Evaluating Oral Communication

BURROWS, Christian
Department of Early Childhood Education
Faculty of Education for Future Generations
次世代教育学部乳幼児教育学科
バロウズ・クリスチャン

キーワード：英会話能力，会話を評価すること

Abstract：著者は，学生の英会話技能において一定期間でどれだけ上達したかを確実に測ることを目的として，独自の試験形式を考案し，それを実施，彼らの成績を評価している。
その結果，英会話能力の確実な測定は，なかなかやりがいのある（ある意味，困難な）研究領域にあると考える。
正確に学生の会話能力を測るには，試験の内容が正当性を持ち，確実なものでなければならない。
この論文では，確実な評価につながる試験とはどのようなものなのかを説明している。又，試験の形式が正当であると言い切る理由と，試験内容の考案から作成，実施に至るまでの過程についても言及している。

Ⅰ. Introduction

Of all the language skills learners acquire when studying a foreign language, perhaps the most problematic one to accurately measure is oral proficiency. With factors such as test validity and reliability prominent concerns, such difficulty may help explain why oral testing is the least developed and practiced. Despite the difficulties, in order to meet International Pacific University's goal of measurable improvements in speaking ability for its Eikaiwa course, the author devised, implemented, and assessed an evaluation which aimed to measure student speaking performance. This raised many questions pertaining to how oral proficiency and improvements in linguistic competency are measured. This paper describes the achievement language test which was implemented, and explains the justification for the test, before discussing its validity and reliability. It will also highlight some of the weaknesses in the test format before recommending procedures which could be adopted to help strengthen the test's reliability.

Ⅱ. Justification for the achievement test

Unlike foreign tertiary education, many Japanese universities take student attendance into account when deciding grades. This weighting leads some students to assume that by attending most of the classes a pass is almost assured, regardless of the effort made during the semester. Unfortunately, in order to show real improvement when learning a second language, it is essential that students apply themselves throughout the entire semester. The standard grading structure at the author's university, with attendance consisting 30% of the final grade, appeared to do little more than encourage attendance while reinforcing students' assumptions about the ease of passing the course. The remainder of the grade comprised of two group presentations (25% each) and homework (20%). Accordingly, students who attended most of the classes and read for a few minutes during the presentations could comfortably attain a pass grade, regardless of their English proficiency. Such grading systems help to perpetuate the 'false beginners' label (Wadden, 1993：38) that allows Japanese students to pass English language courses despite poor linguistic

skills. Therefore, the demand for measurable improvements necessitated an assessment which aimed to evaluate students' speaking ability. It was felt that a grading structure which rewarded students for their ability and efforts would be fairer, while at the same time encouraging active participate through the positive benefits of the backwash effect.

The grading structure implemented to achieve this included two oral achievement tests composing 40% of the final grade. These tests attempted to evaluate linguistic ability, not proficiency, in addition to providing the motivation for students to engage in activities during class. As students were being asked to reproduce precisely the forms that the author wished to measure, direct testing appeared to offer the most accurate way of assessing performance. This 'dependable measure' (Hughes, 2003 : 4) allowed the author to determine if the students had acquired the linguistic targets from each class. This type of direct testing appeared to offer the most practical way of measuring students' speaking skills which other direct evaluations may not accurately assess. For example, group presentations allow students to report their findings yet in many cases only requires them to read their answer or learn it by heart. Such indirect evaluation, used to determine linguistic ability, would appear to seriously undermine the validity of any test (Weir, 1990 : 75). In addition, by only testing what the students had encountered in the classroom it could be argued it allowed for a fairer assessment of students' ability. However, the author does recognize that successful performance on the test may not truly indicate successful achievement of course objectives, and that some students may have already possessed the linguistic ability.

Ⅲ. Format of the test

Most first year Japanese university students have little experience of direct English language speaking tests. In order to overcome this unfamiliarity affecting their test performance, the author explained the format of the test and role-played the test situation in the classes prior to the test day. The actual dates of the tests were included in the course outline distributed at the beginning of the semester. Students were also provided with a guideline or 'representative sample' (Hughes, 2003:116) of what answers constitute a 'full' and an 'incomplete' answer, along with their corresponding grades. Despite general acceptance in the literature (Brown, 1987) that students perform best if certain phases exist (e.g. warm up ; getting accustomed ; checking level; wind down), as the author did not directly participate in the test it was impossible to ensure these took place.

