International Journal of English Linguistics; Vol. 10, No. 6; 2020
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

# Investigating the Validity of a University-Level ESL Speaking Placement Test via Mixed Methods Research

Becky H. Huang[1], Mingxia Zhi[1] & Yangting Wang[1]

[1] Department of Bicultural-Bilingual Studies, College of Education and Human Development, The University of Texas at San Antonio, San Antonio, TX, 78249, U.S.A.

Correspondence: Becky H. Huang, Department of Bicultural-Bilingual Studies, College of Education and Human Development, The University of Texas at San Antonio, San Antonio, TX, 78249, U.S.A. E-mail: becky.huang@utsa.edu

Received: June 10, 2020 Accepted: July 12, 2020 Online Published: July 16, 2020
doi:10.5539/ijel.v10n6p1 URL: https://doi.org/10.5539/ijel.v10n6p1

**Abstract**

The current study investigated the validity of a locally-developed university-level English as a Second Language (ESL) speaking placement test using a mixed-methods design. We adapted Messick's integrative view of validity (1996) and Kane's interpretation argument framework (2013) and focused on two sources of validity evidence: relations to other variables, and consequences of testing (AERA, APA, and NCME, 2014). We collected survey data from 41 student examinees and eight teacher examiners, and we also interviewed the teacher examiners about their perceived validity of the test. Results from the study provided positive evidence for the validity of the speaking test. There were significant associations between student examinees' speaking test scores, their self-ratings of speaking skills, and their instructors' end-of-semester ratings of student examinees' English language proficiency. Both the examinees and examiners also perceived the format and questions to be appropriate and effective. However, the results also revealed some potential issues with the clarity of the rubric and the lack of training for test administration and scoring. These results highlighted the importance of norming and calibration in scoring for the speaking test and entailed practical implications for university-level ESL placement tests.

**Keywords:** consequences of testing, ESL speaking test, mixed-methods research, placement test, validity

## 1. Introduction

Each year, thousands of non-native English speakers travel to English-speaking countries to study English as a second language (ESL) at universities (Wall, Clapham, & Alderson, 1994). To place these students into appropriate levels of ESL classes, ESL programs either use standard English language proficiency tests, such as the Test of English as a Foreign Language (TOEFL) or International English Language Testing System (IELTS) (Kokhan, 2012), develop their own placement assessments, or use a combination of both. Placement is an important component because of the need to group students and provide instruction appropriate for their language proficiency (Wall et al., 1994). Results from the placement tests may also be used to make decisions about students' admission to universities and financial aid applications. Despite the stakes associated with placement tests, research on the validity of these tests are relatively limited (Roever & McNamara, 2006).

In particular, validation research on speaking placement test receives even less attention. In this study, we adapted Messick's integrative view of validity (1996) and Kane's interpretation argument framework (2013). Validity is the degree to which the inferences drawn are valid. Therefore, validation research should include empirical evidence for or against the justification of inferences made from scores. Although university-level ESL placement tests generally evaluate four language skills, i.e., listening, speaking, reading, and writing, not all ESL placement tests have a speaking component. Despite high demand for communication skills in speaking and writing, the focus of the majority of the ESL services in tertiary education is mainly on academic writing (Ransom, Larcombe, & Baik, 2005). It is also more labor-intensive to administer and score speaking tests than other tests. However, research has shown that the lack of a speaking subtest could have undesirable consequences on the reliability and validity of the placement decisions for listening and speaking classes.

To fill this gap in the literature, the current study investigated the validity of a university-level speaking placement test using a quantitative-dominant mixed-methods (MM) design. As stated by the *Standards for*

*Educational and Psychological Testing* (AERA, APA, & NCME, 2014), the five aspects of validity evidence are (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) consequences of testing, and the current study specifically focused on *relations to other variables* and *consequences of testing*. Results from the present study contribute to the research on consequential validity from stakeholders' perspectives. Since many placement tests are developed and administered by teachers in the ESL programs, understanding teacher examiners' perspectives is critical to improving the reliability and validity of the tests. The use of a MM approach also contributes to the diversity of research methodology in language testing as it has been dominated by quantitative and psychometric research methods (Jang, Wagner, & Park, 2014; Lee & Greene, 2007). Finally, the study can serve as an example for university ESL programs that are interested in conducting an in-house validation study for accreditation or program development and improvement purposes.

### *1.1 The Validity of University-Level ESL Speaking Placement Tests*

Not all ESL placement tests include a speaking subtest because speaking classes are not always offered in university language programs. There are also practical concerns over the labor-intensive process for administering and scoring the speaking test. The lack of a speaking subtest could compromise the accuracy (Note 1) and the reliability of the placement results for speaking/listening classes. Johnson (2012) investigated the validity of a university-level placement test that includes a standardized multiple-choice test (in English and Math) and a writing test. Results of the multiple-choice and the writing tests were used for placement decisions for all ESL classes, including speaking/listening classes, despite the absence of a speaking/listening component in the placement test. The author found that more than two-thirds of the students in the speaking/listening class reported that the class was too easy. Based on instructors' end-of-semester evaluations of students' proficiency levels, approximately half of the students in the speaking/listening classes were misplaced.

Validation studies of university-level ESL placement tests often focus on evaluating the entire placement test (Johnson, 2012; Li, 2015) rather than validating a subtest such as speaking (Jamieson, Wang, & Church, 2013) or writing (e.g., Johnson & Riazi, 2017). University-level placement tests usually either use a commercially developed speaking test like the TOEFL (Kokhan, 2013) or develop their own speaking test locally (Jamieson et al., 2013). Research has shown that each test has its pros and cons. For example, Jamieson and colleagues (2013) examined the concurrent validity of an in-house speaking test developed for ESL placement purposes. They compared the curriculum coverage, statistical distributions, and practicality between the in-house speaking test and Versant, a commercial speaking test. The results revealed that the in-house test discriminated better among students of mid-level proficiency and Versant discriminated better between low and high proficiency levels. However, the in-house test is much more affordable, and thus more practical, than the commercial test.

### *1.2 The Role of Consequential Validity in Test Validation*

Test validity is generally considered the most important quality of a test in language assessments. Validity refers to "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (Messick, 1989, pp. 29–30). The current study adapted Messick's integrative view of validity (1996) and Kane's interpretation argument framework (2013) and focused on two aspects of validity evidence: relations to other variables and consequences of testing. Testing consequences, also known as consequential validity, relates to the implications of the inferences made from test results and the societal consequences of test use (Chapelle, 1999). Researchers have argued against including consequential validity in validation frameworks because of its inherent subjectivity and because it imposes a great burden on test developers to consider all the possible consequences of the test uses (Borsboom, Mellenbergh, & van Heerden, 2004). However, Messick (1989), as well as other researchers, have claimed that consequential validity should be an integral part of test validation (Brown & Abeywickrama, 2010; Kane, 2002; McNamara & Roever, 2006; Winke, 2011). Brown and Abeywickrama (2010) posit that consequential validity is a multi-facet construct that entails the effect of tests on test takers and stakeholders' perceived validity of the test.

