論　文

Creating the DIS Speaking Test for Placement: Considering Validity, Reliability, and Practicality

[1]沖　キャサリン　[2]Floyd H. Graham III

[1]同志社女子大学・学芸学部・国際教養学科・助教（有期）
[2]元同志社女子大学・学芸学部・国際教養学科・助教（特別契約教員）

[1]Catherine OKI　[2]Floyd H. Graham III

[1]Department of International Studies, Faculty of Liberal Arts,
Doshisha Women's College of Liberal Arts, Assistant professor
[2]Department of International Studies, Faculty of Liberal Arts,
Doshisha Women's College of Liberal Arts, Former assistant professor (contract)

Abstract

Placement testing presents a number of challenges for university language programs. Although there are a variety of benefits for creating one in-house, for practical purposes, commercial assessment tools are often chosen to accomplish the task of dividing students into groups of similar abilities. The goals and objectives for each program should inform the types of testing involved and appropriately delineating student abilities should typically be achieved through measuring all or a combination of several skills: reading, writing, speaking, listening, grammar and vocabulary. This paper details the steps used in the Department of International Studies (DIS) when creating the DIS Speaking Test for Placement, a speaking component integrated into the existing placement structure. This procedure is described and evaluated to show how developers tried to adhere to principles of reliability and validity in concert with practical issues. This paper does not purport that the test is in fact valid or reliable as further research will need to be conducted to better establish the degree to which those principles were achieved; however, the researchers do hope that this study can serve as an example of one department's attempt, highlight their efforts to identify where immediate improvements can be made to benefit the test's raters, and contribute to further dialogue about speaking placement test creation.

Keywords: placement test, oral assessment, validity, reliability, practicality

Students often see evaluations as instruments lending insight into their levels, progress, and proficiency, yet many teachers view them differently. As Hughes and Hughes (2020) pointed out "Language tests too often fail to measure accurately whatever it is that they intended to measure. Teachers know this. Students' true abilities are not always reflected in the test scores that they obtain" (p. 1). A common frustration felt by many language teachers occurs after students' English levels have been assessed upon entrance into a university and they are placed into classes where they should be with others of a similar level. Inevitably, and because of the varying factors that affect the accuracy of assessment, some students are misplaced, which can exacerbate classroom management challenges among other problems. At these researchers' university, speaking and writing skills class teachers began to vocalize their concerns about several students who were seemingly in a notably different level class than their demonstrated abilities. This initiated conversations amongst the faculty regarding a need to amend the placement exam which, at that time, consisted of only a reading, listening, and writing portion. The first revision was to place more emphasis on the writing scores because the reading and listening (receptive) classes were grouped into larger classes divided into just two levels ― the top half and bottom half students ― which yields less level homogeneity, whereas the writing and speaking (productive) classes were purposely divided into eight levels resulting in much smaller classes and more individualized teaching. In addition, the EAP curriculum of the Study Abroad program was being revisited based on a needs analysis conducted on over 100 students who had studied abroad, which had determined the need for more communicative based lessons. Therefore, when the April 2020 Freshman orientation was changed to a solely on-campus event and with the transition to an IELTS focus (away from TOEFL iBT), an opportunity presented itself and the department approved its addition.

Placement testing for language programs

Streaming university language students into classes of similar levels is widespread, so much so that many educators take it for granted. Although some, like Rodriguez-Yagi and Rupp (2020), declared benefits in certain cases for mixed groupings in language classes, much research points to the overwhelming advantages of grouping students into language lessons with others of similar abilities (e.g. McMillan and Joyce, 2011; Wu, Tsai, and Chu, 2018) or the difficulties posed by mixed grouping (Al-Shammakhi and Al-Huamidi, 2015). Hille and Cho (2020) observed,

> Given that the premise of placement testing is to match students' instructional needs with a level of instruction for optimal teaching and learning, the effectiveness of placement testing has both instructional and financial ramifications for educational institutions and students. Accurate placement is expected to optimize teaching and learning because a placement result indicates the level of instruction that a student needs. By the same token, when students are misplaced, both teachers and students are likely to struggle in the classroom as their expectations are in discord in terms of the level of instruction (p. 454).

