

【資料】

Ensuring Equivalence of Alternate Forms of Achievement Speaking Tests

Judith Runnels

Abstract

Oral achievement tests aim to give students the opportunity to demonstrate how well they can use language they were previously exposed to and practiced using in the classroom. Literature in the field of language assessment however, focuses largely on proficiency or placement testing, much to the dismay of educators. The current study expands on the work of Fulcher (2004) and Luoma (2004) and presents some novel steps that can be taken in the preliminary stages of achievement test development to ensure assessment is more representative of classroom conditions and leads to equivalent alternate forms of a test. The equivalence of the 8 forms of the sophomore General English speaking test at Hiroshima Bunkyo Women's University was examined and statistical analyses revealed no significant differences across test forms. Involving teachers in the development stages appears to be one way to ensure equivalency across alternate test forms. Consulting teachers appears to contribute to producing achievement tests that are representative of the curriculum.

Introduction

In language assessment, tests can be classified as either achievement or proficiency. In contrast to the latter, whereby the main purpose is to assess what a speaker's existing language ability is and how that reflects their underlying knowledge of language, an achievement test differs in that its main goal is to assess the knowledge and skills that were presented in a designated classroom situation (Gronlund, 1998). This may be in the form of an end of year exam or a mid-term assessment based on the content of the previous units. The test scores can be used as part of the grade of the course or to determine if the student is prepared to advance through the system - high-achievement scores typically indicate mastery of the material of prior instruction. Essentially, achievement tests aim to give students the opportunity to demonstrate how well they can use language they were previously exposed to and practiced using in the classroom, and not simply, how good their English is.

When developing an achievement test, developers usually begin with a list of standards that include information about what the students are expected to learn. The challenge of achievement tests however, is to ensure that the test developers can write items that accurately reflect what happened in the classroom in their assessment. This is

no easy task since not only is there often a gap between test developers and the educators themselves, but speaking tests are notoriously complicated because of the many factors that can influence one's impression of another's proficiency in a language. For speaking assessments in particular, this might include pronunciation, pausing, recasting, overall accuracy or speed of speech (fluency). Furthermore, even if the educators themselves are the ones designing the test, very little time is devoted to this in teacher training courses and assessment in general often falls to other areas of language education such as materials design or curriculum development (Lindholm-Leary, 2001).

Given that speaking tests usually make up a percentage of the students' final grades and that test design is time consuming, competing with grading and lesson preparation for example, it should not be sacrificed for other aspects of course assessment. Luoma (2004) and Fulcher (2004), among many others, provide step-by-step guides on how to develop a test. Nonetheless, these texts are not specific to achievement testing and teachers do not always have time to read numerous texts before developing their assessment. The goals of the current paper are to adapt this work within a more achievement-test-focused framework and to provide a teacher-friendly procedure for test development. Some preliminary steps that can be taken for teachers who are interested in making their tests more representative of the curriculum are: summarizing the lessons based on approximate time spent in class, presenting teachers with a survey of the curriculum's major tasks to determine what the teachers themselves deemed to be the most important aspects of the curriculum to determine what they thought students should be able to do by the end of the semester. The survey also includes questions about task number and rating styles. Based on the feedback from the survey, test prompts are developed and in following the well-established processes of Fulcher (2004) and Bachman and Palmer (2010), piloting and norming sessions held. Finally, scores for all of the students are measured and compared across test versions to test for equivalency. By following this process, it is expected that all test versions will be scored evenly across all raters and there should be no significant differences across test prompts.

Method

The following section describes the participants, materials and procedures for the survey for which the test used as a basis. Proceeding this, the second sub-section provides information about the test itself, including details about the test-taking population.

Survey

Participants

Eight Sophomore English teachers from the Bunkyo English Communication Centre (BECC) at Hiroshima Bunkyo Joshi Daigaku (HBJD) were surveyed on task-content, rating scales and task number.

Materials

Due to the communicative task-based nature of the General English curriculum, the test was performed in a paired-format, mirroring the classroom environment. The

Ensuring Equivalence of Alternate Forms of Achievement Speaking Tests

tasks teachers chose from were based on major foci of the lessons. Teachers, using a Likert Scale of 1-5, indicated their agreement to the inclusion of 7 potential tasks on the test (selected since they were the major tasks from the lesson summary). The teachers also indicated their agreement to the number of tasks to be included on the assessment (2, 3 or 4), the task length (1 minute to 4 minutes) and whether they preferred using a holistic or analytical rubric. This resulted in 3 two minute tasks for the final assessment.

