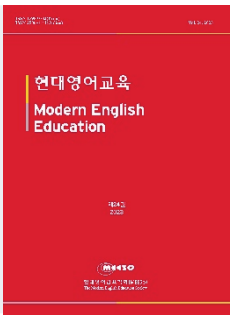
Modern English Education, Vol. 24, (2023) pp. 91-109
doi: https://doi.org/10.18095/meeso.2023.24.1.91

Modern English Education

ISSN: 1598-0782 / e-ISSN: 2586-6141
http://journal.meeso.or.kr



Validity Argument for Post-Entry Oral Performance Assessment for International Students in Korean Higher Education

Jiyoung Han (Hankuk University of Foreign Studies)
Kilryoung Lee (Hankuk University of Foreign Studies)

Received: 28 April 2023
Revised: 25 May 2023
Accepted: 1 June 2023

Han, Jiyoung, & Lee, Kilryoung. (2023). Validity argument for post-entry oral performance assessment for international students in Korean higher education. *Modern English Education*, *24*, 91-109.

Keywords
International students' English speaking ability, diagnostic oral performance test, test validation
국제학생의 영어말하기능력, 영어 말하기 진단평가, 시험 타당도 검증

Jiyoung Han (First author)
PhD Candidate
Department of TESOL
Hankuk University of
Foreign Studies
hufsmonica@gmail.com

Kilryoung Lee (Corresponding author)
Professor
Department of TESOL
Hankuk University of
Foreign Studies
klee@hufs.ac.kr

* This work was supported by the 2023 research fund of Hankuk University of Foreign Studies.

Abstract
With an increasing number of international students in Korean universities, communication breakdowns among international students, domestic students, and faculty members have been an important issue. However, their language ability and difficulties have received little attention. Since English is a lingua franca, this study evaluated their English oral communication ability by designing a post-entry oral performance assessment (PEOPA), conducting a pilot test, and investigating the validity of PEOPA based on Toulmin's (2003) argument structure. Firstly, a survey was conducted with 15 international students studying at a Korean university to ask about their English use and identify the target language use (TLU) domain. Among them, 11 took pilot PEOPA. A many-Facet Rasch analysis indicated that PEOPA tasks stably assessed the oral communication ability of participants and reliably separated them into different ability levels. Raters consistently assessed performances according to a scoring rubric reflecting intended construction, with each scoring category measuring distinct aspects of oral communication ability. Along with Rasch analysis, qualitative analysis of participants' transcribed performances provided further backing for test validity. Ensured validity of the assessment allowed us to propose the implementation of PEOPA in universities in similar contexts. Finally, the implications of findings for the diagnosis phase of PEOPA are discussed.

INTRODUCTION

English-speaking Western countries have attracted international students for several decades. More recently, countries in Asia such as China, Japan, South Korea, and Taiwan have been emerging as study abroad destinations (Chang, 2021; Jon, 2013). The increasing number of international students in this region is attributed to the efforts of the governments and of institutions to internationalize their higher education (Chang, 2021) in order to strengthen competitiveness in the globalized academic environment (Altbach & Knight, 2007) coupled with the rising interest in the region as its economy has grown (Jon et al., 2014).

In the case of South Korea, the efforts of the government seem to have borne fruit. The number of international student enrollments steadily increased from 12,314 in 2003 to 152,281 in 2021 (Ministry of Education, 2022). Among them, majority of the population (138,343 out of 152,281) is comprised of students from Asian countries. Consequently, English is not the first language for most of the international students studying in South Korea, but the language they continue to make efforts to improve (Chang, 2021). To be admitted to universities in South Korea, there are several types of entry pathway. For example, a private university in Seoul allows three tracks for undergraduate applicants: English language track, Korean language track, and English and Korean track (Language Requirements, 2022). Some tracks require both Korean Proficiency Test (TOPIK) score and English proficiency test score (e.g., scores above TOEFL iBT 80 or IELTS 5.5) and others require either Korean or English test score. For graduate applicants, either TOPIK score or English proficiency test score is required. This admission system seems to cause dynamic and diverse language use on campus. However, at the same time, the admission system is likely to cause communication breakdowns among students and faculty members since those who are fluent only in English have a good chance of interaction with those who are fluent only in Korean language. Such interaction might cause communication breakdowns, which needs to receive attention since language barrier has been found to affect the international students' adjustment to the host country (Brown, 2008; Campbell & Li, 2008; Poyrazli & Kavanaugh, 2006) and also affect one's self-efficacy and self-confidence (Yu et al., 2019).

It is noticeable that, despite the necessity, few studies investigated the language use and language difficulties of international students in EFL context. Particularly, most of the studies on international students studying in Korea has focused on sociocultural aspect such as adaptation difficulties, power hierarchy (Csizmazia, 2019; Jon, 2013; Jon et al., 2014; Thibault, 2022) whereas language-related issues have received little attention. As it is expected that the number of international students on campus would increase and communication breakdowns become an issue, Korean universities might refer to the post-entry language assessment (PELA).

PELA is a diagnostic test, which was designed to support the students from diverse linguistic backgrounds (Doe, 2014; Read, 2008, 2015) studying in English-speaking countries. Once the students are admitted to the institution, they are required to take a PELA. For those who obtain scores below the cut-off score, language support is provided. Each university can develop a customized PELA to meet their own needs (Read, 2015).

Acknowledging the potential benefits of PELA for international students studying in Korean universities, the study developed a pilot post-entry oral performance assessment (PEOPA). The purpose of the study is to validate the PEOPA based on Toulmin's (2003) validity argument model. To enhance the validity of the test, the survey was conducted with international students prior to test development, which asked the needs of English oral communication ability on campus and identified the TLU (target language use) domains for English use. The test was designed based on the survey results and assessed a small sample of international students' English oral communication ability using it.

LITERATURE REVIEW

Validity Argument

The argument-based framework enables one to develop a valid test and to justify the uses of observed scores through collecting evidence over the whole testing procedures (Kane, 2013). Toulmin (2003) suggested the basic argument structure and terminology from which further argument-based frameworks could be developed. According to him, an argument comprises claims, data and warrants. A claim is the interpretation we make based on the data. The data is the information we collect, such as test taker's performance and responses. A warrant is a statement that we propose to justify the inference, which is based on backing. The backing assures the warrants as providing authority to the warrant. Theory, previous studies and evidence collected during testing process can be the backing (Bachman, 2005). Drawing on this framework, Kane (2012, 2013, 2021) addressed the notion of interpretive argument, which involves the link between interpretation and assumptions,

which require evidence to support. Moreover, Kane extended the linkages to the decisions (Bachman, 2005) we make based on the test scores.

Post-entry Language Assessment

Acknowledging the struggles that post-secondary students from diverse linguistic background had in Australia and New Zealand, the Diagnostic English Language Assessment (DELA) was designed in 2002 at the University of Melbourne in collaboration with the University of Auckland (Doe, 2014; Read, 2015). It aimed to provide the students at risk with diagnostic feedback and further remedial language support. It is now known as post-entry language assessment (PELA) in Australia (Read, 2015). To identify students' language learning needs and to provide helpful guidance, clarifying target language use (TLU) domain of the students in their context is necessary (Y. Lee, 2015).