When referencing students' language abilities, one must consider the goals and objectives of the program itself ― for example, is it a general English course or an EAP one with students intent on study in universities abroad ― and understand that there can be a wide disparity in each language skill of an individual student. Although many have outlined recommended practical steps and standards to follow when designing tests (e.g. American Educational Research Association AERA, American Psychological Association APA, National Council on Measurement in Education NCME, 2014; Hughes & Hughes, 2020), Murray (2001, pp. 28-38) provided comprehensive guidelines for the creation of placement tests in particular. Included in these steps are to:

1) Create an assessment team ideally made up of administrators, curriculum-coordinators, teachers, and even students.
2) Define a linguistic and biographical profile of the test-takers, or "participant identification," to assist in establishing suitable materials and test

procedures.

3) Define placement test objectives, specifically what will be tested and how it will be done.
4) Decide on the type and content of the test from among direct or indirect, discrete point or integrative, and norm- or criterion-referenced test-types followed by content reflecting curricular objectives.
5) Make the test heeding recommendations for the language being tested.
6) Decide the scoring system or rubric for rating the test, choosing between less-reliable holistic or more-reliable analytical scoring schemes, so as to create an assessment that clearly matches the student performance to the criteria required.
7) Pilot the test in order to analyze whether the test will create accurate results to aid in differentiating students' language abilities.
8) Train the test raters and administrators charged with administering and scoring the test in order to maintain consistency and accuracy in test delivery and scoring.

Despite the ubiquity of placement testing, there is a dearth of research on the process, particularly of developing a speaking component in Japanese contexts, partially because speaking is least frequently measured (Shimizu, 2002) and poses a variety of difficulties (Ockey, 2017) in these assessments. Hence, this paper attempts to add to that discussion by elaborating on how one program developed a speaking portion for their existing placement test, while considering how to balance the three principles of validity, reliability, and practicality in its initial iteration.

Considerations for Developing Oral Assessments

Assessing speaking/oral assessment

The main objective of a speaking test is to elicit spoken English samples from students sitting the placement test, and as Ockey and Li (2015) pointed out, the most widely used test of spoken English is "oral proficiency interviews" (p. 7) which require the test-takers to listen and respond to questions posed by the interviewer. This interview format is considered a type of direct testing of oral proficiency, making it an integrative activity because more than one skill is required (Hughes & Hughes, 2020). There is "no script and no preparation on the learner's part for any special activity" (Underhill, 1991, p. 31) and is thus recommended along with a "comprehensive marking scheme" (Murray, 2001, p. 43) which Underhill (1991) stated must be accurately-worded to benefit speaking test administrators. Additionally, it was suggested that administrators be class teachers who can give the test in a place which is comfortable for candidates (Murray, 2001).

As for scoring applications, analytical scoring, which requires examiners to assess a multiplicity of tasks, offers numerous advantages for evincing greater reliability and as Hughes and Hughes (2020) remarked, "the very fact that the scorer has to give a number of scores will tend to make the scoring more reliable" (p. 103). In speaking, Ockey and Li (2015) identified four oral proficiency constructs which can be assessed. First, interactional competence, which is displayed, for example, when in real-time interaction stimuli from someone else is orally responded to appropriately. Next, in this response, appropriate prosodic and segmental features, in other words, phonology is used whereby language is segmented so words are imbued with meaning through their articulation, including pitch, intonation, and stress among other prosodic features. Additionally, the speaker must show the extent to which they can use grammatical and lexical depth and range to express themselves effectively. Finally, they must express themselves effectively with appropriate speech rate, use of pauses, and minimal repetition or language repair demonstrating proficient fluency. The IELTS academic Speaking paper is an example of a commercial test which assesses these. It is a criterion-referenced test (CRT) (British Council, n.d.a), meaning it measures students' oral proficiency according to 4 criteria — fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation (British Council, n. d.b) — parsed into bands from 0-9 to describe candidates' proficiency level (IELTS, n.d.a).

Validity

There are certainly many challenges in creating accurate and fair assessments. One of the most prominent ones, often declared the most important one, is validity. Although Messick's (1988, 1989, 1998) seminal work refocused the validity framework, there is still considerable debate about what exactly constitutes validity (Newton and Shaw, 2013) and how these measures should be used. This paper adopts Akbari's (2012) definition as, "whether a test measures what it is supposed to measure…in other words, a test should measure the intended skill, ability, or components, and should provide enough proof to its claim of relevance" (pp. 30-31).