Procedures

The survey was administered electronically, using online surveying software from www.surveymonkey.com.[©]

Test

Participants

Test takers were 291 second year (sophomore) General English students from HBJD who are required to complete two years of the General English program to be eligible to graduate. The assessment was worth 15% of their semester grade and was taken at the end of the first semester in their regular class room and class time.

Materials

The feedback from the teacher survey was incorporated into the test by selecting the top 3 ranked tasks. Four different test versions were created, with each version having different prompts for Student A and Student B (essentially, 8 different versions of the test).

Procedures

Following Fulcher (2004) and Luoma (2004), after the test was created, it was piloted using 20 students (10 pairs) from two different classes of the sophomore General English cohort. Using the samples from the pilot session, the rubric was altered, to better reflect the types of utterances produced by the pilot students (see Luoma, 2004 for rubric development). Following this, a norming session was held for all Sophomore English teachers (Weigle, 1994). Finally, the test was implemented during the final two weeks of the semester and the scores collected for analysis.

Results

The factors being measured were Test Form (8 in total), Test Version (1 through 4) and Prompt (A vs. B) as the independent variables and overall test score as the dependent variable. ANOVAs for Test Form, Test Version (1 through 4) and Prompt (A vs. B) were run across all raters for the dependent measure test score to determine any significant differences between tests. Additional analyses compared test scores across raters and tasks.

In total, 291 test takers were rated. Table 1 shows the frequency, mean score (%) and standard deviations for all test forms. A univariate ANOVA, to seven degrees of freedom, indicated that there were no significant differences between test forms ($F = 0.52$, $p = 0.81$).

Table 1

Frequency, Mean (%) Score and Standard Deviation are shown for each Test Form

Test Form	<i>N</i>	<i>M</i>	<i>SD</i>
1A	50	79.6	8.37
1B	49	76.7	9.34
2A	51	77.9	8.45
2B	52	78.6	8.12
3A	24	78.2	9.88
3B	23	76.2	11.03
4A	21	78.1	7.79
4B	21	78.3	10.09
Overall	291	77.9	9.13

In order to reduce the degrees of freedom for a more specific comparison across test version, further analyses indicated that across versions 1 through 4 (3 degrees of freedom), there were also no significant differences ($F = 0.15$, $p = 0.93$).

A final analysis was run for test prompt A and B although there were no significant differences between prompts ($F = 0.91$, $p = 0.94$). Essentially all test versions were shown to be equivalent across all raters. However, there were some significant differences within raters and within tasks. Across the 8 raters, there were significant differences in mean scores within raters, but as can be seen in Figure 1, this was not related to the number of tests the teachers rated ($F = 9.861$, $p < 0.05$). The differences in mean scores is likely more reflective of the different abilities of the students as there were also significant differences between the scores of different classes overall ($F = 11.412$, $p < 0.05$).

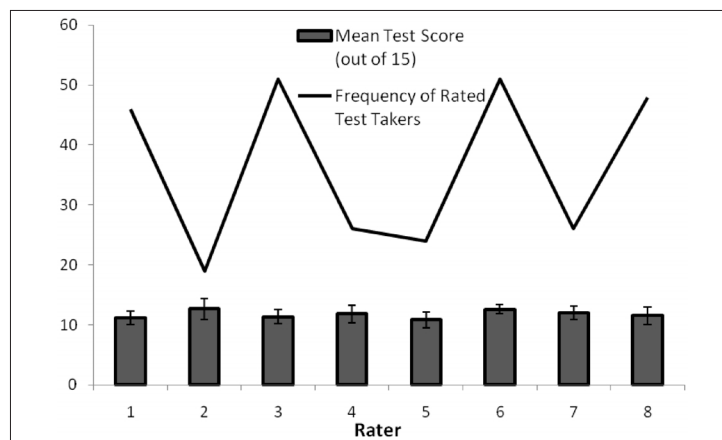


Figure 1

The mean test score (out of 15) and the number of test takers that each rater scored is shown (no statistically significant differences shown between rater and number of tests rated). Y-error bars indicate standard deviation.