PELAs take several different forms in terms of organization, design, content, mode, and target cohort (Read, 2015). Usually, a PELA consists of a two-phase process: screening and diagnosis. The students who are placed below cut-off scores in the screening phase are required to take the diagnosis phase. Regarding diagnostic solutions, studies have discussed various language development activities and their effectiveness. D'Silva and Kinnear (2021) described the case of implementing PELA with first-year engineering students in Canada. Through the screening phase, students were classified into three bands based on their scores. Students in Band 1 require additional academic English language support; those in Band 2 have a good foundation in English, but still require some additional support; and those in Band 3 have sufficient academic English skills and do not need further. Language support programs include assignment-specific workshops, skill-based workshops, and in-course support, where instructors guide students to improve academic and communication competencies.

While the PELAs target academic language ability required to study in English medium instruction (EMI), the present study aimed to develop a test, which assessed the speaking ability required for communications that occur on campus, since not all the international students in South Korea are enrolled in an EMI program. The target construct will be referred as oral communication ability in the remainder of the study.

English Language and International Students in Non-English Speaking Countries

Substantial body of research has claimed that English proficiency affected the quality of international students' life in English-speaking countries (Brown, 2008; Muller, 2011; Poyrazli & Kavanaugh, 2006; Yu, 2013) as facilitating their academic, sociocultural, and psychological adjustment (Campbell & Li, 2008; Roberts et al., 2018). In spite of the growing number of international students in EFL contexts, however, few studies have investigated their English language proficiency, language difficulties, and learning needs in the region. Among the few studies, Yu et al. (2019) conducted a survey with international students in universities in Hong Kong to examine the relationship between language proficiency and adaptation. They found that English language proficiency contributed to both psychological and sociocultural adaptation while Cantonese proficiency was interrelated only with sociocultural adaptation.

In Korea, little attention has been paid to international students' English use while several studies investigated the use of Korean language of international students (Choi, 2018; Park, 2016; Thibault, 2022). Although the focus of the study was not solely on English language, two studies indicated that English fluency of international students might affect their life on campus. Jon (2012) investigated power dynamics between domestic students and international students attending a university in Korea. Interviews with domestic students suggested that English language use would empower international students. Domestic students who had participated in the campus buddy program displayed more positive attitudes toward international students who spoke English. Similarly, Park (2016) found that international students who were fluent in English had better relationship with domestic students. It was in contrast to the researcher's assumption that international students who were fluent in Korean language would have better relationship with domestic students. It was also found that those who were fluent in English had better relationship with other international students who were from different countries as well. Meanwhile, Choi (2018) found that the level of Korean language proficiency of international students has decreased and explained that some of international students were accepted in Korean universities despite the insufficient level of Korean fluency. It is worthwhile, therefore, to investigate their English language ability since the students who are fluent in neither English nor Korean might have difficulties in communication with other students and professors. In addition, aforementioned studies suggest that English language proficiency might assist international students' adaptation to the host country as well as empower them in relationships with domestic students.

Present Study

As described in the previous section, few studies paid attention to English use of international students studying in Korean universities despite the necessity. Therefore, to identify the needs of English use and TLU domain, the study conducted survey with international students in a Korean university. Then, a pilot PEOPA was developed based on the survey results and assessed a small sample of NNS international students' English oral communication ability using it. The study took the procedures presented in Figure 1.

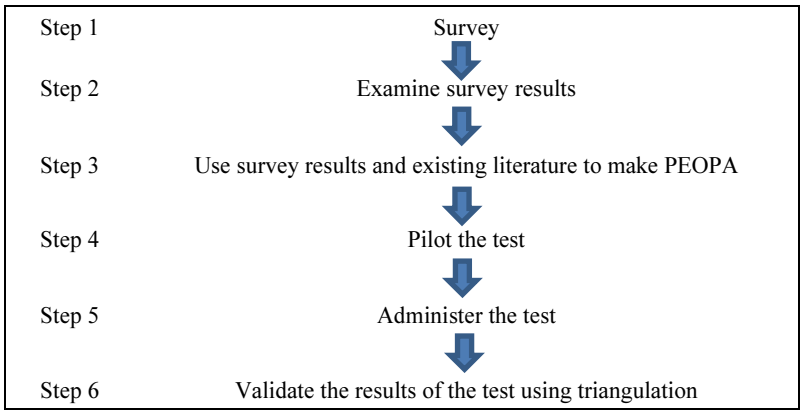


FIGURE 1
Validating Procedures of PEOPA

Figure 2 illustrates how interpretation of the PEOPA scores is linked to the data, backings, and warrant. Preliminary, the survey was conducted with international students to identify the needs of English oral communication on campus. The responses to the questionnaires were analyzed to obtain evidence of the needs of English oral communication ability on campus. (Backing 1). Then, to examine the validity of test components including PEOPA tasks, rating criteria, and raters' consistency, a many-Facet Rasch measurement (MFRM) analysis was used (Backing 2). In addition, to examine the language use of the participants in-depth, transcribed performances were analyzed qualitatively (Backing 3). All these backings can be used to rationalize the use of test results and to interpret that international student who got scores below cut-score would need English language.

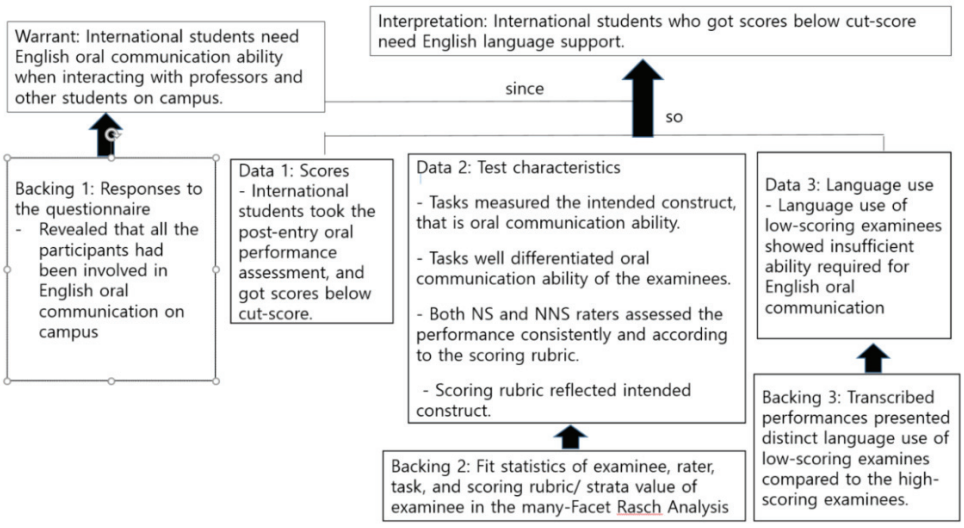


FIGURE 2
Structure of the Validity Argument for Post-Entry Oral Performance Assessment

The following research questions guided the study.

Research Question 1. What are the English language needs of NNS international students at a Korean university?

Research Question 2. Do PEOPA scores reliably separate the participants into different levels of oral communication ability?

Research Question 3. Do PEOPA tasks stably assess oral communication ability?

Research Question 4. Do the raters assess the performances consistently and according to the scoring rubric?

Research Question 5. Does the scoring rubric reflect the same and intended construct, which is oral communication ability?