One pair of students was taken to an empty classroom where the chairs were arranged to face each other, as shown in figure 1.1.

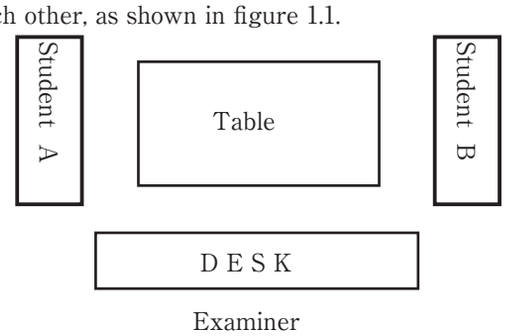

Figure 1.1－Layout of the classroom

After a quick review of the format one student chose a number from 1－12 from the randomly mixed test questions. The number corresponded with the order of the question papers rather than the lesson number, so students were unable to choose a preferred target. The students were informed again that student A first asks the questions from the paper, and then it is the turn of student B. Each student had 10 minutes, which was sufficient enough for them to answer the questions adequately, although it is slightly less than the 15 minutes Hughes recommends (2003 : 124) as the minimum time needed for collecting a representative sample of a test candidate's language ability.

Students questioning each othern (not the author)

allowed the easiest opportunity for them to directly elicit the answers required from their partner, but more importantly lessened any discomfort they might have experienced had the author participated. Although it was probable students felt more comfortable interacting with a classmate or friend, it raised the potential problem if there was a difference in communicative competence. This, along with the prospect of one student dominating were overcome by allocating each student a set time to only ask the questions on the paper, thereby negating any possibility that one student may have spoken more than the other. As the test's contents were contained in the syllabus, with students asking from a set list of questions, to some extent it made it easier to make comparisons with other students' responses. The question papers included several questions along the same theme allowing students opportunities to answer separate 'items' in case they encountered trouble. Also, by testing the teaching targets practiced in the classroom the test was much easier to prepare since the framework within which they operated had already been established. This further helped to reduce any unease and reduce the effect of potential factors (psychological, non-linguistic) affecting performance.

Ⅳ. Grading

By assessing students' ability to answer specified linguistic targets the test allowed the author to compare answers against the scoring criteria and assign a particular mark from 1-10, depending on accuracy, fluency and appropriateness of the answers. Students able to satisfactorily answer the questions received a grade of 5-7 out of 10. Those who gave more detailed, expansive or sophisticated answers were rewarded with a grade of 7-9. This meant the grades were not an accurate attempt to assess in detail, but were merely a tool that provided a general assessment of students' overall performance within the framework of the target. This type of open ended or expressive performance allowed students the choice of linguistic output within the framework of the question. As the test questions were based on a lesson's grammatical target, with a limited number of answers, the author, to some extent, could predict the lexical items but not the structures the student may use. This does not mean that the lessons studied only for the test, rather the relationship was a partnership where testing was supportive of teaching.

Ⅴ. Test validation

Test validity is how well the results of the test reflect the ability it is testing. The four main types are: content validity, construct validity, criterion-related validity, and face validity. Since validity is an argument which cannot be estimated mathematically (Gorsuch, 1997), the difficulty in achieving it may allow less accurate measures to claim acceptable levels, leading some to point out that, 'it may not always be the prime consideration' (Moller in Weir, 1990:33) as low reliability may not adversely affect the overall validity. However, the literature is in agreement that in order to achieve high validity certain practical elements should exist. These include:

1. Appropriate level of difficulty.
2. Adequate discrimination between the different levels of performance.
3. The test assesses the skills/abilities as defined by the objectives of the examination.
4. Assessment of abilities is qualitative rather than quantitative.
5. Tasks are clear so that the examinees understand what is being asked.