Relatively little research has been devoted to investigating consequential validity, and most consequential validity studies have focused on K-12 high-stakes tests (e.g., Brewer, Knoeppel, & Lindle, 2015) and/or large-scale standardized assessment such as TOEFL (e.g., Fox & Cheng, 2016). Existing studies have mainly focused on examinees' perceptions of the test (Cheng & Deluca, 2011). Other non-testing-professional users' opinions, such as teacher users' perceptions of the tests, have been undervalued in test validation research (East, 2015; Winke, 2011). In the case of local university-level ESL placement tests, teachers usually serve an important role in the development and/or administration of placement tests. They may also serve as raters for productive tests, such as scoring students' essays and speech production. We argue that teachers' perceptions of the placement test's validity can have an effect on their role as examiners and should be included in test validation processes (Winke, 2011). Because they are also directly affected by the placement test results as class instructors, an investigation of their perceptions can also complement technical evidence for the validity of the test (Chapelle, 1999; Kane, 2002).

### *1.3 Mixed-Methods Approach in Language Assessment Validation*

Test validity consists of many interrelated qualities, and thus requires diverse types of data to understand the construct (Creswell, 2013). Some test validation studies have used the MM design to investigate the validity of ESL placement tests (Johnson & Riazi, 2017; Lee & Greene, 2007). Lee and Greene (2007) used a MM design to examine the predictive validity of a university-level ESL placement test. The authors collected 100 graduate students' ESL placement test scores and three measures of their academic performance. They also administered surveys on and interviews with graduate students and faculty members at the university. Although the quantitative analyses did not yield any significant correlations between students' ESL test scores and their academic performance indicators, qualitative results showed that English skills contributed to students' academic performance. Additional MM analyses also helped unveil the complex relationships between students' English skills and their academic performance. MM has also been used to examine human raters' decision-making process in speaking and writing language assessments (Barkaoui, 2010). For example, Barkaoui (2010) used a MM design to investigate whether raters differ, as a function of their levels of ESL teaching and rating experience, in the holistic and analytical scores they assigned to ESL essays as well as their rating process. Experienced raters in the study were stricter than novice raters, and they also valued linguistic accuracy more than novice raters.

### *1.4 The Current Study*

The three research questions are listed below by validity dimension:

1) To what extent do student examinees' scores on the speaking placement test correlate with their self-assessment of English language proficiency? (relations to other variables)

2) To what extent do student examinees' scores on the ESL speaking placement test correlate with their instructors' evaluations of their English language proficiency at the end of the semester? (relations to other variables)

3) How do student examinees' and teacher examiners' responses to the surveys and interview questions contribute to a more comprehensive understanding of the validity of the speaking test, via mixed-methods analysis? (consequences of testing)

The study was conducted at an Academic English (AE) program at a university in the United States. The AE program is an academic unit at the university, and it offers English language courses in four different domains, i.e., Oral Communication, Reading/Vocabulary, Writing/Grammar, and TOEFL Preparation. The program uses three placement tests to place students to the appropriate level: a standardized language proficiency test that consists of listening, grammar, vocabulary, and reading sections, and a speaking and a writing test that were developed by the staff and teachers in the AE program. The placement tests are administered by the teachers at the beginning of each semester and summer sessions.

The current study focused on the speaking test. The speaking test was a face-to-face interactive interview between a student examinee and a teacher examiner. A second teacher examiner observed the interview silently and rated the interview based on a holistic rubric (see Appendix A). The two examiners rated students individually using a six-level holistic rubric and then discussed their scores after the interview to reach a consensus. After a brief introduction and greeting, the examiner delivered the prompts verbally. The interviews lasted between 5−10 minutes. The prompts of the speaking test comprised of a series of everyday questions such as "what do you like to do in your free time?" followed by academically-oriented questions such as "how can the English language help you advance in your career?" All teachers received general training about the placement tests at the beginning of each semester, but the training did not include practice scoring and calibration.

All student examinees completed the three placement tests on the same day. Based on the placement test results, they were placed into one of five levels of ESL classes in the four domains. The speaking test results were used to place students in the appropriate level of oral communication classes. Student examinees could challenge their placement results and petition to switch to a different level within the first two weeks of class. According to the program, the petition rate in the past three years was approximately 1% for the oral communication classes, suggesting high accuracy of the speaking placement test.

## 2. Method

### *2.1 Participants*

Participants included eight teacher examiners and 41 student examinees. The eight teacher examiners administrated and scored the speaking placement tests in an academic year. They were all females with a mean age of 40 (*SD* = 11). All teacher examiners held a master's degree and had a minimum of 2 years of teaching experience (*Range* = 2 to 24). At the time of participation, half of them had taught at the AE program for approximately 2 years. The rest had been there longer than 4 years, with one experienced teacher who had worked for the AE program for 24 years.

The 41 student examinees took the placement tests in either Fall (*n* = 26) or Spring (*n* = 18). Approximately 60% of the student examinees were males (*n* = 24). There was a large diversity in their native language backgrounds, which consisted of Spanish (27%), Arabic (24%), Japanese (12%), Chinese (10%), Vietnamese (10%), and Turkish (7%), as well as Farsi, French, and Portuguese. Because the same placement tests were administered in the same way, and data collection procedure was also standardized in both semesters, we combined data from the two cohorts to increase statistical power.

### *2.2 Instruments*

#### 2.2.1 Teacher Individual Interview

We conducted a semi-structured interview with five out of the eight teacher examiners who administered the tests. Five interviews were conducted in Fall and two in Spring. The interviewer asked the teacher participants six open-ended questions about their perceptions of the effectiveness of the rubric and their suggestions for the placement tests. The individual interview sessions were all audio-recorded and fully transcribed for analysis.

#### 2.2.2 Teacher Perception Survey

The teacher perception survey was adapted from prior research (e.g., Winke, 2011). We created the main constructs first and developed questions for each construct. We asked two experienced ESL teachers and program directors to review earlier drafts of the survey. The final survey included a 1−5 Likert scale (1 = Strongly Disagree; 5 = Strongly Agree) type of questions about the appropriateness and effectiveness of the placement tests, the placement results, and the use of the holistic rubric. The survey also included two open-ended questions: 1) If you could change the placement tests, including the rubric, what changes would you make? 2) Do you have any additional comments about the placement tests?

#### 2.2.3 Teacher's Ratings of Students' English Proficiency

At the end of the fall and spring semester, teachers were asked to evaluate their students' English proficiency in their respective language classes from a scale point of 1 (poor) to 7 ("like a native speaker" for listening and speaking and "excellent" for reading and writing"). Teachers were asked to evaluate domains of students' language skills that were relevant to the course they taught. For example, teachers of the oral communication class rated their students' speaking and listening skills.