METHOD

Survey

Participants

The questionnaires were e-mailed to international students who were participating in a mentoring program of a private university located in Seoul with the help of Graduate Student Council of the university. A total of 15 international students were asked to answer the questionnaires. The mentoring program has been carried out since 2017 to facilitate the adjustment of international students to campus life by matching them with domestic students. At the time of the study, 10 international students were involved in the program as mentees. Five mentees responded to the questionnaire. With snowball sampling, along with five mentees, a total of 15 international students participated in the survey. The age range of the participants was from 22 to 37. Two were undergraduate students, one was taking a 10-week translation and interpretation course, and the rest of the participants were taking graduate course. Six students responded that they had taken more than one EMI classes. Three participants had L1 background of Vietnamese, seven with Chinese, two with Cantonese, one with Uzbek, one with Spanish and one was French. Among participants with Chinese L1 background, three were from Taiwan. The length of stay in Korea varied from two months to 7 years. Their major included TESOL, teaching Korean as a foreign language, translation, international studies, global culture and contents, and beauty art. Four participants were more fluent in English than Korean language whereas 11 were more fluent in Korean language. The L2 of the students who were more fluent in English than Korean language was French, Spanish, Uzbek, and Chinese.

Instrument

The questionnaire was designed in order to identify the needs of English oral communication ability and TLU domain of international students studying in a university in Seoul. The items consisted of three sections: (a) demographic information, (b) occasions and interlocutors of English conversation on campus (TLU domain), and (c) need for further English conversation class. (see Appendix A). The first section of the questionnaire asked about respondents' L1, age, length of residence in Korea, institutional status, and in which language the students were more fluent. To identify their English oral communication needs, the second section asked whether they have been involved in English conversation on campus, and about the occasions and interlocutors of the English conversation. The participants could choose multiple responses to the question asking about interlocutors and they were asked to describe the situation and topic of the English conversation they were involved. The final item asked whether they were willing to take extra English conversation class. Total 13 items were included in the questionnaire. The survey items were provided both in English and Korean language.

PEOPA

Participants and Raters

Among survey participants, 11 participants volunteered to participate in taking the PEOPA and signed consent forms. Only one participant (E7) had English speaking test scores (IELTS Speaking 7.5). Their self-perceived English oral proficiency

varied as seen in Table 1. All participants have learned English in public school since the third grade in elementary school except for E2, who started English learning in kindergarten, E4, in middle school, and E8, who first learned English in university. E3 had taken English Medium Instruction (EMI) program for six months in Malaysia, and E4 had lived in the Philippines for a month to learn English. The others had no experience of living in English-speaking countries. E8 was male and the others were female. Academic status was not deemed to affect the scores because the task topics did not relate to the specific academic area as presented in Table 2.

Given that international students encounter both a native speaker (NS) of English and a non-native speaker (NNS) on campus, raters consisted of two native English speakers and three non-native English speakers. They were asked to provide information on experiences in L2 assessment, TESOL (Teaching English to Speakers of Other Languages) experience, and rater training on a questionnaire drawn from H. J. Kim (2015). The NS raters had a TESOL degree or a certificate. Teaching English in Korea more than 25 years, they had extensive experience of test development and rating, particularly of oral assessments. Three NNS raters were Korean PhD candidates in TESOL with extensive teaching and rating experience. One of them had not taken the assessment class while two had taken it. One of the NS raters and one of the NNS raters were male.

TABLE 1
Demographic Information of the PEOPA Participants

ID	L1 Background	Age	Academic Status	Self-Perceived English Speaking Proficiency
E1	Chinese	28	Graduate	Intermediate
E2	Chinese	26	10-week translation and interpretation course	Intermediate
E3	Chinese	22	Graduate	Intermediate
E4	Chinese	27	Graduate	Low
E5	Chinese	37	Undergraduate	Intermediate
E6	Chinese	25	Graduate	Low
E7	Vietnamese	26	Graduate	Intermediate
E8	Uzbek	32	Graduate	Advanced
E9	Chinese	26	Graduate	Intermediate
E10	Chinese	21	Undergraduate	Low
E11	Vietnamese	24	Graduate	Intermediate

Instruments

Since the participants in the study have already passed admission procedures, administering post-entry language assessment was deemed appropriate. In addition, since PELA has a diagnostic purpose, the test is expected to identify weaknesses of the participants and be able to provide linguistic diagnosis. Whereas the original PELA aims to assess the academic language ability, the present test aims to evaluate the oral communication ability of the participants, not academic English competence as described in the previous section. Those who get scores above cut-off score are regarded to have sufficient speaking ability for daily English communication with professors and other students on campus. On the other hand, those who are placed below the cut-off score are assumed to need language support.

The open role-play, which does not inform situational scenario unlike closed role-play (Kasper & Dahl, 1991) was adapted for PEOPA since it has been found to reflect an interaction that resembles a real-world interaction (Taguchi, 2018). The participants were asked to conduct open role-plays with the interlocutor. In order to avoid the effect of the interlocutor, a PhD candidate in TESOL, whose L1 was Korean language, participated as an interlocutor. She was trained to standardize the interaction with the participants with the same scenario, which was not informed to the participants prior to the test. For

example, when the participants asked the interlocutor to lend a book one more week in T2 (Task 2), she gave same answer ('no') to all the participants.

Tasks were developed based on the identified TLU domain from the responses to the questionnaire and referred to the previous study (Youn, 2015) as described in Table 2. The PEOPA tasks included various speech acts including apologizing, requesting, negotiating schedule, and refusing. Following the suggestion of the literature (Roever & Ikeda, 2022) and reflecting TLU domain, the interlocutor played three roles, that is, roles of a roommate, a professor and a classmate.

TABLE 2
Task Topics

Role-play with your roommate	T1. Introducing each other at the first encounter T2. Apologizing about the delayed- return of a book your borrowed and asking her if you can borrow it one more week. (It has been a month since your borrowed the book) T3. Asking how to use online library to search for academic material in order to write a term paper
Role-play with a professor	T4. Asking if she has time now to discuss your project (you're visiting her office) T5. Requesting an extension of the assignment submission date after finishing the class
Role-play with your classmate	T6. Deciding on a meeting time for the discussion of a group project T7. Refusing her suggestion to meet face-to-face and suggesting meeting online

The test measured: *Content delivery, Interaction, Language use*, *Phonology*, and *Fluency* in a 5-point scale. *Content delivery* assesses to what extent a test-taker can convey the literal and intended meanings of an utterance that are required by PEOPA tasks. *Interaction* measures to what extent the performer can understand the interlocutor's message and respond accordingly by using communicative strategies and preliminaries (May, 2011). *Language use* referred to the degree to which accurate and complex lexical and syntactic forms can be used. For *Phonology,* pronunciation, intonation, and rhythm were evaluated. *Fluency* assessed to what extent the performer can produce fast and smooth production without hesitations and silent pauses.

Test Administration

After conducting a pilot test with two Korean students, PEOPA was administered. Each participant played open role-plays with the interlocutor. Before the test, topics were explained to the students. Planning time of two minutes was given before starting the test. Each participant was given two minutes to complete each task. It took 16 minutes to complete the seven role-plays including two-minute preparation time. The performances were recorded via Zoom. The participants were told to turn off the camera if they wanted to.

Rating

The rater training was given to NS rater group and NNS rater group separately via Zoom. First, they were introduced to the purpose of the test and PEOPA tasks. Explanation of each component of the rating scale (see Appendix B) and rating practice on pilot performances of different participants were followed. The pilot performances were not included in the data set. The raters exercised rating based on the rubric and provided verbal comments on the reasons for their rating decisions. After completing assessment, scores assigned by the raters and the researcher were compared and discussed.