Akbari (2012) quoted Harrison (1983) in declaring that most traditional accounts of test validity consider four specific areas: face validity, content validity, criterion-related (or empirical) validity, and construct validity. He explained that face validity addresses the appearance of the test, along with whether the students and test administrators accept that the test is assessing the skill or aspect it purports. For example, portions of the speaking part of the TOEFL iBT test also require proficiency in reading and listening, which could call into question the face validity degree to which speaking is truly being assessed. Akbari notes, however, that face validity is the factor that holds least relevance for appraising a test, and can be disregarded if other justifications need to take precedence. Content validity refers to the domains of the questions and methodology of the test and whether they are aligned with and adequately representative of objectives, goals, and overall design for which it is being used. Criterion validity is statistically based and locates coherence in a concurrent or predictive fashion, either by correlating scores with a similar, already established assessment of equal construct, or one that takes place sometime in the future.. Some researchers believe construct validity is the only measure of validity, and all others are unified under it (Messick, 1989). Construct validity relies on theory to establish a clear construct that is being measured. If there is no agreement on what, for example, speaking ability at the 4th band of IELTS is, then there can be no validity in how to measure it. There needs to be an agreed upon and established theory of the construct being measured, which "at the end of the day provides the general guidelines as to what goes into a test and what format the test should take" (Akbari, 2013, pg. 32).

To increase test validity, Hughes and Hughes (2020) recommend making explicit test specifications that account for the constructs the test aims to measure so as to then include representative content in the test. In addition to this, they suggest direct testing, whereby the candidates are performing skills "that we are interested in fostering" in tasks which are "as authentic as possible" (p. 58) and having scoring directly related to test content will increase validity.

Reliability

In essence, reliability is the concept that a test will produce scores which are consistent (Farhady, 2012) and is broken into test and score reliability. Test reliability refers to scoring being equivalent regardless of the time or place the test is taken, though, because test-takers are human, slight variability is expected if the same candidate sat the test on different days. Rater or scorer reliability is scoring consistency regardless of who marks the test. Like with test-taker variability, due to the subjective nature of humans rating answers which are open-ended like in speaking, some variation between raters is expected, but can be minimized through training. Having both test and rater reliability gives confidence that the examinees performance is being accurately measured. Rater reliability is further delineated between inter- and intra-rater reliability. Lee (2014) explains that the intra-rater reliability is how uniformly a rater will rate the same tests when given spaced out over time and inter-rater reliability as agreement between different raters marking the same test.

To increase the reliability of candidates performances and rater's scoring, Hughes and Hughes (2020, pp. 49-55) offered several suggestions. First, they recommend having enough items on a test to get a reliable sample of their ability while not making it lengthy enough that test-takers lose focus, affecting their performance. Next, they advocate choosing items which can effectively discriminate between more and less proficient students, with the caveat that tests can start with a few non-discriminating questions if the purpose is to reduce candidates' stress and increase their confidence. Other proposals include, not allowing candidates too much freedom in how they answer, therefore they should not be given the choice of what to answer and the range of what they must address should be narrow; questions must be checked by teachers and then piloted with similar students to be sure they are unambiguous; clearly expressed instructions should be provided and in the case of spoken instructions, a script should be prepared and read from; have test materials made clearly; give candidates test requirements ahead of the test; make administration of the test a uniform process devoid of distraction like background noise or movement; and create scorer reliability by making sure that:

1) Items are sufficiently objective.
2) Candidates are being compared directly so that, for example, they are not being assessed using different questions, thereby reducing freedom.
3) A detailed rating rubric is provided.
4) Raters are trained.
5) There are agreed upon ideas about appropriate answers and scores before the test.
6) There is more than one scorer, each working independently.

Balancing Validity, Practicality, and Reliability

Farhady (2012) explains that validity, reliability and practicality are the three key principles which all tests should meet and explains practicality as, in making, administering, and scoring assessments test developers will have to consider the availability of resources and that choices about the test will depend on the high- and low-practicality factors. Brown (2003, p. 19) defined a practical test as, "one that is not excessively expensive, it maintains appropriate timing, its administration does not represent difficulties and its scoring and evaluation procedures are specific and time efficient." In other words, can the test measure what it needs to when considering the availability of time, people, and the equipment necessary. For spoken tests, Al-Amri (2010) explained that "…using criteria and training in the use of grading rubrics/scales and real-life test formats in their contexts can increase not only validity of the test but also its reliability and practicality" (p. 113) promoting a better balance between these three principles. The next section will describe how the speaking placement test was created in order to then discuss how the developers considered these three concepts in the process, using feedback from raters to highlight how better inter-rater reliability could be established as a first step in evaluating this test for future improvement.