#### 2.2.4 Student Perception Survey

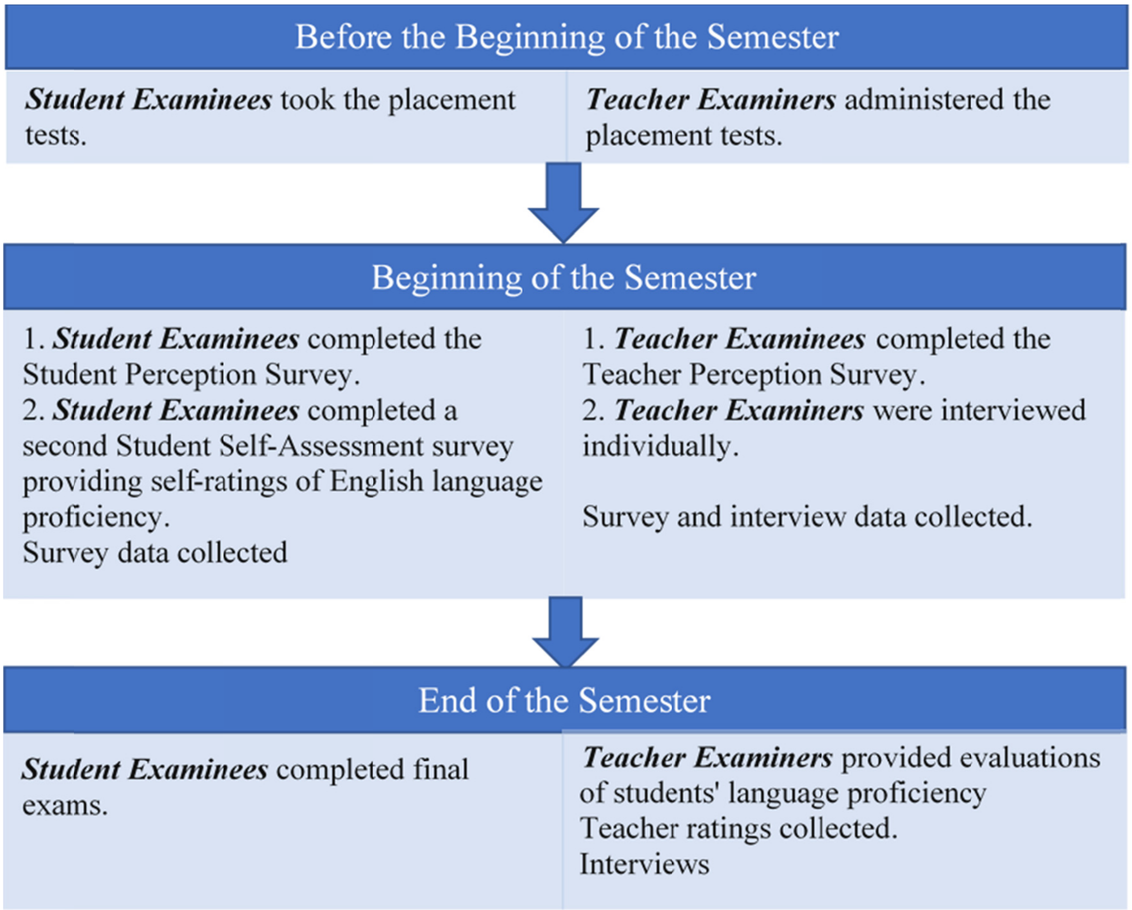
The student survey was designed to parallel the teacher perception survey and included questions about the appropriateness and effectiveness of the placement tests. Student examinees evaluated the content, format, and effectiveness as well as the rubric (where applicable) for each placement test. Students took the perception survey within one week after the placement test.

#### 2.2.5 Student Self-Assessment Survey

A student self-assessment survey was administered each semester within one week after students were placed in a class. To elicit students' evaluation of their English language proficiency, the survey included four self-evaluation items and 16 can-do items. The four self-evaluation items requested students to rate their English proficiency in four domains (listening, speaking, reading, writing) on a 1−7 Likert scale. The 16 can-do statements, consists of four can-do items in each of the four language domains. They were used to measure whether students believed they are able to perform their respective English language skills. The can-do statements examined students' confidence in their English language ability by asking them to rate how much they agree/disagree with the statement on a 1−5 Likert scale.

### *2.3 Data Collection Procedure*

All student and teacher participants completed a perception survey shortly after the placement tests in Fall and in Spring. We also conducted an individual interview with five teacher examiners in Fall and two teacher examiners in Spring. See Figure 1 below for the data collection timeline.



Before the Beginning of the Semester

***Student Examinees*** took the placement tests.

***Teacher Examiners*** administered the placement tests.

Beginning of the Semester

1. ***Student Examinees*** completed the Student Perception Survey.
2. ***Student Examinees*** completed a second Student Self-Assessment survey providing self-ratings of English language proficiency.

Survey data collected

1. ***Teacher Examinees*** completed the Teacher Perception Survey.
2. ***Teacher Examiners*** were interviewed individually.

Survey and interview data collected.

End of the Semester

***Student Examinees*** completed final exams.

***Teacher Examiners*** provided evaluations of students' language proficiency
Teacher ratings collected.
Interviews

Figure 1. Data collection timeline

### *2.4 Data Analysis*

We conducted a series of Spearman rank-order correlations between examinee's speaking placement test scores and their self-ratings of English language proficiency to address the first research question about the relationship between test scores and other assessments, and between student examinees' placement test scores and their instructors' evaluations of their English language proficiency at the end of the semester to answer the second question. We answered the third research question by conducting mixed-methods analysis of the following data: students' and teachers' responses for the perception surveys, teachers' responses to the open-ended survey questions, and semi-structured teacher interviews. Both students' and teachers' responses in the student and teacher perception surveys were analyzed quantitatively and descriptively. Teacher examiners' responses to the open-ended survey questions and interview questions were coded and analyzed qualitatively following Miles and Huberman's inductive approach (1994). One of the researchers in the team conducted descriptive In-vivo coding based on Saldana's (2009) qualitative coding techniques. The In vivo coding method selected a word or a short phrase taken from the participant's response to label a section of the data to represent the participant's perceptions and voice. It was employed in the present study to represent stake-holders' (teachers) perceptions of the placement tests and the test instruments such as rubrics.

## 3. Results

### *3.1 Descriptive Results*

The descriptive results of student examinees' speaking test scores, self-ratings of English proficiency in four domains, and responses to the 16 can-do statements are presented in Table 1. Reliability was high for the self-ratings in the four domains (Cronbach's $\alpha$ = .83) and for the 16 can-do statements (Cronbach's $\alpha$ = .95) (Note 2). The mean for the speaking placement test is 3.19 and there is a good amount of variance in the sample

(*SD* = 1.12; *Range* = 1-5). On average, student examinees perceived their speaking skills to be the weakest in the four domains (*M* = 3.70; *SD* = 1.13). The can-do statement ratings showed that the student examinees were the least confident in making oral presentations about an academic topic than in other non-academic speaking activities (speaking can-do statement No. 1, *M* = 3.56). A similar pattern was observed in their self-evaluations of writing skills; they were less confident in their ability to write academic essays or emails about academic activities than personal letters or non-academic topics (writing can-do statement No. 1, *M* = 3.57). In terms of receptive skills, they reported high confidence in understanding clear speech and both daily conversations and academic discussions (listening can-do statements No. 1, *M* = 3.95; No. 3, *M* = 4.00; No. 4, *M* = 3.91), but they did not appear to comprehend well movies and TV shows (listening can-do statement No. 1, *M* = 3.51) possibly due to the speed/pace of the speech in media. They also reported strong confidence in comprehending text related to daily matters (reading can-do statements No. 2, *M* = 3.92; No. 3, *M* =3.81; No. 4, *M* = 3.97) but less confidence in understanding academic text (reading can-do statement No. 1, *M* = 3.62).