At the actual scoring process, the NS rater group and NNS rater group scored separately. Three NNS raters took part in a rating session via online. They assigned scores to each recorded performance independently. To avoid raters' fatigue, the first session evaluated five participants' performances and the second session, where six participants' performances were assessed, was followed after a 30-minutes break. Two NS raters took the same procedure. The raters evaluated each participant's performance of each task using a five-point scale on each of the five rating criteria. The discussion among raters led to the decision of making cut-off score at 3 points (mean score) out of 5 points (perfect score). In other words, participants who obtained mean score less than 3.00 were deemed to need language support.

Data Analysis

As for the survey responses, responses to question 8, 9 and 10 were examined and analyzed to identify the needs of English

oral communication on campus and TLU domain. The frequencies of the responses to the multiple choice items were calculated whereas the responses to the open-ended question asking TLU domain were examined and categorized into emerging domains. The responses were used to develop PEOPA tasks.

To obtain overall information on the results of POEPA, descriptive statistics including the mean, standard deviation, minimum and maximum scores, skewness, and kurtosis were calculated. Cronbach's alpha (α) was also calculated to examine internal consistency reliability of PEOPA tasks. All of these were estimated using R. Following descriptive statistics, FACETS analysis was employed. To answer the research questions, four-facet Partial Credit Model (PCM) was implemented. Logit and strata value of participant, task, rater, and scoring rubric facet were examined to answer research question 2, 3, 4 and 5 respectively. Infit mean square values, rather than standardized values were used to interpret the results since it is less sensitive to sample size and is information-weighted (Bond & Fox, 2015). Concerning the range of acceptable values for mean-square infit statistics, a lower-control limit of 0.6 and upper- control limit of 1.4 were adapted for the study following Bond and Fox's (2015) suggestion for Likert type tests. In addition to the statistical analysis, all participants' performances were transcribed using software *Happyscribe*, then checked and corrected manually by one of the researchers. Transcribed performances of three low-scoring participants were compared to those of three high-scored participants for each task.

RESULTS

Survey Results

Research Question 1: What are the English language needs of international students at a Korean university?

As for the experience of English oral communication, all the survey participants said they had been involved English conversation. Eleven participants responded the interlocutors were native English students, which ranked the first. Then, eight participants responded that they had English conversation with NNS international students and seven participants with NS professors. Five participants had experience of speaking English with Korean students whereas only three had with Korean professors.

Analysis of responses to the question asking TLU domain resulted in two major domains: academic situation and daily-life situation. Seven participants described academic situation, such as English discussion in class, conversation with professors, conversation about assignment and test with other students. Eleven participants described non-academic situation. In daily-life situation, the interlocutors included professors and other international students, both NS and NNS. The topic included food, shopping, and movie. One participant whose L1 was Cantonese responded that she spoke English when the interlocutor did not speak Korean or when she could not come up with appropriate Korean words. Similarly, the participant with Chinese L1 responded that she spoke English when there was communication breakdown with other international students when engaged in Korean conversation. The participant with Vietnamese L1 described that she lost her towel at a laundry room and talked with other international student who took it by chance. The students who were more fluent in English than in Korean responded that they usually spoke in English with classmates and roommates.

Test Results

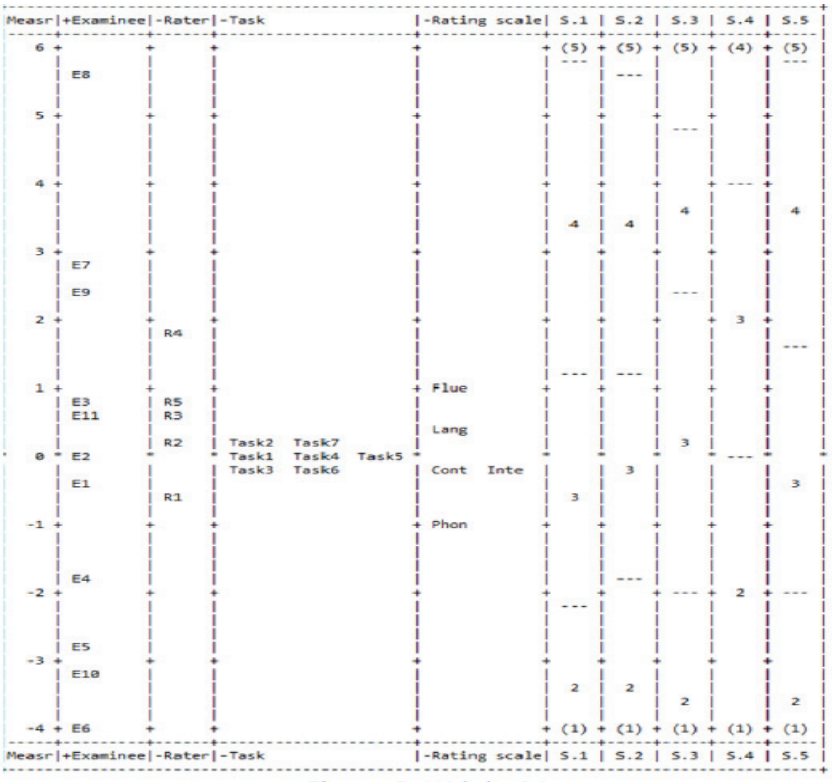
In order to explore overall testing results, descriptive statistics of each task were calculated as shown in Table 3. The mean score of each task was somewhat similar: 2.78, 2.72, 2.8, 2.79, 2.75, 2.79, and 2.77 for T1 through T7, respectively. The standard deviations of all PEOPA tasks were also similar (from 0.76 to 0.84). Both skewness and kurtosis values of each task were within the acceptable range, indicating normal distribution of scores (Bachman, 2004). Although the mean scores over PEOPA tasks were similar, differences of range were identified across PEOPA tasks (from 1.6 of T6 to 2.76 of T2). It seems that T2 (*Apologizing to your roommate about delayed return of the book*) seemed to be particularly difficult for lower-leveled participants than other PEOPA tasks while it was not difficult for a participant in higher-level. Comparing minimum scores and maximum scores, one notable pattern was observed only with minimum scores. Minimum scores showed an increasing pattern as the test proceeded whereas maximum score did not. It is assumed that task anxiety of the participants in low-level had decreased as they felt less nervous as time went by, resulting in such a pattern. Although the maximum score did not display the same pattern, it is noticeable maximum score of T1 was the lowest. Given that T1 was to introduce

oneself to a roommate, which should be familiar to the participants, it is assumed that the high-scoring participants might have felt nervous as well at the beginning of test. Since this can be regarded as a variable to the test validity, it needs to be investigated further. The reliability estimate for the PEOPA was 1.00, a very high level of internal consistency among tasks.