Developing the DIS Speaking Test for Placement

First, once it was decided that adding an oral assessment section to the placement test would enhance the process of dividing the students into streamed groups, meetings were initiated amongst the faculty and administration to determine its overall practicality, but also to address methods to establish pertinent content and issues of validity and reliability. With our curriculum transitioning to include coursework in IELTS preparation, which is now gaining widespread acceptance as a measure of English language readiness for study abroad, and because placement tests should be connected to curriculum objectives, it was determined that the speaking placement test should be based on the IELTS Speaking test.

The IELTS speaking test is an interview conducted face-to-face with a candidate and an examiner, lasting 11-14 minutes and covering three task patterns designed to test examiners' ability to speak first about familiar topics in an interview style pattern, then give an extended answer for 1-2 minutes about a personal experience, incorporating three ideas listed on a topic card given to examinees, and finally, to discuss about more abstract topics with the interviewer (IELTS, n.d.b). It was decided to keep the three speaking sections of the IELTS, but knowing the oral assessment must be done within a three hour and twenty minute window with approximately 96 incoming Freshman, it was determined that the DIS Speaking Test for Placement would be a 10-minute face-to-face interview with eight raters who would then have five minutes to score the interview and transition to the next candidate.

A rubric, using the four IELTS criteria of fluency and coherence, lexical resource, grammatical range and accuracy, and pronunciation, was created and the point system was scaled to the other placement test skill sections (See Appendix). The four criteria were each described, detailing the language profile of the strongest (10-points) and the weakest (2 points). These demonstrated that a DIS "perfect 40" would be a competent user of English but not a native-like one, thus some mistakes would be made and therefore mirror an IELTS speaking band of 6. Conversely, the weakest student was aligned with an IELTS band 3, or someone with limited proficiency who made frequent mistakes and struggled to communicate their ideas effectively.

Initial questions were derived from online IELTS sources and three test sets corresponding to the IELTS format were created. Questions which could elicit simple tenses, present perfect continuous tense, conditionals, frequency words, and idiomatic expressions were sought, while yes/no questions were generally avoided. Then, to choose the best questions from these sets, each was piloted up to six times among eighteen first-year students from each of the eight class levels. The aim was to determine which questions would be easy enough that students could answer with some detail, whilst placing enough language demand on them to allow raters to discriminate between higher- to lower-level students; in other words, questions which elicited enough spoken language at a spectrum of proficiencies for examiners to rate according to the four criteria. Pilotted students were also asked their opinions about the questions. The professors involved in the piloting made written comments after the trials and in a meeting final questions were chosen.

Next, training was planned for the eight raters, seven of whom had never rated an IELTS-type Speaking test nor taught an IELTS-related class, three of whom were new to the department, and two of whom do not teach English in the department. Using the final questions, four interviews were recorded, with four students from the pilot stage who agreed, and a handbook was created to:

1) Make practical test-day details explicit to the raters.
2) Describe the four scoring criteria for the test.
3) Show the 10-point and 2-point descriptor for each criteria.
4) Provide links to the IELTS band 6 video and in-house inter-rater reliability training videos.
5) Show the final test with scripted directions and questions.

Training then began by emailing the raters with links to inter-rater score sheets and instructions to read the attached handbook to understand the criteria and do two activities. The first activity was to watch a video of an IELTS band 6 interview while referencing the provided score feedback table and to rate the three videoed interviews (a higher-, medium-, and lower-level of proficiency), before emailing the coordinator to check their scores, which would be discussed in the interrater reliability meeting.

In the interrater reliability meeting all eight raters discussed the criteria and three videos. Scores of the three videoed students were analyzed and debated in small groups and then as a whole group to clarify reasoning behind final scoring decisions. Then, the fourth videoed interview was watched and assigned points to ascertain score agreement. Finally, an email summarizing the meeting was sent out to all participants. This email first highlighted important points which came up about how to conduct the test, like focussing on students' language, how to keep them talking, and to save the question "Why" for when they did not give details on their own. It then explained tips for scoring, like possibly concentrating on scoring the pronunciation criteria in Part 1 and then moving to other criteria in the second and third parts, or how to better distinguish between a score falling below or above a 6.