Table 1. Means, standard deviations, and ranges of examinees' speaking test scores and self-evaluation of English language proficiency

| | Min | Max | M | SD |
|---|---|---|---|---|
| Speaking test scores | 1.00 | 5.00 | 3.19 | 1.12 |
| Self-Rating-Speaking | 2.00 | 6.00 | 3.70 | 1.13 |
| Self-Rating-Writing | 1.00 | 7.00 | 3.89 | 1.24 |
| Self-Rating-Listening | 1.00 | 6.00 | 3.84 | 1.28 |
| Self-Rating-Reading | 2.00 | 6.00 | 4.00 | 1.18 |
| Speaking can-do statements | | | | |
| 1. I can make oral presentations about an academic topic in class. | 2.00 | 5.00 | 3.56 | 0.79 |
| 2. I can communicate effectively with teachers and classmates in class. | 1.00 | 5.00 | 3.77 | 0.93 |
| 3. I can talk about topics with simple words and phrases which are familiar to me. | 2.00 | 5.00 | 3.77 | 0.84 |
| 4. I can participate in simple conversations about everyday matters and everyday needs, such as interacting with people for housing, banking, shopping, etc. | 1.00 | 5.00 | 3.67 | 1.04 |
| Writing can-do statements | | | | |
| 1. I can write English academic essays. | 2.00 | 5.00 | 3.57 | 0.73 |
| 2. I can write personal letter, diary in English to express my feelings or describe my day. | 2.00 | 5.00 | 3.76 | 0.93 |
| 3. I can write emails to school personnel about academic activities. | 2.00 | 5.00 | 3.54 | 0.93 |
| 4. I can write simple notes and text messages about everyday matters and every need. | 1.00 | 5.00 | 3.81 | 0.94 |
| Listening can-do statements | | | | |
| 1. I can understand classroom instructions, lectures, presentations, and discussions. | 2.00 | 5.00 | 3.95 | 0.91 |
| 2. I can understand English movies, TV shows, videos, etc. | 1.00 | 5.00 | 3.51 | 0.96 |
| 3. I can understand speech that is slow and clear. | 3.00 | 5.00 | 4.00 | 0.74 |
| 4. I can understand the main points of clear standard speech on familiar matters connected with work, school, leisure, etc. | 2.00 | 5.00 | 3.91 | 0.83 |
| Reading can-do statements | | | | |
| 1. I can understand academic textbooks and essays in English. | 2.00 | 5.00 | 3.62 | 0.76 |
| 2. I can understand simple informational texts and short simple descriptions, especially if they contain pictures that help explain the text. | 2.00 | 5.00 | 3.92 | 0.85 |
| 3. I can understand simple personal letters in which the writer tells or asks me about aspects of everyday life. | 3.00 | 5.00 | 3.82 | 0.77 |
| 4. I can understand everyday signs and notices in public places, such as streets, restaurants, airports, and schools. | 3.00 | 5.00 | 3.97 | 0.75 |

### *3.2 Research Question 1: To What Extent Do Student Examinees' Scores on the Speaking Placement Test Correlate with Their Self-Assessment of English Language Proficiency?*

The speaking placement test scores significantly correlated with students' self-assessment of Listening, Speaking, and Writing skills, but not with Reading skills (*rs* = .28, NS). The strength of correlation ranges from .28 to .56, with the strongest correlation being with Speaking, providing support for the validity of the speaking test (See Table 2).

Table 2. Correlations (*Spearman's rho*) between examinees' speaking test scores and self-evaluation of English language proficiency (n = 41)

| | Speaking Placement | Self-Speak | Self-Listen | Self-Read | Self- Write |
|---|---|---|---|---|---|
| Speaking Placement | -- | .56** | .35* | .28 | .36* |
| Self-Rating – Speak | | -- | .65** | .52** | .56** |
| Self-Rating – Listen | | | -- | .56** | .59** |
| Self-Rating – Read | | | | -- | .76** |
| Self-Rating – Write | | | | | -- |

Students' responses to the 16 can-do statements, however, told a slightly different story (See Table 3). The speaking test scores significantly correlated with two of four can-do statements related to their speaking skills: "I can make oral presentations about an academic topic in class." (*rs* = .37, *p* = .022) and "I can talk about topics with simple words and phrases which are familiar to me." (*rs* =. 40, *p* = .010). However, speaking test scores did not yield significant correlations with student examinees' self-assessments of their communication skills in class ("I can communicate effectively with teachers and classmates in class") or their ability to participate in simple conversations about everyday needs ("I can participate in simple conversations about everyday matters and everyday needs, such as interacting with people for housing, banking, shopping, etc."), suggesting that the speaking placement test may actually orient more toward academic-related domains of proficiency than toward everyday conversation proficiency. Additionally, speaking test scores significantly correlated with several reading and writing can-do statements. Counter to our expectation, however, speaking placement test scores did not yield any significant correlations with the four listening can-do statements.

### *3.3 Research Question 2: To What Extent Do Student Examinees' Scores on the ESL Speaking Placement Test Correlate with Their Instructors' Evaluations of Their English Language Proficiency at the End of the Semester?*

Table 4 presents the descriptive statistics of teachers' average ratings of students' four skills. As shown in Table 5, students' speaking test scores at the beginning of the semester yielded significant correlations with the teachers' end-of-semester ratings of students' four skills. The strength of correlation ranges from .38 to .66. We interpreted the results to provide some support for the validity of the speaking placement test, though the strength of correlation between the speaking test score and teachers' ratings of students' speaking skills is weaker compared to that between the speaking test score and ratings of other skills.

Table 3. Correlations (*Spearman's rho*) between the speaking test and student examinees' self-evaluation of can-do statements by domain (n = 41).