TABLE 3
Descriptive Statistics of the Post-entry Oral Performance Assessment

	N	Mean	Median	Std. Deviation	Skewness	Kurtosis	Range	Minimum	Maximum
T1	11	2.78	2.72	0.76	-0.13	-0.93	2.68	1.32	4
T2	11	2.72	2.68	0.82	0.18	-1.01	2.76	1.48	4.24
T3	11	2.8	2.76	0.79	0.42	-1.05	2.6	1.72	4.32
T4	11	2.79	2.96	0.79	-0.04	-1.3	2.56	1.56	4.12
T5	11	2.75	2.8	0.81	0.32	-0.97	2.6	1.76	4.36
T6	11	2.79	2.96	0.84	0.04	-1.32	1.6	2.68	4.28
T7	11	2.77	2.8	0.77	0.42	-0.72	1.64	2.72	4.36

Figure 3 presents a Wright map of many-FACET Analysis. The first column presents measurement, the logit scale. The second column presents the estimates of the participants' ability. The highest scoring participant is located at the top of the column, while the lowest scoring participant is located at the bottom of the column. The results displayed a wide distribution of participants' oral communication ability spanning 9.65 logits (M = 0, SD = 2.88). The third column displays the severity and leniency of raters. The raters located higher in the column exercised more severity when scoring the performances. The fourth column displays the difficulty level of seven PEOPA tasks. Tasks located in higher place were more difficult for participants than tasks appearing lower. The fifth column represents the relative difficulties of five scoring criteria. The last column presents the five-point rating scale, with category thresholds, which indicates the point at which the score moves onto the next higher level in the given criteria.



Note. S.1 = Content delivery; S.2 = Interaction; S.3 = Language use; S.4 = Phonology; S.5 = Fluency.

FIGURE 3
Wright Map

Research Question 2: Do PEOPA tasks reliably separate the participants into different levels of oral communication ability?

As seen in Table 4, all participants' infit mean square values were within the acceptable range (0.6 – 1.4) except for E5, whose infit value indicates .51 logit, slightly below the low-limit. Observation of her performance did not show any erratic behavior. When applying a less robust infit value range, 0.5 to 1.5, (Lunz et al., 1990), her infit value is not likely to be misfit. Furthermore, participants misfit might not be a major issue in this kind of assessment because there are no chances of random guessing or careless answering (Bonk & Ockey, 2003). The strata value was 25.61 with a reliability of 1.00. This can be interpreted that the seven tasks separated the oral communication ability of the participants into approximately 26 levels.

TABLE 4
Measurement Report for Participants

Participants	Ability Logit	Model Error	Infit Mean Square	Observed Score
E8	5.66	.16	1.10	4.24
E7	2.71	.15	.84	3.51
E9	2.44	.15	.92	3.45
E3	.77	.15	1.10	3.05
E11	.55	.15	.97	3.02
E2	-.08	.14	1.17	2.79
E1	-.39	.14	1.04	2.70
E4	-1.79	.14	.97	2.28
E5	-2.71	.14	.51	1.99
E10	-3.17	.14	1.05	1.95
E6	-3.99	.14	1.25	1.61

Note. Strata = 5.61; Reliability = 1.00; Fixed chi-square = 3788.1 (df = 10; p < .01).

Research Question 3: Do PEOPA tasks stably assess oral communication ability?

As seen in Table 5, the infit mean-square indices of all PEOPA tasks were within the acceptable range from .83 to 1.18. This can be interpreted that PEOPA tasks measured the same construct. Regarding with difficulties, *Apologizing your roommate about delayed-return of the book* was the most difficult one with the highest logit value of 0.17 and *Refusing her suggestion of meeting face-to-face and suggesting meeting through online* was similarly difficult with the logit value of 0.15. *Requesting a professor for an extension of the assignment submission date* followed next with the logit value of 0.05. The logit value of T1, *Introducing yourself at the first encounter,* was higher than T4, T6, and T3. As previously mentioned, this is likely due to the initial test anxiety of participants rather than due the difficulty of the task. *Asking how to use online library to search for the academic material* was the easiest one (logit = -.15), followed by T6, *Deciding on a meeting time for the discussion of a group project with the roommate* (logit = -.11).

TABLE 5
Measurement Report for PEOPA Tasks

Tasks	Difficulty Logit	Model SE	Infit Mean Square
T2	.17	.11	1.00
T7	.15	.11	.83
T5	.05	.11	.95
T1	-.02	.11	1.13
T4	-.10	.11	.93
T6	-.11	.11	1.18
T3	-.15	.11	.89

Note. Strata = .65; Reliability = .05; Fixed chi-square = 7.4 (df = 6; p = .65).

Research Question 4: Do the raters assess the performances consistently and according to the scoring rubric?

As presented in Table 6, the mean square infit value of the raters indicated that all of them were internally consistent and rated according to the rubric appropriately. No one showed a central tendency or halo effect. The strata index of 10.91 with a reliability of .98 indicates that the raters' severity varied despite the rater training. Two NS raters, R1 and R2 showed more lenient tendency whereas three NNS raters presented more severity. However, the degree of severity, which is regarded as a variable to threaten the validity (Eckes, 2015) was not a problem of this test because the same raters assigned all participants' scores.

TABLE 6
Measurement Report for Raters

Raters	Severity Logit	Model SE	Infit Mean Square
R4	1.70	.10	1.21
R5	.88	.10	1.00
R3	.67	.10	.88
R2	.21	.10	.79
R1	-.65	.10	1.06

Note. Strata = 10.91; Reliability = .98; Fixed chi-square = 315.7 (*df* = 4; $p < .01$).

Research Question 5: Does the scoring rubric reflect the same and intended construct, which is oral communication ability?

The infit statistics of the rating criteria showed all the criteria were within the acceptable fit as seen in Table 7. That is, each criteria functioned as intended and measured the same construct. Moreover, differing levels of difficulty logit for each rating scale showed that each criteria tapped into a distinct feature of oral communication ability. The most difficult criterion for the participants to get high scores was *Fluency* with the logit value 1.09 followed by *Language use* (.47 logit) and *Content delivery* (-.20 logit). On the contrary, *Phonology* was the least difficult criterion with the logit of -1.07 followed by *Interaction* (-.29 logit).

TABLE 7
Measurement Report for Rating Scale

Rating Scale	Difficulty Logit	Model SE	Infit Mean Square
Fluency	1.09	.10	.95
Language Use	.47	.10	.83
Content Delivery	-.20	.091	1.16
Interaction	-.29	.091	.95
Phonology	-1.07	.11	1.03

Note. Strata = 10.26; Reliability = .98; Fixed chi-square = 267.9 (*df* = 4; $p < .01$).

Qualitative Results

To identify the weaknesses of participants who obtained lower scores, their transcribed performances were compared to those of higher-scoring participants. Qualitative analysis showed that language use of lower-scoring participants were pragmatically less appropriate compared to that of higher-scoring participants. The following Excerpts compare how participants with different ability requested for an extension of assignment submission date from the professor at T5 (see Appendix C for transcription conventions). Excerpt 1 shows the performance of E8, who achieved the highest score of 4.24 whereas Excerpt 2 illustrates the performance of E5, whose scores ranked 9th with score of 1.99.

1) Excerpt 1 (E = Participant, P = Professor)

1 E: Yeah (.) I was wondering if you (.) could you know
2 extend the-ah-the deadline for (.) un assignment
3 submission?
4 P: oh::
5 E: you know I know today's the deadline (.) I mean the
6 Wednesday but-if it is possible for me to submit it on
7 on by Friday because (.) uh I'm kind of you know
8 late I couldn't finish it (0.2) so I need to review once
9 again and
10 P: hm [let me see
11 E: [develop for the presentations

2) Excerpt 2

1 E: Hi (0.2) professor::
2 P: Oh hi?
3 E: Yes (0.2) we-ah-we-(0.2) we have the:: have(more)
4 project? I have to:::have to update in list Friday?
5 right?
6 P: That's right. You have to submit (.) it by Friday
7 E: Oh yes (0.2) but (0.2) I have a little (problem)(.hhh)
8 so am:: I will (0.2) I can (0.4) update (???) list
9 Friday maybe I (can) maybe I
10 P: uh huh
11 E: Maybe I can finish it (0.2) update like a Monday?

