Assessing Second Language Speaking

Maria Barbera, Shihua Chen, Nosheen Malik, & Kayla Rakita

Northern Arizona University

Abstract

Assessing second language production, particularly in the area of speaking, is very important for educators to know of their students' preparedness for what is to come and their mastery of previously learned material. However, many considerations should be made in developing a second language speaking test, such as the reliability and validity of the instrument. This paper details the creation of a speaking test by amateur test developers, the test's administration, and finally, an analysis of test results to see whether the test is valid for its context—all of which were conducted in the Program in Intensive English at Northern Arizona University. Though experimental, this test was relatively successful. This analytical process is important for guiding the development of tests and building an understanding of how tests do or do not achieve their purposes, as the impact of a test has great power to change the course of a language learner's aspirations and future opportunities.

Keywords: assessment, test, rubric, speaking, validity, reliability

Background

Speaking is an essential form of communication in an ESL context, in which learners are immersed in a world that otherwise could not understand them. However, speaking has a reputation of being the most difficult, elusive language skill to assess (Ginther, 2012). It involves the careful development of construct and content to ensure generalizability to the real-world language skill, as well as rater training to increase reliability of scoring. Issues of generalizability and score interpretation are more relevant to speaking tests than any other skill assessment (Fulcher, 1997). This project was an investigation of test techniques that research has shown are important in the design and analysis of speaking assessments. To make this investigation more meaningful, a short speaking test was developed by four graduate students. This project investigated five research questions:

1. What is the most difficult component of the speaking test?
2. How did students perform on the test, and what is the distribution of scores?
3. What argument, if any, can be made for this speaking test’s validity?
4. How reliable is this speaking test?
5. Is this speaking test effective for its purpose?

This report will cover the test’s methods, results, and the relevance to PIE and second language learning.

Methods

This test was administered to twenty second language (L2) students in the Program in Intensive English (PIE) at Northern Arizona University. However, the number of students whose scores were analyzed was later narrowed down to twelve students who agreed to participate in research. The age range of participants was from 18 – 29 years old. All students were at the beginning level of English proficiency, and had been placed into two different sections of Level 2 in the PIE.

Task 1, the individual task, begins with a listening passage (Food for America) about fast food. Once the listening is over, the test taker has 3 minutes to prepare their answer and 1 minute to record a response about their own opinion of fast food (see Appendix C for the full test). Also on Task 1, there is a box provided with fast food-related vocabulary words, three of which must be included in the test-taker's response. The rubric for this task has three bands, for delivery, content, and language use, all of which are weighted equally. Task 2 in this test is a paired speaking task that requires two test takers to interact. For this task, examinees have 3 minutes to prepare with their partner and then must produce a 2 minute response. In addition, the prompt gives directions for each step of the conversation, and explains what kind of information the exchange should include from each participant. The rubric for this task has four bands, for content, delivery, language use, and conversational skills.

For the scoring of this test, two raters scored Task 1, and two different raters scored Task 2. The set of raters for Task 1 scored them independently, and did not talk about the reasoning behind assigned scores when they were more than .5 points apart, while the raters for Task 2 scored together at the same time, and discussed responses that were scored more than .5 points apart. To determine the total score of each test-taker, the scores of both raters were averaged for each task and then added together. The scores that would have hypothetically been reported to students would be (a) the total score of both tasks transformed to a percentage of total points possible; and (b) the scores for each subconstruct, averaged between rater 1 and rater 2 for each task and added together, and then transformed to a percentage of the total points possible.

This test was administered on Thursday, November 14th and Friday, November 15th to two different sections in the PIE. Examinees were assessed once, and the test's administration took 30 minutes. Two of the test developers administered the test to the two sections. Recorders were used for each test-taker's response and then were collected afterward. Sound files were then downloaded for the test-takers who agreed to participate in research. These were then scored by the four raters the next week, on November 18th.

Results

Once responses were scored, the subconstructs for each task were analyzed for item difficulty (P) and item discrimination (D). These types of item analyses are usually used to represent a norm-referenced test, but can still inform a criterion-referenced test such as this one (Miller et al., 2009). The most difficult component of the speaking test was language use on Task 2, with a mean value of 3.22, while the least difficult component was language use for Task 1, with a mean of 3.22. In addition, item discrimination refers to the extent that items with high scores correspond with the higher group within the sample of participants; positive proportions indicate that the item is discriminating the same way as the test as a whole (Miller et. al, 2009). The discrimination values for subconstructs on each task all fell between .71 and .92, with consistently higher values for Task 2. These are all strong proportions for item discrimination.

For the twelve participants who took this test, there is a negatively skewed distribution. The cut-off score of 20 resulted in 2 participants' scores falling below the cut score as non-masters, and 10 participants exceeding the cut score as masters. The highest total score was 27.7 out of a possible 28 points, while the lowest total score was 18.25. The standard deviation for the total score was 3.08, and the mean for the total score was 24.31 out of 28. Descriptive statistics are also presented for the subconstructs for both tasks combined, averaged between the two raters, as this is the information given to test-takers on their score report forms. There was the greatest spread between scores on the content subconstruct, as the standard deviation is .94. The standard deviation for the total test scores is 3.08.

The test was also examined for its validity. In terms of content validity, the variety within the two tasks show that they give an accurate representation of the TLU domain—first the integrated, individual task which is representative of the classroom setting domain as speaking is almost always accompanied with listening, and second, the paired task which is representative of a real face-to-face interaction between two speakers. In addition, there seems to be sufficient evidence for criterion validity, as the teachers of the students, upon looking at the scores of different students, were not surprised at participants' relative performance as indicated by their scores.

The test reliability is high, and so is the inter-rater reliability on both tasks. Even the first task, which had lower inter-rater reliability than the second task, had a minimum value of .70, which is still relatively high. The standard error of measurement is small as well, and would only include one more test-taker as falling below the cut-off score if applied. Finally, the distribution of scores is highly appropriate for this test's purpose, achievement, as it is negatively skewed and a greater proportion of students achieved higher scores than lower scores.

Relevance to PIE and Second Language Learning

This project is relevant to PIE and second language learning because of its strong results as well as its experiential significance for the test developers. First, the test revealed which students in these two sections were at a point in their speaking ability that meant they were meeting course objectives. Further, this project allowed graduate students to have the opportunity to pilot a test of their design and find out where improvements needed to be made. Although much of the purpose was hypothetical, the test developers came to the conclusion that with necessary revisions, this test could be used in an authentic testing setting, with real consequences to students playing out. Thus, it furthers the field in that it provided second language teachers a hands-on experience that resulted in a better understanding of second language assessment, and allowed them to create an effective testing instrument. In addition, the piloted test used methods very similar to the testing methods of PIE, and was shown to be effective—which supports the credibility of assessment practices at PIE.

References

Fulcher, G. (1997). The testing of speaking in a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 75-86). Norwell, MA: Kluwer Academic Publishers.

Ginther, A. (2012). Assessment of speaking. In C. Chapelle, *The Encyclopedia of Applied linguistics* (7 pages).