| Speaking & Listening Can-do Statements | SP | S1 | S2 | S3 | S4 | L1 | L2 | L3 | L4 | R1 | R2 | R3 | R4 | W1 | W2 | W3 | W4 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S1. I can make oral presentations about an academic topic in class. | .37* | -- | | | | | | | | | | | | | | | |
| S2. I can communicate effectively with teachers and classmates in class. | .29 | .68** | -- | | | | | | | | | | | | | | |
| S3. I can talk about topics with simple words and phrases which are familiar to me. | .41* | .58** | .64** | -- | | | | | | | | | | | | | |
| S4. I can participate in simple conversations about everyday matters and everyday needs, such as interacting with people for housing, banking, shopping, etc. | .28 | .76** | .49** | .58** | -- | | | | | | | | | | | | |
| L1. I can understand classroom instructions, lectures, presentations, and discussions. | .10 | .36$^{*}$ | .45$^{**}$ | .51$^{**}$ | .37$^{*}$ | -- | | | | | | | | | | | |
| L2. I can understand English movies, TV shows, videos, etc. | .15 | .55$^{**}$ | .32 | .29 | .51$^{**}$ | .62$^{**}$ | -- | | | | | | | | | | |
| L3. I can understand speech that is slow and clear. | .16 | .34$^{*}$ | .40$^{*}$ | .39$^{*}$ | .34$^{*}$ | .66$^{**}$ | .44$^{**}$ | -- | | | | | | | | | |
| L4. I can understand the main points of clear standard speech on familiar matters connected with work, school, leisure, etc. | .18 | .65$^{**}$ | .46$^{**}$ | .56$^{**}$ | .60$^{**}$ | .74$^{**}$ | .74$^{**}$ | .63$^{**}$ | -- | | | | | | | | |
| Reading & Writing Can-do Statements | SP | S1 | S2 | S3 | S4 | L1 | L2 | L3 | L4 | R1 | R2 | R3 | R4 | W1 | W2 | W3 | W4 |
| R1. I can understand academic textbooks and essays in English. | .18 | .44$^{**}$ | .38$^{*}$ | .31 | .42$^{*}$ | .41$^{*}$ | .47$^{**}$ | 0.28 | .42$^{*}$ | -- | | | | | | | |
| R2. I can understand simple informational texts and short simple descriptions, especially if they contain pictures that help explain the text. | .41$^{*}$ | .53$^{**}$ | .51$^{**}$ | .63$^{**}$ | .58$^{**}$ | .69$^{**}$ | .56$^{**}$ | .67$^{**}$ | .71$^{**}$ | .51$^{**}$ | -- | | | | | | |
| R3. I can understand simple personal letters in which the writer tells or asks me about aspects of everyday life. | .46$^{**}$ | .50$^{**}$ | .41$^{*}$ | .58$^{**}$ | .53$^{**}$ | .50$^{**}$ | .52$^{**}$ | .54$^{**}$ | .66$^{**}$ | .41$^{*}$ | .68$^{**}$ | -- | | | | | |
| R4. I can understand everyday signs and notices in public places, such as streets, restaurants, airports, and in schools. | .16 | .53$^{**}$ | .45$^{**}$ | .51$^{**}$ | .44$^{**}$ | .66$^{**}$ | .52$^{**}$ | .63$^{**}$ | .85$^{**}$ | .43$^{**}$ | .72$^{**}$ | .64$^{**}$ | -- | | | | |
| W1. I can write English academic essays. | .11 | .28 | .37$^{*}$ | .33$^{*}$ | .31 | .32 | .33$^{*}$ | .36$^{*}$ | .36$^{*}$ | .46$^{**}$ | .47$^{**}$ | .40$^{*}$ | .41$^{*}$ | -- | | | |
| W2. I can write a personal letter, diary in English to express my feelings or describe my day. | .30 | .52$^{**}$ | .43$^{**}$ | .49$^{**}$ | .58$^{**}$ | .51$^{**}$ | .40$^{*}$ | .52$^{**}$ | .66$^{**}$ | .38$^{*}$ | .55$^{**}$ | .65$^{**}$ | .62$^{**}$ | .40$^{*}$ | -- | | |
| W3. I can write emails to school personnel about academic activities. | .55$^{**}$ | .70$^{**}$ | .67$^{**}$ | .67$^{**}$ | .56$^{**}$ | .46$^{**}$ | .53$^{**}$ | .44$^{**}$ | .57$^{**}$ | .46$^{**}$ | .63$^{**}$ | .70$^{**}$ | .49$^{**}$ | .46$^{**}$ | .56$^{**}$ | -- | |
| W4. I can write simple notes and text messages about everyday matters and every need. | .42$^{**}$ | .66$^{**}$ | .55$^{**}$ | .78$^{**}$ | .69$^{**}$ | .53$^{**}$ | .51$^{**}$ | .53$^{**}$ | .71$^{**}$ | .33$^{*}$ | .69$^{**}$ | .76$^{**}$ | .58$^{**}$ | .49$^{**}$ | .67$^{**}$ | .79$^{**}$ | -- |

*Note*. SP = Speaking Placement.

Table 4. Means, standard deviations, and ranges of instructors' ratings of student examinees' English language proficiency at the end of the semester

| | Min | Max | M | SD |
|---|---|---|---|---|
| Teacher Rating: Listen | 2.00 | 7.00 | 5.44 | 1.42 |
| Teacher Rating: Speak | 1.00 | 7.00 | 5.50 | 1.50 |
| Teacher Rating: Read | 1.00 | 7.00 | 5.44 | 1.41 |
| Teacher Rating: Write | 3.00 | 7.00 | 5.19 | 1.10 |

*Note.* The rating scale is a 1–7 Likert scale (1 = poor; 7 = like an educated native speaker).

Table 5. Correlation matrix (*Spearman's rho*) between speaking placement test scores and end-of-semester teacher ratings of students' language proficiency

| | Speaking placement test | Teacher rating-Listen | Teacher rating-Speak | Teacher rating-Read |
|---|---|---|---|---|
| Teacher rating- Listen | .50** | -- | | |
| Teacher rating- Speak | .38* | .82** | -- | |
| Teacher rating- Read | .66** | .68** | .49* | -- |
| Teacher rating- Write | .46** | .23 | .35 | .75** |

*Note.* The rating scale is a 1–7 Likert scale (1 = poor; 7 = like an educated native speaker).

### *3.4 Research Question 3: How Do Student Examinees' and Teacher Examiners' Responses to the Surveys and Interview Questions Contribute to a More Comprehensive Understanding of the Validity of the Speaking Test, via Mixed-Methods Analysis? (Consequences of Testing)*

Generally speaking, student examinees tended to report that the speaking test's format and questions were appropriate and effective (See Table 6). For the most part, they enjoyed taking the test and believed that the placement decision was accurate. They expressed some confidence in the construct validity of the test. However, to our surprise, not everyone saw the benefits of previewing the assessment rubric prior to the speaking test ($M$ = 3.14; $SD$ = 1.13). Only 15 out of the 41 students (circa 36%) agreed with the statement that "I think it would be better if I could see the rubric before the interview."

Table 6. Student examinees' responses to perception survey items (n = 41)

| | 1 | 2 | 3 | 4 | 5 | M (SD) | Range |
|---|---|---|---|---|---|---|---|
| 1. The questions during the interview (i.e., oral proficiency test) are appropriate and effective. | 0 | 2 | 14 | 15 | 10 | 3.80 (.87) | 2–5 |
| 2. The face-to-face interview format of the oral proficiency test is appropriate and effective. | 1 | 1 | 8 | 16 | 15 | 4.04 (.94) | 1–5 |
| 3. I think it would be better if I could see the rubric before the interview. | 4 | 6 | 16 | 10 | 5 | 3.14 (1.13) | 1–5 |
| 4. I enjoyed talking to the examiner/interviewer during the oral proficiency test. | 1 | 2 | 10 | 7 | 19 | 4.05 (1.09) | 1–5 |
| 5. There was no real connection between the oral proficiency test and my oral language proficiency. | 9 | 12 | 14 | 2 | 4 | 2.51 (1.18) | 1–5 |
| 6. The oral proficiency test worked really well placing me in the appropriate class.[a] | 0 | 3 | 8 | 15 | 14 | 4.00 (.93) | 2–5 |

*Note*. Strongly disagree = 1. Strongly agree = 5. [a] Missing data = 1; Unlike the other items in the survey, question 5 is a reverse item and the results should be interpreted carefully.