While E8 could use more pragmatically appropriate expressions by using modal verbs *could* in the interaction with a professor as seen at line 1 in Excerpt 1, E5 had difficulties to make a request to the professor. She attempted to express her idea by using *we have*, *I have to*, and *I will* at the first time, then she could finally say *I can* at line 11. However, this sounds like a statement rather than a request. Without using *could* or *would*, E5 conveyed her message in somewhat direct way. Excerpts 3 and 4 display how a higher-scoring participant and a lower-scoring participant responded to the interlocutor's refusal at T2 (*apologizing about delayed-return of the book and asking her if you can borrow it one more week*). E2 who achieved the second highest score of 3.51 used acknowledging tokens "I see" as shown in Excerpt 3. On the contrary, E6, who achieved the lowest score of 1.61 did not use proper acknowledging tokens and then insisted she still needed the book as in Excerpt 4.

3) Excerpt 3 (R = Roommate)

1 R: Well:: one thing (.) I'd like to do (0.2) but I actually
2 need the book for my term paper
3 E: Uh huh:: so I see (0.2) um (0.2) I see OK:: I will
4 return you today? Is that OK?

4) Excerpt 4

1 R: So I'd like you to borrow it for another week (.) but
2 I'm sorry to say this (.) but I need the book now
3 E: ah::if I havn't (0.4) if I haven't (0.2) also I can not
4 (0.2) finish my paper

Although it is not sure if such a strict way of speaking was resulted from her lack of proficiency or her attitude, the interlocutor might feel unpleasant with this response, which might affect further interaction.

DISCUSSION

This section discusses how the findings supported the validity of the test. Furthermore, identified weaknesses of the participants and the diagnostic phase of the test are discussed based on the survey and test results. The first research question was about the needs of English language use. All the participants had been involved in English conversation on campus with international students, domestic students, and NS professors. The questionnaire responses showed that they were involved in more conversations that are in English with foreign students and professors than they were with Korean students and professors. The identified domains included both academic and non-academic situations. The results indicated that, therefore, international students studying in a Korean university needed English oral communication ability although they did not take EMI courses. The second research question asked how reliably the seven PEOPA tasks could separate participants into different levels of oral communication ability. The logits of participants and strata value in MFRM analysis indicated PEOPA tasks could differentiate participants' oral communication ability into 26 levels. Given the different language requirement for admission, it seems to be a natural consequence. The substantial disparity of oral communication ability among participants will be likely to cause communication breakdowns. For instance, there is a fair chance that E8, who achieved a score of 4.24 and E1, who achieved a score of 2.70 would communicate with each other in class or out of class because they belonged to the same department. Indeed, analysis of the response of E1 to the question that asked her experiences of English conversation on campus revealed that she had English discussions in class. It can be assumed that she might have had difficulties in discussion due to her insufficient English proficiency while E8 might have not. In addition, since English speaking ability can empower international students in a relationship with domestic students (Jon, 2012), the lack of English oral communication ability of E1 might affect her social adjustment and self-efficacy as well as academic achievement (Campbell & Li, 2008; Roberts et al., 2018). The relationship between English oral proficiency and adjustment of international students in EFL contexts needs further in-depth investigation.

The third research question asked about the unidimensionality of PEOPA tasks. Infit values (at or near 1.0 for all tasks) of task facet suggested PEOPA tasks measured one construct, thereby providing valid evidence of the unidimensionality of the tasks (Bonk & Ockey, 2003). In terms of the task difficulty, *Apologizing about delayed-return of the book and asking her if you can borrow it one more week* and *Refusing her suggestion to meet face-to-face and suggesting meeting through online* (T7) were most difficult for the participants. As explained in the literature (H. Kim, 2015), face threatening acts seem to be challenging to L2 learners. In addition to T2 and T7, T5, *Requesting for an extension of the assignment submission date from a professor* was similarly difficult to the participants. The difficulty in interacting with a professor was reported in Youn (2015). This might be due to the interlocutor's social status. Particularly, transcribed performances showed that the low-scoring participants tended to make a request in a blunt manner. Following T5, T1, *Introducing yourself to your roommate* was considered difficult. Along with MFRA, descriptive statistics indicated that this might be due to the test anxiety rather than due to the task difficulty. Although participants were given two-minute preparation time, the anxiety from the initial interaction with the interlocutor might have affected their performances. Therefore, it would enhance the validity of the test to exclude the score from T1 and keep the task as a warm-up.

Research question four was about consistency of the raters. Rater fit indicated all the raters consistently rated by using scoring rubric appropriately. However, the differing levels of raters' severity needs to be addressed. Discrepancies between raters in the rating severity even after rater training has been found in the literature (Eckes, 2015; Myford & Wolfe, 2000; H. J. Kim, 2015) due to the different background and characteristics of the raters (H. J. Kim, 2015; J. Lee et al., 2014; Lim, 2011). Thus, it has been suggested that the rater training needs to emphasize internal consistency rather than making efforts to diminish inter-rater variability. The rater infit values in the study, despite different degree of severity, suggested internal consistency of each rater, thus supporting the validity of score inference. It is noteworthy that three NNS raters were stricter than NS raters were. L1 of raters could have affected scoring. Interviews with raters about this issue could have confirmed L1 effect on rating. Considering that the international students have chances of interacting with students from diverse L1 background, it is suggested that the raters with more diverse L1 background are included in rater group

Research question five was about validity of scoring rubric. Infit value of each criteria provided objective backing for the criteria's appropriateness. Regarding with the difficulty of criteria, MFRM analysis showed *Fluency* was the most difficult to get high scores followed by *Language use* whereas *Phonology* and *Interaction* were easier. Although no one obtained a perfect score in *Phonology*, which means no one had native-like pronunciation, overall, participants had good and adequate control of pronunciation. Taking into consideration the increasing emphasis on L2 comprehensibility rather than L1-like accuracy (Crowther et al., 2016; Saito, 2020), this is an important finding that needs to be considered at the diagnosis phase. For this group of participants, L2 learning should focus on development of fluency and language use including vocabulary learning rather than on pronunciation improvement.

It is noteworthy that the MFRM analysis of *Interaction* criteria, which has been increasingly emphasized in oral assessment recently (May, 2011; May et al., 2020; Ockey & Chukharev-Hudilainen, 2021; Roever & Ikeda, 2022; Roever & Kasper, 2018), yielded a stable infit value. The study drew its operationalization from May (2011) and the descriptor seems to function as intended. The logits indicated that interaction was the second least difficult for the participants. It is likely that the participants might have had chances to meet new people in new circumstances, through which they could develop interaction ability. However, as qualitative analysis illustrated, the low-scoring participants need to expand their linguistic resources required for appropriate interaction.