Additionally, prior to the test day, the student leaders, who are chosen every year to help run and organize the Freshman orientation, were trained to assist examiners by making sure that test-takers were in the right place and ready with their pencils when it was their turn. Test-takers were also given a description of the test two days prior to the test and on the morning of the test the coordinator spoke to all freshman students to remind them about what to expect. Student leaders led students to waiting rooms, where the examiner would retrieve and lead them to the test room. Examiners had their scripted directions and questions, Part 2 question cards to give test-takers for that portion of the test, and score sheets where they were to rate each student in real-time or shortly after the 10-minute interview ended.

Finally, the six raters filled in a survey asking about the interrater reliability training experience and their thoughts about scoring on test day to help determine areas where improvements were necessary. Two raters were not included in this questionnaire because one was an IELTS examiner with extensive inter-rater experience and the other created all materials including adapting the IELTS rubric to a department rubric, making the training videos, and writing the handbook, making them unable to objectively evaluate the process.

Discussion

This section illustrates how the developers considered validity, reliability and practicality in creating, administering, and interpreting the results of the test and what rater feedback revealed about ways to improve inter-rater reliability as a first step in evaluating and refining this assessment tool.

The speaking portion of the placement test could be regarded as relatively low-stakes because it is one of four scores used in the process and egregious mistakes can be corrected easily. Nonetheless, in order to achieve an acceptable level of validity in this first attempt, recommendations from Hughes and Hughes (2020) were followed in several places. First, the explicit test specifications were derived from an analysis of the construct by stakeholders including students, teachers, and administrators: speaking skills with questions which were trialed, modified, and chosen in order to elicit specific and ample language samples representative of the necessary oral proficiency constructs. Moreover, the testing was direct, through a face-to-face interview, using accepted criteria based on IELTS parameters and related to the skill being measured to score the test with content gleaned from the curriculum and its objectives. As a result, the students' performance on the test can yield a relative expectancy for how the students would perform in the classes they would be taking.

The reliability of the test conditions created for candidates to perform consistently were also considered in a number of ways. Although our time limitations required us to shorten the length of the individual interviews, Hughes and Hughes (2020) suggest that 5-10 minutes is considered a sufficient length "to prevent gross errors in assigning students to classes" (p. 129) and we believe that maintaining three sections of questions which are varied enough but challenge test-takers to display different speaking abilities could support the test's reliability. That said, whether these were discriminatory enough or not requires further investigation. Moreover, interviewers controlled the questions being answered and in Part 2 what should be addressed in their speech is explicitly stated on the task card, asking all students to respond to the same prompts. Additionally, test items were checked and piloted and feedback was solicited about what the students thought made some questions confusing or too difficult for the freshman. The examiner script containing the questions and directions for candidates was also piloted to be sure directions were clear. Finally, prior to the test, bilingual instructions were disseminated to test-takers, reviewed the morning of the test, and student leaders were enlisted to support the smooth administration of the speaking assessment on orientation day.

Rater reliability was also considered. While questions could not be objective because this test is a direct speaking interview with open-ended questions, the same questions were provided for all examinees and they were all rated according to the same criteria using a detailed rubric with trained raters who met prior to testing and agreed upon appropriate answers and scores. That said, scoring was only done by one independent rater, which compromised scoring reliability. However, for practical reasons, this was unavoidable. While double-rating of interviews would have likely improved reliability, scores had to be entered immediately in order to meet the time constraints of rating the writing portion and then holding the placement meeting that afternoon. This, however, was seen as a place where practicality could override reliability in that students could be moved to a different class level should a clear mismatch be identified.

Because reliance on a single rater was the most obvious weakness in the application of the speaking component of the placement test, establishing inter-rater reliability is even more important. Consequently, in order to get a clearer perspective of the reliability of their assessments, raters completed a survey about their confidence in scoring each of the four criteria, the usefulness of the training materials and activities, and any other issues to illuminate areas where improvement could be made. Based on several comments, it seemed that the methods employed to ensure inter-rater reliability were effective. For example, one examiner claimed, "Taking into account that some of the raters have no experience in IELTS test, having these workshops makes a huge difference," while another added, "The fact that we had established a rating range, specific for DIS students, made me feel less doubtful about my evaluation." On the other hand, another rater wrote, "Establishing the grading criteria for pronunciation was very helpful, so perhaps establishing rating ranges for other areas could be helpful as well" revealing some issues in the understanding of those criteria.