On the other hand, teacher examiners also tended to report positive opinions about the speaking placement test (See Table 7 for the descriptive results). To a large degree, they perceived the content and format of the tests to be appropriate and effective. They also tended to believe that the placement test results correlated with student examinees' ability to meet the university's English language requirements and that the test worked well in placing examinees in appropriate levels. However, they appeared to have some concerns about the rubric and/or the rubric use when determining examinees' language proficiency; half of the teacher examiners disagreed with the statement that "the rubric for the oral proficiency test is appropriate and effective."

Table 7. Teacher examiners' responses to teacher perception survey items (n = 8)

| | 1 | 2 | 3 | 4 | 5 | M (SD) | Range |
|---|---|---|---|---|---|---|---|
| 1. The questions of the oral proficiency test are appropriate and effective | 0 | 1 | 0 | 2 | 5 | 4.38 (1.06) | 2–5 |
| 2. I found it difficult to place students into the appropriate levels based on the speaking rubric. | 3 | 1 | 1 | 1 | 2 | 2.75 (1.75) | 1–5 |
| 3. The rubric for the oral proficiency test is appropriate and effective. | 1 | 3 | 0 | 3 | 1 | 3 (1.14) | 1–5 |
| 4. The format of the oral proficiency test (i.e., an interview) is appropriate and effective. | 0 | 0 | 0 | 3 | 5 | 4.63 (0.52) | 4–5 |
| 5. There was no real connection between the oral proficiency test and students' ability to meet the university's English language requirements. | 6 | 0 | 1 | 1 | 0 | 1.63 (1.19) | 1–5 |
| 6. The oral proficiency test worked really well placing students in appropriate classes. | 0 | 1 | 0 | 5 | 2 | 4.00 (0.92) | 2–5 |

*Note*. Strongly disagree = 1. Strongly agree = 5.

Results from their responses to the open-ended survey questions and individual interviews corroborated the quantitative findings of teachers' concerns with the rubric. Four out of the eight teacher examiners made suggestions to revise the rubric for clarity and simplicity and to match the institution course levels. The two excerpts below illustrate their concerns. In Excerpt 1, instructor 6 saw the value of using a rubric as documentation of why a score was assigned based on students' proficiency level. This required the use of an analytic rubric or diagnostic rubric. The current holistic rubric, on the other hand, did not support this function. This may imply the need for professional training in teachers' assessment literacy and teachers' participation in the rubric development process.

Excerpt 1 (open-ended questions)

Instructor 6: "For the oral presentation, I would suggest a way to circle or check the bullet marks as to why you placed them at that level that can be used as support and documentation."

Other rubric-associated potential problems were also revealed, including lack of standardization in administration, insufficient training for the teacher examiners to use the rubric, and the disconnection between the test rubric standards and the institution course levels. Teacher examiners reported during the individual interviews that the proctoring process should be standardized and to ensure the consistency in the administration of the test. In Excerpts 3, Instructor 4 shared the concern that oral interviews during the speaking test should be conducted in the same fashion. While there was no objective observation of the lack of reliability in the exam administration, stake-holders' voice should be valued, and training should be provided to ensure the reliability of the test.

Excerpt 3 (open-ended questions)

Instructor 4: "In my opinion, instructors should all be executing oral proficiency exam the same way with in the same time frame. There should be strict guidelines that indicate how the exam should be proctored and followed by the instructors who are doing this exam."

While the quantitative results showed that teacher examiners tended to perceive the speaking test as effective in placing examinees in appropriate levels, the interview data (e.g. Excerpt 4) revealed that some teachers may rely more on their own experience and intuition than using the rubric and the guidelines to determine student examinees' proficiency. One potential explanation for teachers relying on past teaching experience rather than the rubric was that the proficiency levels may have shifted over time due to uneven student enrollment across levels. As a small language program, sometimes a certain minimal number of students is required for a class to make. This means that sometimes the beginner level students may be forced into the low intermediate level due to lack of low enrollment of the students in the beginner level. The average proficiency level of the students in a class, therefore, may not always represent the level of proficiency in the rubric standards. As a result, the actual proficiency in each level of class, as was stated by Instructor 5, was "not completely clear to all" instructors, due to the fact that the proficiency levels of the classes "have changed over time" based on student enrollment (Excerpt 5). Therefore, without an updated rubric and norming sessions each semester, the rubric standards can be confusing to the raters. Instructors may thus rely on their experience with students in previous classes instead of the rubric standards when rating the students.

Excerpt 4 (interview)

Instructor 2: "… (to determine student level) sometimes I go back to the **students that I had**, I taught level one and presentations and compare. ***It is intuition***."

Excerpt 5 (open-ended questions)

Instructor 5: "Levels have **not been completely to clear** with all of us… Levels have **changed** over time.

## 4. Discussion

The current study set out to investigate the validity of a university-level ESL speaking placement test via a MM approach. For our first research question, results showed significant associations between students' test scores and self-assessments of overall proficiency, and the strength of correlation between the test scores and self-ratings of speaking skills is stronger than that between the test scores and self-ratings of other skills. It is worth noting that, despite the significant correlations between test scores and a few of the can-do statements about reading activities, there is no significant correlation between test scores and examinees' self-assessment of overall reading proficiency. The discrepancy in correlation patterns appears to suggest that examinees' interpretations of "overall reading proficiency" may be different from their reading ability for specific activities, such as reading academic texts. Results from validation studies of self-assessments have been mixed. While some studies showed encouraging results of the validity of inferences from self-assessments (Ross, 1998), other studies challenge their validity (Suzuki, 2015). Although we found high reliability in the self-assessment measures of proficiency, more research is needed to examine the validity of self-assessments. Future research should also use standardized and objective measures, such as TOEFL or IELTS speaking test scores as other variables.

The correlation analyses between the scores and the can-do statements revealed a more nuanced picture; there were only significant associations between the scores and academically-oriented activities but not between the scores and non-academic activities, such as participating in conversations about daily needs. We interpreted the results to be positive validity evidence for the speaking test as the construct targets academically-oriented speaking proficiency than interpersonal communication skills. Counter to our expectation, there were no significant correlations between speaking test scores and the listening can-do statements. Although the speaking test is an interactive task that involves listening skills, the teacher examiner provides extensive prompting if the student examinee does not respond (Nakatsuhara, 2018). The extensive prompts may have prevented students from active listening. More research is needed to understand the relationship between student examinees' listening skills and their performance in interactive speaking tasks.

For the second research question about the placement test scores' relation with end-of-semester English language proficiency, the bivariate correlational analyses showed significant correlations between test scores and instructors' ratings of examinees' English language proficiency in all four domains. Interestingly, the strength of the correlation is the highest between the speaking placement test scores and teacher ratings of reading proficiency ($r$s = .66) instead of speaking proficiency ($r$s = .38). The counter-intuitive results may be attributed to differences between the target construct of the speaking placement test and the curricular objectives of the oral communication class.