Despite the efforts to enhance the validity of the test, the study has several limitations. First of all, findings cannot be generalized due to the small number of participants. Since all the classes within the university were conducted online to prevent the spread of COVID-19 at the time of the study, opportunities to meet students were extremely limited. Only five international students out of ten, who were participating in mentoring program, responded to survey via e-mail. Although more international students took part in the survey and PEOPA with snowball sampling, the number of participants is still small. Therefore, as for the test results, the study used infit mean square values instead of standardized values to interpret the PEOPA results because it is less sensitive to sample size. Based on the present pilot study, further study with more participants would be able to investigate the needs of English language and evaluate their English ability more accurately. Second, the survey items and response scales need to be elaborated. All the participants responded they had been involved in English conversation on campus. It implies that English language as well as Korean language is important for international students studying in Korean universities. Consequently, it is suggested to conduct a thorough needs analysis to measures the degree of English language needs, difficulties, and satisfaction or unsatisfaction of current English use of international students. The results would show how the universities could facilitate their English language needs. Third, the participants of the study were confined to Asian students. The results can be different when the students from different regions are included as participants. Raters with more diverse L1 backgrounds will also enhance validity to the test since the participants interact with interlocutors of diverse L1 background on campus. Lastly, although the present study limited TLU domain to outside the classroom, classroom situation will capture broader English use of the participants.

CONCLUSION

The purpose of the study was to identify English language use of international students who studying in Korean universities and validate the pilot PEOPA, which aimed to assess the oral communication ability of the participants and examine their weaknesses. Analysis of questionnaire responses, MFRM analysis and qualitative analysis of transcribed performances provided the convincing backings to the test validity. As the purpose of the PELA is to diagnose the language ability of the participants, the results of PEOPA revealed several weaknesses of the participants. Overall, scores at rating scale of Fluency and Language use were lower than other criteria. The participants felt more difficulty to apologize and to refuse. In addition, lower-scoring participants showed lack of linguistic resources required for appropriate L2 interaction. It is likely that providing English language support focusing on identified difficulties will improve participants' oral communication ability. Regarding the practical activities, previous studies have illustrated useful methods such as generic workshops, seminars, credit-bearing language development units, e-learning, and peer-to-peer program, which have been practiced in other institutions. These activities can serve as preceding examples for universities in Korea.

References

Altbach, P. G., & Knight, J. (2007). The internationalization of higher education: Motivations and realities. *Journal of Studies in International Education, 11*(3-4), 290-305.

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge University Press.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*, 1-34.

Bond, T. G., & Fox, C. M. (2015*). Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*, 89-110.

Brown, L. (2008). The incidence of study-related stress in international students in the initial stage of the international sojourn. *Journal of Studies in International Education, 12*(1), 5-28.

Campbell, J., & Li, M. (2008). Asian students' voices: An empirical study of Asian students' learning experiences at a New Zealand university. *Journal of Studies in International Education*, *12*(4), 375-396.

Chang, S. (2021). English medium instruction, English-enhanced instruction, or English without instruction: The affordances and constraints of linguistically responsive practices in the higher education classroom. *TESOL Quarterly, 55*(4), 1114-1135.
Choi, Hyeon-sil. (2018). A Study on the adaptation of foreign students to Korean universities. *Studies in Humanities and Social Sciences, 61*(4), 71-94.
Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*(2), 160-182.
Csizmazia, R. (2019). Challenges and internationalization of higher education in South Korea. *International Journal of Humanities and Social Science, 6*(6), 11-18.
Doe, C. (2014). Diagnostic English language needs assessment (DELNA), *Language Testing, 31*(4), 537-543.
D'Silva, F., & Kinnear, P. (2021). Developing disciplinary discourse in a first-year engineering course: The DELNA initiative. *Discourse and Writing/Rédactologie*, *31*, 126-135.
Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater- mediated assessments.* Peter Lang.
Jon, Jae-Eun. (2012). Power dynamics with international students: From the perspective of domestic students in Korean higher education. *Higher Education, 64*(4), 441-454.
Jon, Jae-Eun. (2013). Realizing internationalization at home in Korean higher education. *Journal of Studies in International Education*, *17*(4), 455-470.
Jon, Jae-Eun., Lee, J., & Byun, Kiyoung. (2014). The emergence of a regional hub: Comparing international student choices and experiences in South Korea. *Higher Education, 67*(5), 691-710.
Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 3-17.
Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1-73.
Kane, M. (2021). Articulating a validity argument. In G. Fulcher & L. Harding (Eds.), *The Routledge handbook of language testing* (pp. 32-47). Routledge.
Kasper, G., & Dahl, M. (1991). Research methods in interlanguage pragmatics. *Studies in Second Language Acquisition, 13*(2), 215-247.
Kim, Hyekyeng. (2015*). The effects of pragmatic instruction on Korean university students with different language proficiency levels*. [Unpublished doctoral dissertation]. Hankuk University of Foreign Studies.
Kim, Hyun Jung. (2015). A qualitative analysis of rater behavior on an L2 speaking assessment. *Language Assessment Quarterly, 12*, 239-261.
Language Requirements. (2022). *Hankuk University of Foreign Studies.* https://sites.google.com/view/internationalhufs/English-Admission-Guide/languagerequirements
Lee, Ji-oo, Lim, Hyun-woo, & Kim, Hyun Jung. (2014). An investigation into native English-speaking and Korean raters' judgments of Korean English learners' pronunciations. *Modern English Education, 15*(1), 195-216.
Lee, Yong-won. (2015). Future of diagnostic language assessment. *Language Testing*, *32*(3), 295-298.
Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28,* 543-560.
Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education*, *3*(4), 331-345.
May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*(2), 127-145.
May, L., Nakatsuhara, F., Lam, D., & Galaczi, E. (2020). Developing tools for learning oriented assessment of interactional competence: Bridging theory and practice. *Language Testing*, *37*(2), 165-188.
Ministry of Education. (2022). *Statistics of foreign students in domestic higher education institutions in 2021.* [Unpublished government document]. https://www.moe.go.kr/sn3hcv/doc.html?fn=48fb0857a562dd3141f0b6062d1fdba8&rs=/upload/synap/202202/
Muller, A. (2011). Addressing the English language needs of international nursing students. *Journal of Academic Language and Learning*, *5*(2), A14-A22.
Myford, C. M., & Wolfe, E. W. (2000). *Monitoring sources of variability within the Test of Spoken English assessment system* (TOEFL Research Report No. 65). Educational Testing Service.
Ockey, G. J., & Chukharev-Hudilainen, E. (2021). Human versus computer partner in the paired oral discussion test. *Applied Linguistics*, *42*(5), 924-944.
Paul, T. H. (2007). *Doing conversation analysis: A practical guide.* Sage.
Park, Soon-young. (2016). Foreign students' adaptation to college life in Korea: Focusing on the social relations. *Journal of Regional Studies, 24*(2), 75-102.
Poyrazli, S., & Kavanaugh, P. R. (2006). Marital status, ethnicity, academic achievement, and adjustment strains. *College Student Journal*, *40*(4), 767-780.
Read, J. (2008). Identifying language needs through diagnostic assessment. *Journal of English for Academic Purposes, 7*(3), 180-190.
Read, J. (2015). Issues in post-entry language assessment in English-medium universities. *Language Teaching*, *48*(2), 217-234.
Roberts, P. A., Dunworth, K., & Boldy, D. (2018). Towards a reframing of student support: A case study approach. *Higher Education*, *75*(1), 19-33.
Roever, C., & Ikeda, N. (2022). What scores from monologic speaking tests can (not) tell us about interactional competence. *Language*