Finally, a one-way ANOVA to two degrees of freedom was performed for the independent variable task and the dependent variable task score (each task out of 5 points). Task 1 was rated the highest ($M= 4.074$, $SD= 0.47$), followed by task 2 ($M= 3.876$, $SD= 0.64$) and task 3 ($M= 3.756$, $SD= 0.61$). Significant differences were found ($F= 22.535$, $p < 0.05$), and subsequent t-tests revealed the significant differences existed between all 3 tasks (for task 1 and 2, $t=4.270$, $p < 0.05$; task 1 and 3, $t= 7.055$, $p < 0.05$, and task 2 and 3, $t= 2.320$, $p < 0.05$). There were no significant differences for task score across raters or classes, showing that no class did significantly better on any task in particular.

Discussion

As predicted, there were no significant differences across test forms, demonstrating that they can be considered equivalent versions. This was also the case for the more specific analyses of test versions and prompt version (A or B) illustrating that all test versions should be seen as parallel forms - any differences in student performance were not due to having an advantage because of which test version was being used. This is often an issue in speaking tests, because tests can appear to be equivalent before testing and may even be shown to be parallel overall across forms after testing, but can exhibit differences at the individual task level (Weir & Wu, 2006). This was not the case for the current test as differences in task difficulty occurred across all test forms and for all raters. In fact, the task difficulty analyses revealed unexpected results: not only did each task score a statistically different mean difficulty level but also that the difficulty increased with task number : the students answered the easiest question first and the hardest question last for all versions. This task order has a well-established association with higher performance scores than any other combination of difficulty (McDowd & Craik, 1988). That being said, there were differences across raters. This could be due to either class ability or rater severity. The difference in mean scores between the lowest and the highest class was 18.4%, while the difference in mean scores for each rater was only 12% (as can be seen in Figure 1, the difference between raters overall is small), suggesting that the differences across raters are more likely due to overall class ability. These differences do not impact the argument of equivalency of test forms directly: they have been included to demonstrate that there are differences between mean class scores and that this was uninfluenced by the overall number of pairs scored by each rater, thus supporting the argument of test form equivalency - if all raters and classes had been found to be without significant differences, the lack of significant differences for test form analyses could have been attributed to a lack of differences overall, thus eliminating the evidence of equivalency of test forms.

Essentially, the results described herein illustrate a test with wholly equivalent forms and it is suggested that this could be partly due to the influence of the teacher's survey. At the Bunkyo English Communication Center, in addition to teaching, instructors also develop the curriculum, lesson materials, and rate all student assessment. They are rarely however, involved in test development and involving them in task selection

at the early developmental stages is not something that has been before suggested by any previous publications in the oral testing literature. It is possible that in getting feedback on prompts and rubrics from instructors, they felt they had contributed to the test development, felt familiar with the tasks on the test prior to norming, or were generally comfortable with the tasks and the scoring system (since they had selected it themselves) during implementation of the test. At this stage it is not clear exactly what the precise influence of the survey was but it certainly is worth considering adding to the design stages of oral achievement tests, especially for developers looking to make their tests representative of the curriculum or if nothing else, as a formalization of the test development process.

References

- Bachman, L. & Palmer, A. (2010) *Language Assessment in Practice*. Oxford University Press:USA.
- Gronlund, N. (1998). *Assessment of Student Achievement*. Allyn & Bacon Publishing: MA.
- Fulcher, G. (2004) *Testing Second Language Speaking*. Harlow: Pearson Longman.
- Lindholm-Leary, K. (2001) *Dual language education*. Multilingual Matters: Bristol.
- Luoma, S. (2004) *Assessing Speaking*. New York: Cambridge University Press.
- McDowd, J. & Craik, F. (1988) Effects of aging and task difficulty on divided attention performance. *Journal of Experimental Psychology: Human Perception and Performance*, 14(2): 267-280.
- Weigle, S. (1994) Effects of training on raters of ESL compositions. *Language Testing*, 11(2): 197-223.
- Weir, C. & Wu, J. (2006) Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing*, 23(2): 167-197.

—平成23年11月 2 日 受理—