Turning now to the third research question about consequential validity, the results revealed that, to a large degree, both the examinees and examiners perceived the format and questions to be appropriate and effective. For the most part, the examinees also believed in the accuracy of the placement decision was accurate. However, only about one-third of them expressed interest in seeing the scoring rubric prior to the speaking placement test. While we did not interview the examinees directly to understand their rationale, we believe that it may be attributed to the influence of the testing practices in their native countries, where test-taking procedure is highly structured and authoritative. Examinees may be ready to accept the AE program's testing practice and thus did not consider the benefits of previewing the scoring rubric. Alternatively, examinees' perceived stakes for the speaking test may be low, and they thus do not see the need to preview the rubric.

Although most teacher examiners perceived the test to be appropriate and effective and reported that the test worked well in placing examinees in levels aligned with their proficiency, they were divided in their opinions about the quality of the rubric. We triangulated their survey responses with their responses to the open-ended questions and individual interviews to cross-check data in the search for regularities. While their open-ended responses corroborated their concerns with the appropriateness and clarity of the rubric, they revealed other concerns with the test, specifically the disconnection between the test and course levels, lack of standardization across administration, and insufficient training for teacher examiners in test administration. Possibly due to their reservations about the rubric, teacher examiners appeared to rely more on their personal teaching experience to evaluate and determine examinees' speaking proficiency. The reliance on personal experience rather than using the rubric compromises the reliability and validity of the scoring process.

A number of researchers have argued that teachers should be involved in assessment (Bachman & Palmer, 1996;

Hamp-Lyons, 1997). Because many ESL service programs develop their placement tests locally, teachers are likely to be involved in test development and administration. However, as shown in the current study, teacher examiners, like any other human examiner, need training in test administration and the use of rubrics. Without systematic training and calibration, teachers may be subject to their own preferences and biases when assessing examinees' language proficiency (Huang, 2013; Isaacs & Thomson, 2013) or academic achievement (Riley, 2015). Teachers in the current study only received general training on the placement test at the beginning of the semester. They did not have an opportunity to practice scoring using the rubric and calibrate their scores. The training they received may not be sufficient to achieve high intra- and inter-rater reliability. Research on the rater effect has shown that human raters tend to draw on their past experience with the second language (L2) speech when evaluating L2 learners' speaking proficiency, leading to a potential bias (Isaacs & Thomson, 2013). Specifically, some studies have found that raters' familiarity with the speakers' non-native accents and/or raters' ESL teaching experience may have an (objective and/or self-reported) effect on their ratings of second language speech (Carey, Mannell, & Dunn, 2011). Furthermore, raters' interpretation and use of the scoring rubric, which may serve as "de facto test constructs" (McNamara, Hill, & May, 2002, p. 229), could also vary as a function of their demographic backgrounds and experiences. As shown in the current study, teacher examiners appeared to rely on their personal experience, rather than applying the scoring rubric in the evaluation of speaking proficiency. To ensure the quality of the placement tests and accuracy of the results, ESL services/programs should provide regular training and calibration in scoring for teacher examiners. Future investigations of teacher examiners' rating processes and application of the scoring rubrics (or lack thereof) when assigning ratings would also provide a better understanding of the validity of the speaking test.

We further argue that teacher examiners would also benefit from professional developments that develop their language assessment literacy (Scarino, 2013) as well as their identity as an assessor rather than as a teacher during the evaluation (Looney, Cumming, van Der Kleij, & Harris, 2018). We define language assessment literacy as the knowledge of measurement and assessment and the application of this knowledge to classroom assessment practices and language assessments. Popham (2011) pointed out that many teachers have limited knowledge of the fundamental concepts in assessment, such as reliability and validity. Given the prominent role teachers play in ESL placement tests as well as in classroom/formative assessments, increasing teachers' language assessment literacy will improve the quality of the assessment that teachers are involved in. Furthermore, as observed in the current study, teacher examiners appeared to rely on their intuition and teaching experience rather than the scoring rubric in evaluating examinees' speaking proficiency. Developing teachers' identity as an assessor when they assume the role of an examiner would also help ensure the reliability and accuracy of their ratings.

Limitations of the study must be acknowledged and addressed in future studies. Data for the study were collected from one ESL program and thus limited in sample size, statistical power, and generalizability. The correlational test results should thus be interpreted conservatively and await replication from future research with a larger sample size. The results for consequential validity could also have been strengthened by interviewing the student examinees. Furthermore, we did not examine content validity and construct validity. Future research also needs to explore the relationships between construct-related learning experiences, such as student examinees' prior instruction and length of residence in the United States and their speaking test scores.

To conclude, results from the study provided some validity evidence for a locally-developed ESL speaking test. Although the results were overall positive, the analyses of teacher examiners' perceptions revealed some potential issues with the clarity of the rubric and the lack of standardization in and training for test administration. Instead of applying the rubric, teacher examiners drew on their own teaching experience in evaluating the examinees, thus compromising the quality of the speaking test. These results entailed practical implications for university-level ESL placement tests as many placement tests were also developed and administered locally by teachers and administrators in the program. The results highlighted the importance of norming and calibration in scoring for the speaking test. Teachers can benefit from targeted training on the test administration protocol, the use of the rubric, and practice scoring.

Continuous professional developments to promote teachers' assessment literacy and develop their professional identity as assessors will also in turn help improve the reliability and validity of the speaking placement test. Given the increasing program accreditation requirements and the need for ESL program development, the study also serves as an example for programs that are interested in validating their assessments.

## References

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

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly*, *7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review*, *111*(4), 1061–1071. https://doi.org/10.1037/0033-295X.111.4.1061

Brewer, C., Knoeppel, R. C., & Lindle, J. C. (2015). Consequential validity of accountability policy: Public understanding of assessments. *Educational Policy*, *29*(5), 711–745. https://doi.org/10.1177/0895904813518099

Brown, H. D., & Abeywickrama, P. (2010). Principles of language assessment. In *Language Assessment: Principles and classroom practices* (2nd ed.). White Plains, NY: Pearson Education.

Carey, M. D., Mannell, R. H., & Dunn, P. K. (2011). Does a rater's familiarity with a candidate's pronunciation affect the rating in oral proficiency interviews? *Language Testing*, *28*(2), 201–219. https://doi.org/10.1177/0265532210393704

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics*, *19*, 254–272. https://doi.org/10.1017/S0267190599190135

Cheng, L., & DeLuca, C. (2011). Voices from test-takers: Further evidence for language assessment validation and use. *Educational Assessment*, *16*(2), 104–122. https://doi.org/10.1080/10627197.2011.584042

Creswell, J. W. (2013). *Research design: Qualitative, quantitative, and mixed-methods*. Thousand Oaks, CA: Sage.