Testing, *39*(1), 7-29.
Roever, C., & Kasper, G. (2018). Speaking in turns and sequences: Interactional competence as a target construct in testing speaking. *Language Testing, 35*(3), 331-355.
Saito, K. (2020). Multi- or single-word units? The role of collocation use in comprehensible and contextually appropriate second language speech. *Language Learning*, *70*(2), 548-588.
Taguchi, N. (2018). Data collection and analysis in developmental L2 pragmatics research: Discourse completion test, role play, and naturalistic recording. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 7-32*).* John Benjamin.
Thibault, N. (2022). *Student-faculty interaction experiences of international students in Korea: A phenomenology study.* [Unpublished doctoral dissertation]. The American College.
Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press.
Youn, Soo Jung. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing*, *32*(2), 199-225.
Yu, B. (2013). Asian international students at an Australian university: Mapping the paths between integrative motivation, competence in L2 communication, cross-cultural adaptation and persistence with structural equation modeling. *Journal of Multilingual and Multicultural Development, 34,* 727-742.
Yu, B., Bodycott, P., & Mak, A. S. (2019). Language and interpersonal resource predictors of psychological and sociocultural adaptation: International students in Hong Kong. *Journal of Studies in International Education*, *23*(5), 572-588.

Appendix A

Questionnaire

Hi, everyone. I am Jiyoung Han, Ph.D. student at TESOL department. I am conducting a survey for my research. The study aims to examine English interaction between non-native speakers of English. I would appreciate it if you answer to these questions and send me back if you are willing to: hufsmonica@hufs.ac.kr. This will be kept confidential and used only for the purpose of the research. Thank you so much!

Korean Version	English Version
1. 당신의 모국어는 무엇입니까?	What is your first language?
2. 나이는 어떻게 됩니까?	How old are you?
3. 한국에서의 체류기간은 얼마나 됩니까? 총 체류기간을 적으시오.	How long have you stayed in Korea? Please write down total number of years.
4. 현재 어떤 코스를 공부하고 있나요? (학사, 교환학생, 석사, 박사, 기타이면 서술)	Which course are you taking currently? (e.g., undergraduate, MA, Ph.D., exchange student, language program) Else: (please write down)
5. 한국에서 어떤 코스를 공부했었나요? 없으면 6번 문항으로 이동 (학사, 교환학생, 석사, 박사, 기타이면 서술)	Which course did you take before? If this course is your first academic experience in Korea, please move to question six.
6. 한국어와 영어 중 어떤 언어가 더 편합니까?	In which language, are you more fluent or confident, English or Korean?
7. 영어로 대화가 가능합니까? (1) 영어로 대화가 가능하다 (2) 영어로 대화가 불가능하다	Can you orally interact in English? (1) Yes (2) No
8. 한국에서 영어로 대화를 한 적이 있습니까? (1) 있다 (2) 없다	Have you ever been involved in English conversation since you came to Korea? (1) Yes (2) No
9. 한국에서 영어로 대화를 했을 때 상대는? (해당사항 모두 표시해 주세요) (1) 한국인 학생 (2) Native speaker 가 아닌 외국인 학생 (3) Native speaker 학생 (4) 한국인 교수님 (5) Native speaker 교수님 (6) 기타 (기술해 주세요)	Whom were you talking with when you spoke English? (1) With Korean students (2) With students who were not native speakers of English (3) With native speaker student (4) With Korean professors (5) With professors who were native speakers of English (6) Else ( )
10. 한국에서 영어로 대화를 했다면 어떤 상황, 어떤 주제였습니까? 생각나는 대로 모두 적어주세요. (예: 친구에게 서류를 어떻게 제출하는지 물었다, 등, 미국인 교수님에게 수업에 참석할 수 없다고 이야기했다, 등등)	If you have been in English conversation in Korea, please explain what the topic was and what the situation was. (e.g., I asked my Italian classmate how to submit some documents to school; I told my American professor that I could not attend the next class, etc.)
11. 현재 영어 (언어)수업을 받고 있습니까? 그렇다면 과목명을 적어주세요.	Are you currently taking English language classes? If yes, please write down the subject.
12. 영어로 하는 다른 과목 수업을 받고 있습니까? 그렇다면 과목명을 적어주세요.	Are you currently taking English-medium classes? If yes, please write down the subject.
13. 학교에 영어회화 수업이 있다면 들을 의향이 있습니까? (1) 있다 (2) 없다	If the university provides English conversation classes, do you want to take the course? (1) Yes (2) No

Appendix B

Analytic Scoring Rubric of PEPOA

Analytic Scoring Rubric of PEPOA					
Components	Contents Delivery	Interaction	Language Use	Phonology	Fluency
Description	To what extent can a performer convey literal and intended meanings of an utterance as required by the task?	To what extent the performer can understand the interlocutor's message and respond accordingly by using communicative strategies and preliminaries? To what extend can a performer take a turn?(e.g., initiating, responding, maintaining interaction)	To what extent does a performer use accurate and complex lexical and syntactic forms?	To what extent can a performer pronounce accurately? Segmental features: vowel, consonant, articulation Prosodic features: stress, rhythm	To what extent the performer can produce fast and smooth production without hesitations and silent pauses.
5	Completely clear Fully elaborated as required by the task	Fully understands the interlocutor's message and is able to respond accordingly. Successfully uses communicative strategies and preliminaries Effective turn taking	Grammatically accurate, a broad range of forms, simple & complex	Excellent control of segmental & prosodic features.	Very fluent, native-like speed, very smooth.
4	Generally clear – Fairly well elaborated as required by the task	Able to understand the interlocutor's message and is mostly able to respond accordingly. Generally able to use effective communicative strategies and preliminaries Appropriate turn taking	Mostly accurate. A relatively broad range. Relatively complex sentences.	Good control of s segmental & prosodic features.	Generally fluent, fast, and smooth.
3	At times unclear – Adequately elaborated and/or at times irrelevant elaboration to the task	Able to understand the interlocutor's message and may respond accordingly but with occasional lapses. May use communicative strategies but not always effectively, or appropriately. Acceptable turn taking	Fairly accurate, a somewhat narrow range, somewhat limited to simple sentences	Adequate control of segmental & prosodic features	Sometimes fluent, with some hesitations and pauses.
2	Often unclear – Inadequately elaborated and/or generally irrelevant elaboration to the task	Limited ability to understand the interlocutor's message. May not be able to respond. Use of communicative strategies and preliminaries is noticeably limited Inadequate turn taking	Several major errors, a narrow range of forms, often limited to simple sentences	Little control of segmental & prosodic feature	Limited fluency, a little slow, with hesitations, long silent pauses.
1	Generally unclear – Poorly elaborated	Unable to understand the interlocutor's message. Often cannot respond. Inadequate use of communicative strategies Not being able to take a turn	Almost always inaccurate, an extremely narrow range, only simple sentences	No control of segmental & prosodic feature	Not fluent, too many silent pauses.

Appendix C

Transcription Conventions (Paul, 2007).

?	Rising intonation
(.)	Tiny ‘gap’ within or between utterances.
(0.0)	Numbers in parentheses indicate elapsed time in silence by tenth of seconds, so (7.1) is a pause of 7 seconds and one-tenth of a second
::	Prolongation of the immediately prior sound. Multiple colons indicate a more prolonged sound.
-	Cut-off.
(word)	Parenthesized words are especially dubious hearings
.hhh	Inbreath