In fact, a question on rater confidence for the rating criteria illuminated that the majority of raters were not confident in their measurement of grammatical range and accuracy. Furthermore, half felt "confident" and the other half felt "not very confident" for lexical resource. One mentioned, "I felt it was easier to rate discourse competence than to check the small grammatical mistakes," and the comment "not being able to" about rating lexical resource indicates that these were more difficult for raters to distinguish. Ockey and Li (2014) stated that grammar and lexis could arguably be combined as the relationship between these scores is strong and other studies illustrated that raters do not distinguish between these when assessing oral communication (e.g. Batty, 2006); however, as teachers must come to understand these differences in order to better communicate them in their classes reserved for IELTS preparation, it might be more beneficial to amend the rater training for these criteria instead. Moreover, including a number of different scoring tasks maintains the range of criteria for analytical assessment purposes and contributes to reliability.

Additionally, some survey answers revealed that the pre-meeting portion needs to be refined. Firstly, two raters remarked that the handbook was "not very useful." Others commented, "The reason why I wrote that the interrater meeting was not very useful is that, personally, it confused me a lot. I would have graded much stricter, but in the meeting I was surprised to see and hear that I am expected (or at least that was my impression) to grade less strictly," and "I think what I may be confused on is what the speaking score represents and to who are we comparing the student's level to." These indicate that the handbook was perhaps not clear enough in explaining how we aligned our highest scores with the IELTS band 6 score. Therefore, more emphasis on the IELTS band 6 video activity's place in the training would likely benefit the raters.

Overall, examiners found watching and assessing the three videos ahead of the meeting most useful to understanding the rubric and criteria, followed by the summary email after the meeting, and then the interrater meeting. However, one wrote, "I think they need to watch maybe five examples of different levels of speaking." This might reduce the degree to which raters relied on other determinants in their scoring assessments as they would get more practice prior to test day. Although raters were expected to only refer to the criteria provided to score interviews, some raters indicated that other factors influenced their scoring, which likely compromised the scoring reliability. For instance, raters stated that students' "confidence, ability to communicate well, and body language," or "…their ability to establish communication, that is to say, to respond in a conversational way to the questions asked," guided some of their grading. Mentioning these types of comments in the edited handbook and in future inter-rater reliability meetings may reduce reliance on factors outside of students' language use.

Lastly, creating a training overview video to introduce the handbook and emphasize some important points about the criteria, scores, and the relationship with the IELTS band 6 video might enhance rater's understanding of the scoring. Additionally, after receiving raters' initial training scores, immediately sending feedback rather than waiting for the meeting may be beneficial. Doing so could lend them time to think about comments and look back at the criteria in the handbook in order to better distinguish between them, especially lexical resource and grammatical range and accuracy. This understanding could then be further refined in the meeting. Better integrating enhancements like these should assist in establishing placement results with greater levels of reliability.

Future Research

The study does not show the actual validity or reliability of the test, but this is a step in the processes of the ongoing evaluation of it. Therefore, the focus of future research should include studies of the test's validity and reliability. For example, placement levels should be compared and correlated to initial IELTS scores taken between July and August to yield a clearer demonstration of the degree to which the test results may have achieved a satisfactory criterion validity. An additional way to support this would be to look at the predictive validity where the teachers and/or students would be asked their thoughts on student placement. In terms of reliability, determining the degree of decision consistency and finding ways to better communicate these distinctions to raters in an effort to enhance rater reliability, as raters will vary from year to year, would be beneficial.

Conclusion

This paper described the method used to create the DIS Speaking Test for Placement in order to evaluate the process and understand the place of validity, reliability, and practicality in its development. By looking closer at inter-rater reliability through the eyes of novice raters, areas where immediate improvements can be made in the training of scorers were identified. The raters' feedback showed that weaknesses exist in the understanding of some scoring criteria and therefore more clarification of some training elements is needed. Finally, a more thorough reliability and validity battery is also required to determine how accurate the test is. Ultimately, this process has helped these researchers reflect on the role of testing principles and illuminated the need to continue exploring these issues to benefit the Department of International Studies' students and teachers.

