it detects the noise with a microphone attached to the tree. Third, it sends a signal to the cloud. A forest guard receives the signal.

Speaker N

The device is called RFCX. Let me explain how it works. First, the device detects logging of electric saws with a sensitive microphone. Then they send a signal to the cloud, then the rainforest ranger receives the logging detection signal from the cloud. After that, rainforest rangers can go to block the logging.

Speaker O

First, as a man is doing illegal logging with his chainsaw, a solar powered microphone attached the tree detects a noise from the chainsaw. Second, It sends a signal to clouds and then forest rangers receive a signal. They can know the location of illegal logging, and they go to the illegal logging spot immediately.

Speaker P

When the sound of an electric saw is heard on a solar cellphone, the system sends a signal to the cloud. A real time alarm is sent to a nearby security guard, and the guard can be dispatched immediately.

Speaker Q

People randomly lumber trees. The rainforest system detects them and send a signal to the security guard through the cloud and quickly dispatches them to the site.

Speaker R

The first, illegal logging is hear. And the second, RFCX detect the sound. Third, we're sending a signal to the forest ranger. Fourth, forest ranger are aware of the location of illegal logging.

Speaker S

Install rainforest connection on the trees. Then when rainforest connection does illegal logging, the microphone sensitively catches the sound of the electric saw, finds the location and receive the location from the forest guard through the cloud to have the forest guard move quickly.

Speaker T

RFCX hears the sound wave and send a signal to the administrator through the cloud. And the administrator received the signal and dispatches it.

Speaker U

First, RFCX catches the noise from the chainsaw. Second, it sends a signal including the location of the illegal loggers to your phone via the cloud. Finally, you can go the place and arrest the illegal loggers.

Speaker V

If someone is logging, the microphone attached to the tree will detect a noise and send a signal to cloud with solar power and send a signal to forest rangers.

Speaker W

If you try to do illegal logging, first the invention attached around here informs the forest ranger and come to the site.

Speaker X

First, when you hear a saw cutting wood, the machine detects it and sends it to the cloud and the cloud knows the guard, the guard goes to the place where he cuts trees and stop them.

Speaker Y

When loggers cut a tree with a chainsaw, the machine recognize the sound and sends our radio wave. Radio wave transmits to the forest guard of cloud app. The location of the logger can be determined.

Speaker Z

First, the sounds of illegal logging is caught on the cellphone microphone. Additionally, cellphone receive electricity from solar energy. Second, software send a signals to cloud. Third, a nearby forest rangers get a warning calls.

Appendix 10. Main text (p.83~87, High School English, YBM, Han et al., YBM Holdings, 2018)

Lesson 4. Invention for the rainforests

RFCx: the Rainforest Savior

Imagine you are standing in a rainforest. You are surrounded by tall trees, many of which are more than 40 meters tall. You are a hundred kilometers away from the nearest city. What do you hear? Do you think it is a quiet, peaceful place? If

so, you are wrong. The rainforest is actually a very noisy place. Insects, birds, and monkeys are responsible for much of this noise. And sometimes there is another sound, one that does not belong in the forest at all. It is the buzz of a chainsaw. Every year some 13 million hectares of rainforest, an area about the size of England, disappears.

This loss destroys the habitats for millions of species and has a major effect on the jungle's biodiversity. Also, it increases the amount of CO2 in the air. Destruction of the rainforest is caused by logging, farming, mining, and other human activities. Among these, logging is the main reason for nature's loss. Some 70 to 80 percent of the logging in the rainforests is thought to be illegal. To address this problem, a young American engineer has invented a simple device that detects illegal logging the moment it occurs.

It all started in 2011, when Topher White visited Indonesia as a volunteer. One day, he and some of the other volunteers set out from the ranger station on a walk into a protected rainforest. After walking only five minutes, his group came upon people who were cutting down trees illegally. The surprised loggers fled, but White was shocked. Despite the fact that they were still fairly close to the ranger station, it had been impossible to hear anything from back there. It is because the forest was so full of other sounds.

White started thinking about ways to help. He knew that even in the jungle, far from the city, there was good cell phone service. He thought that perhaps cell phone technology could solve the problem. After he returned home to the U.S., in his father's garage he developed a small listening device using an old cell phone. He attached a sensitive microphone to the cell phone so that it could detect chainsaw noise from up to three kilometers away. This device would be placed high up in a tree. When it picked up the buzz of a saw, it would send a message to a ranger's cell phone.

White knew that he had to protect the cell phone so that it could survive in the hot and wet rainforest environment. His solution was to put the phone in a plastic box. Since there was no electricity where the phone needed to be placed, the device had to be able to power itself. White attached solar panels to the cell phone. He was sure that the panels would work, even under the shade of the thick tree leaves.

How the Device Works

1. It all starts here! Sound of chainsaws is picked up by microphones in solar-powered cell phones.
2. Software sends a signal to cloud.
3. Real-time alert is received by a ranger on the ground nearby.
4. That enables the rangers to go to the site immediately.

White returned to Indonesia to test the device. Surprisingly, on only the second day after he installed the device, it picked up chainsaw noises. An alert message was immediately sent to White and the forest rangers. When they approached the logging spot, the illegal loggers ran away.

White published his story on the Internet and word quickly spread. People living in other countries contacted White and asked if they could use the device. Others, from around the world, started sending him their old cell phones so he could build more devices. These devices, called Rainforest Connection (RFCx), are now being used in the rainforests in Africa and South America.

One RFCx can protect 300 hectares of forest. If a forest of this size is cut, 15,000 tons of CO2 are released into the air. Preventing this amount of CO2 from being released has the same effect as taking 3,000 cars off the road for a year. These devices are saving rainforests and providing new life for thousands of discarded cell phones. Thanks to Topher White and his RFCx devices, the earth is now a better place to live.

Appendix 11. One way ANOVA test result

Descriptive statistics

		Levene Statistic	df1	df2	Sig.
4_grams	Based on Mean	13.752	2	21	.000
	Based on Median	10.583	2	21	.001
	Based on Median and with adjusted df	10.583	2	14.563	.001
	Based on trimmed mean	13.187	2	21	.000

Robust Tests of Equality of Means

4-grams	Statistic[a]	*df1*	*df2*	*Sig.*
Welch	34.990	2	10.563	.000
Brown-Forsythe	49.636	2	10.636	.000

a. Asymptotically F distributed.

Multiple Comparisons

Dependent Variable: 4-grams

Games-Howell

(I) level	(J) level	Mean Difference (I-J)	Std. Error	*Sig.*	95% Confidence Interval	
					Lower Bound	Upper Bound
low	med	-.040	.016	.080	-.085	.005
	high	-.262*	.031	.000	-.353	-.172
med	low	.040	.016	.080	-.005	.085
	high	-.222*	.034	.000	-.317	-.129
high	low	.262*	.031	.000	.171	.354
	med	.222*	.034	.000	.129	.317

*. The mean difference is significant at the 0.05 level.