East, M. (2015). Coming to terms with innovative high-stakes assessment practice: Teachers' viewpoints on assessment reform. *Language Testing*, *32*(1), 101–120. https://doi.org/10.1177/0265532214544393

Huang, B. H. (2013). The effects of accent familiarity and language teaching experience on raters' judgments of non-native speech. *System*, *41*(3), 770–785. https://doi.org/10.1016/j.system.2013.07.009

Isaacs, T., & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. *Language Assessment Quarterly*, *10*(2), 135–159. https://doi.org/10.1080/15434303.2013.769545

Jamieson, J., Wang, L., & Church, J. (2013). In-house or commercial speaking tests: Evaluating strengths for EAP placement. *Journal of English for Academic Purposes*, *12*(4), 288–298. https://doi.org/10.1016/j.jeap.2013.09.003

Jang, E. E., Wagner, M., & Park, G. (2014). Mixed-methods research in language testing and assessment. *Annual Review of Applied Linguistics*, *34*, 123–153. https://doi.org/10.1017/S0267190514000063

Johnson, R. C. (2012). *Assessing the assessments: Using an argument-based validity framework to assess the validity and use of an English placement system in a foreign language context*. Unpublished doctoral dissertation. Macquarie University, Sydney, Australia.

Johnson, R. C., & Riazi, A. M. (2017). Validation of a locally created and rated writing test used for placement in a higher education EFL program. *Assessing Writing*, *32*, 85–104. https://doi.org/10.1016/j.asw.2016.09.002

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, *21*(1), 31–40. https://doi.org/10.1111/j.1745-3992.2002.tb00083.x

Kokhan, K. (2012). Investigating the possibility of using TOEFL scores for university ESL decision-making: Placement trends and effect of time lag. *Language Testing*, *29*(2), 291–308. https://doi.org/10.1177/0265532211429403

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing*, *30*(4), 467–489. https://doi.org/10.1177/0265532213475782

Lee, Y.-J., & Greene, J. (2007). The Predictive Validity of an ESL Placement Test: A Mixed-methods Approach. *Journal of Mixed-Methods Research*, *1*(4), 366–389. https://doi.org/10.1177/1558689807306148

Li, Z. (2015). *An argument-based validation study of the English Placement Test (EPT)–Focusing on the* *inferences of extrapolation and ramification*. Unpublished Ph.D. dissertation, Iowa State University.

Looney, A., Cumming, J., van Der Kleij, F., & Harris, K. (2018). Reconceptualising the role of teachers as assessors: teacher assessment identity. *Assessment in Education: Principles, Policy & Practice*, *25*(5), 442–467. https://doi.org/10.1080/0969594X.2016.1268090

McNamara, T., Hill, K., & May, L. (2002). Discourse and assessment. *Annual Review of Applied Linguistics*, *22*, 221. https://doi.org/10.1017/S0267190502000120

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1996). Validity and washback in language testing. *Language Testing*, *13*(3), 241–256. https://doi.org/10.1177/026553229601300302

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis: An expanded source book*. Thousand Oaks, CA: Sage.

Nakatsuhara, F. (2018). Investigating examiner interventions in relation to the listening demands they make on candidates in oral interview tests. In G. J. Ockey & E. Wagner (Eds.), *Assessing L2 listening: Moving towards authenticity* (pp. 205–225). Amsterdam, the Netherlands: John Benjamins. https://doi.org/10.1075/lllt.50.14nak

Nevo, B. (1985). Face validity revisited. *Journal of Educational Measurement*, *22*(4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265–273. https://doi.org/10.1080/08878730.2011.605048

Ransom, L., Larcombe, W., & Baik, C. (2005). *English language needs and support: International-ESL students' perceptions and expectations*. Language and Learning Skills Unit, The University of Melbourne, Australia.

Riley, T. (2015). "I Know I'm Generalizing but…": How Teachers' Perceptions Influence ESL Learner Placement. *TESOL Quarterly*, *49*(4), 659–680. https://doi.org/10.1002/tesq.191

Roever, C., & McNamara, T. (2006). Language testing: The social dimension. *International Journal of Applied Linguistics*, *16*(2), 242–258. https://doi.org/10.1111/j.1473-4192.2006.00117.x

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing*, *15*(1), 1–20. https://doi.org/10.1177/026553229801500101

Saldana, J. (2009). *Coding Manual for Qualitative Researchers*. Thousand Oaks, CA: Sage.

Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing*, *30*(3), 309–327. https://doi.org/10.1177/0265532213480128

Suzuki, Y. (2015). Self-assessment of Japanese as a second language: The role of experiences in the naturalistic acquisition. *Language Testing*, *32*(1), 63–81. https://doi.org/10.1177/0265532214541885

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing*, *11*(3), 321–344. https://doi.org/10.1177/026553229401100305

Winke, P. (2011). Evaluating the Validity of a High‐Stakes ESL Test: Why Teachers' Perceptions Matter. *TESOL Quarterly*, *45*(4), 628–660. https://doi.org/10.5054/tq.2011.268063

## Notes

Note 1. By "accuracy," we meant the accuracy of the placement decisions. That is, students are placed in classes appropriate for their language proficiency levels.

Note 2. We also ran a reliability analysis for the Can-do statements in each of the four different domains, and the Cronbach's α is also high for the separate domains (Speaking = .86; Writing = .84; Listening = .88; Reading = .86).

## Appendix A

### Oral Rating Rubric

**Foundational Level:** The student…

- communicates in one or two words or basic three to four-word phrases.
- cannot communicate due to the lack of vocabulary and usage of grammar.
- cannot express or elaborate on their responses when promoted (repeatedly).
- cannot produce logically structured responses.
- cannot be understood because response is incomprehensible.

**Level 1:** The student…

- is able to form basic formulaic responses.
- communicates using simple vocabulary and simple grammar tenses.
- has a difficult time getting ideas to connect in the response.
- has a difficult time elaborating on explanations and examples (when prompted).
- is difficult to understand because of hesitation and pauses.

**Level 2:** The student…

- is able to maintain a basic conversational dialogue.
- communicates using appropriate vocabulary and basic grammar tenses.
- has a difficult time connecting and forming ideas into a structured response (hesitates, circles, or repeated ideas).
- can expand on explanations and examples when prompted
- is understandable, but their speech is segmented due to pronunciation.

**Level 3:** The student…

- is able to sustain a longer conversational dialogue (intermediate).
- communicates using meaningful vocabulary and has understanding of grammar usage.
- is able to connect structure and organized ideas, but some responses may be illogical.
- can elaborate on explanations and examples without prompting.
- is understandable despite pronunciation errors.

**Level 4:** The student…

- is able to communicate in an advanced dialogue.
- communicates using academic vocabulary and has command of grammar usage.
- can elaborate on explanations and examples using facts and specific details.
- is able to produce logical responses, but could improve clarity in organization and structure.
- is understandable with minor segmental errors.

**Level 5:** The student…

- is able to communicate in a very native like academic dialogue.
- communicates using advanced academic vocabulary and uses grammar accurately in responses.
- can elaborate using accurate and factual details during explanations.
- is able to produce complex, clear, logical flow of interchangeable and connected ideas in responses.
- is clearly understood with little to no pronunciation.

### Copyrights