References

Akbari, R. (2012). Validity in language testing. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 30-36). Cambridge University Press.

Al-Amri, M. (2010). Direct spoken English testing is still a real challenge to be worth bothering about. *English Language Teaching*, *3*(1), 113-117. https://doi.org/10.5539/elt.v3n1p113

Al-Shammakhi, F., & Al-Humaidi, S. (2015). Challenges Facing EFL Teachers in Mixed Ability Classes and Strategies Used to Overcome Them. *World Journal of English Language*, *5*(3), 33-45. https://doi.org/10.5430/wjel.v5n3p33

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

Batty, A. (2006). An analysis of the relationship between vocabulary learning strategies, A Word Association Test, and the KEPT. *Studies in Linguistics and Language Education: Research Institute of Language Studies and Language Education*, *17*, 1-22.

British Council. (n.d.a). *Criterion-referenced test*. TeachingEnglish. https://www.teachingenglish.org.uk/article/criterion-referenced-test.

British Council. (n.d.b). *Evaluating speaking - the IELTS speaking test*. https://www.teachingenglish.org.uk/article/evaluating-speaking-ielts-speaking-test.

Brown, H.D. (2003). *Language Assessment: Principles and Classroom Practices*. Longman.

Farhady, H. (2012). Principles of Language Assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 37-46). Cambridge.

Harrison, A. (1983). *A Language Testing Handbook*. McMillan Press.

Hille, K., & Cho, Y. (2020). Placement testing: One test, two tests, three tests? How many tests are sufficient? *Language Testing*, *37*(3) 453-471. https://doi.org/10.1177/0265532220912412

Hughes, A. & Hughes, J. (2020). *Testing for Language Teachers* (3rd ed.). Cambridge University Press.

IELTS. (n.d.a). Speaking: Band Descriptors (public version). https://www.ielts.org/-/media/pdfs/speaking-band-descriptors.ashx?la=en

IELTS (International English Language Testing System) Test format. (n.d.b). Cambridge Assessment English. https://www.cambridgeenglish.org/exams-and-tests/ielts/test-format/

Johnson, R. C., & Riazi, A. M. (2016). 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

Lee, J. (2014). Rater Reliability on Criterion-referenced Speaking Tests in IELTS and Joint Venture Universities. *English Teaching in China, 4*, 16-20.

McMillan, M. & Joyce, B. (2011). Teacher perspectives on student placement in university EFL programs. *Journal of Nepal English Language Teachers' Association* (*NELTA*), *16*(12), 70-81.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer (Ed.), *Test validity* (pp. 33-45). Erlbaum. https://doi.org/10.1002/j.2330-8516.1986.tb00185.x

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). McMillan.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(4), 35-44.

Murray, J. R. (2001). Steps and Recommendations for More Accurate Placement Test Creation. (ERIC Document Reproduction Service No. ED 453 655).

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods*, *18*(3), 301-319. https://doi.org/10.1037/a0032969

Ockey, G. J., & Li, Z. (2015). New methods and not so new methods for assessing oral communication. *Language Value*, *7*, 1-21. https://doi.org/10.6035/LanguageV.2015.7.2

Ockey, G. J. (2017). Approaches and challenges to assessing oral communication on Japanese entrance exams. *JLTA Journal*, *20*, 3-14. https://doi.org/10.20622/jltajournal.20.0_3

Rodriguez-Yagi, M., & Rupp, M. J. (2020). Gaining an emic perspective on streamed versus mixed-level EFL classes at the tertiary level in Japan. *Proceedings of the School of Agriculture, Tokai University 39*, 1-7.

Shimizu, Y. (2002). Survey research on the use of placement tests at four-year universities in Japan. *Ritsumeikan Studies in Language and Culture*, *14*(1), 231-243.

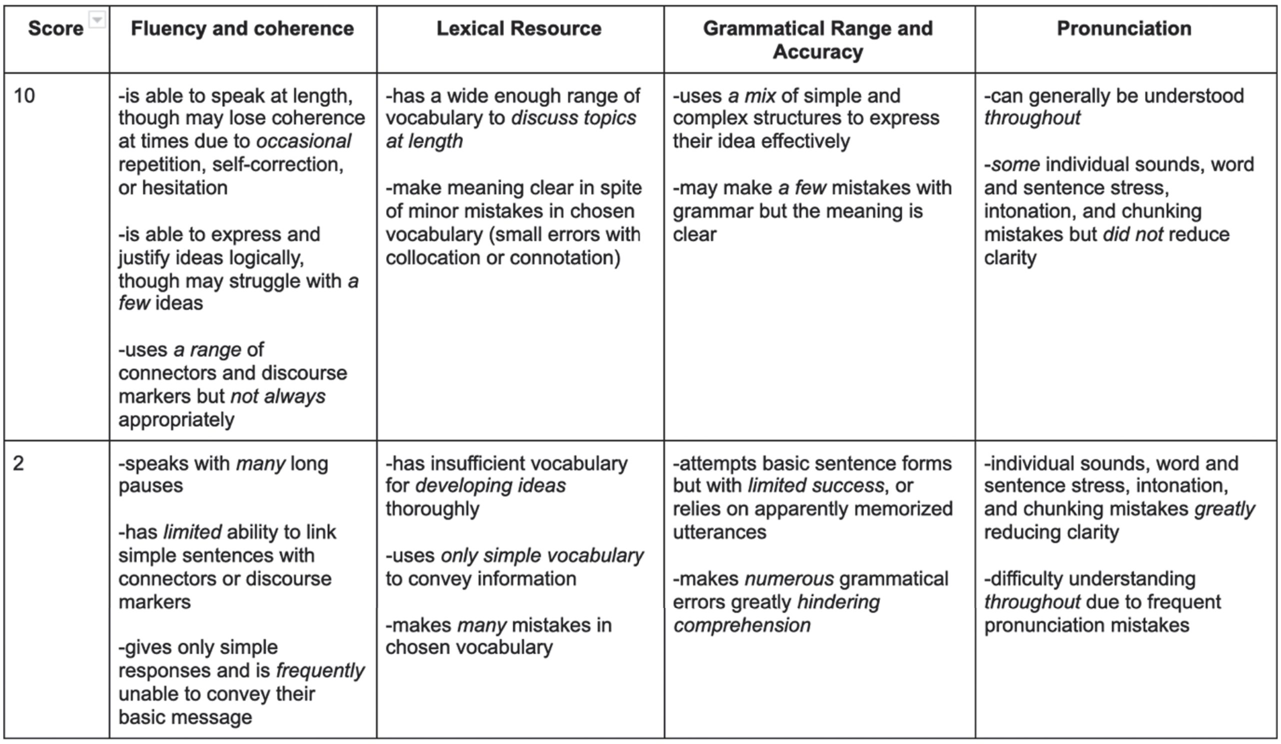
Underhill, N. (1991). *Testing spoken language: A handbook of oral testing techniques*. Cambridge University Press.

Wu, C., Tsai, C., & Chiu, Y. (2018). A longitudinal analysis of ability grouping with college EFL learners. *TESOL International Journal*, *13*(1), 20-32.

Appendix

DIS Speaking Rubric

A score of 10 and a score of 2 are defined here:



Score	Fluency and coherence	Lexical Resource	Grammatical Range and Accuracy	Pronunciation
10	-is able to speak at length, though may lose coherence at times due to *occasional* repetition, self-correction, or hesitation -is able to express and justify ideas logically, though may struggle with *a few* ideas -uses *a range* of connectors and discourse markers but *not always* appropriately	-has a wide enough range of vocabulary to *discuss topics at length* -make meaning clear in spite of minor mistakes in chosen vocabulary (small errors with collocation or connotation)	-uses *a mix* of simple and complex structures to express their idea effectively -may make *a few* mistakes with grammar but the meaning is clear	-can generally be understood *throughout* -*some* individual sounds, word and sentence stress, intonation, and chunking mistakes but *did not* reduce clarity
2	-speaks with *many* long pauses -has *limited* ability to link simple sentences with connectors or discourse markers -gives only simple responses and is *frequently* unable to convey their basic message	-has insufficient vocabulary for *developing ideas* thoroughly -uses *only simple vocabulary* to convey information -makes *many* mistakes in chosen vocabulary	-attempts basic sentence forms but with *limited success*, or relies on apparently memorized utterances -makes *numerous* grammatical errors greatly *hindering comprehension*	-individual sounds, word and sentence stress, intonation, and chunking mistakes *greatly* reducing clarity -difficulty understanding *throughout* due to frequent pronunciation mistakes