(Weir, 1990 : 30 ; Hughes, 2003 : 86-8)

Due to the level of the students (high-beginner) it would have been inappropriate to have used this test to measure their general speaking proficiency; so by restricting the test to the coursework it was hoped that the students would have acquired enough ability to be able to express themselves and speak about various topics. Thus, to be able to express themselves intelligibly, reasonably accurately, and without too much hesitation (Byrnes

in Hughes, 2002 : 67), knowledge of English was less of a consideration than their ability to use English to meet the linguistic demands of the question. Although it is claimed (Weir, 1990) that a clear distinction between performance and competence will always be difficult to maintain:

In testing communicative language ability we are evaluating samples of performance, in certain specific contexts of use, created under particular test constraints for what they can tell us about a candidate's communicative capacity or language ability (Weir, 1990 : 7).

Ⅵ. Construct and content validation

If the test assesses a subject matter (i.e. the lesson targets) about which conclusions are to be drawn, it is said to have content validity (Brown, 1987 : 222). Assuming one wants to measure these targets, the closer the relationship between test and teaching the stronger the construct validity is likely to be. The greater a tests content validity, the more likely it is to be an accurate measure of what it is supposed to be measuring (Hughes, 2003 : 26). As the test questions were fixed and limited to the lesson targets (i.e. no specification of skills that it was meant to cover) it could claim to have some degree of content validity.

Ⅶ. Criterion related

A criterion referenced test is designed to produce a clear description of what an examinee's performance on the test actually means, or determine the extent to which pre-specified program objectives have been met. The degree to which the results agree with dependable assessment of students' ability lets the students know where they stand and where their weaknesses are, while classifying people according to whether or not they are able to perform some tasks satisfactorily. This more subjective grade relates students not to the performance of others but what they can and cannot do. Some criticize the criterion-referenced test in that it is not a comprehensive measure of language mastery, however it is important to remember that no student is likely to attain language mastery as the result of a single course of instruction anyway. And although the test aimed to measure one aspect of students' ability, the author does recognize that it is highly selective and limited in the objectives it measures, also that a 10 minute test cannot give a sufficiently accurate estimate of students' ability. But as a means of measuring specific linguistic functions, which cannot be generalized upon, then it appears to offer a reasonably accurate means of measurement.

Ⅷ. Face validity

Face validity is the extent to which the test 'looks valid' (Weir, 1990 : 26) to those who take it. If the test does not appear to measure what it is supposed to then it will not be accepted as valid by the students which could adversely affect the validity of their responses. It could also be affected by the extent to which the students respond in the manner expected by the author. Some may respond in ways that are counted incorrect, when in fact they knew the correct answers to the question but were misinformed regarding how to respond (Henning, 1987 : 92) . This requires the test to be 'tight' (Weir, 1990 : 38) so that students are clear about which functions they are expected to perform. If these conditions exist (as they did for the test) then high face validity could be claimed but without direct questioning no firm claims can be made.

Ⅸ. Test reliability

Reliability is the extent to which we can depend on the test results to provide similar results on different occasions. That whatever it is that it measures it does so accurately (Davies, 1990 : 6), consistently (Nunan, 1993:119), dependably and fairly (Henning, 1987 : 74). It is dependent on such factors as standard tasks, standard conditions, and standard scoring (Hughes, 2003 : 36). However, despite attempts to make the test conditions scientific, they sometimes yield unreliable results because of

temporary psychological factors or physiological changes beyond the control of the tester. Because these variables seem to have a strong affect on Japanese students, it seems inaccurate to claim that they cannot be ignored without a harmful effect on the validity of the test (Hawking in Weir, 1990：149).

Tests have to ask questions that give us convincing evidence that they accurately and sufficiently measure their particular purpose or objective. Statistical correlation with other observed behavior (correlation coefficient) allows the teacher to quantify the reliability of a test to allow comparison. Oral production tests may be in the 0.70 to 0.79 range (at least 0.9 according to Davies (1990：22)) . Again this discrepancy highlights the difficulty in achieving reliability in the testing of different abilities. However, because tests which assess oral performance cannot automatically claim high standards of reliability it is an area where more reliable measures of communicative abilities are required.

Ⅹ. The influence of backwash

One of the main objectives of the test was to instigate student motivation to practice in English during the lesson. If students are aware they will be tested with virtually identical questions then the backwash is meant to encourage the honing of answers in class. This offers the possibility of exploiting the tests to bring about desired outcomes in the classroom which encourage learning and lead to learners acquiring targeted skills. Weir (1990：13) goes even further and claims that such is the influence of backwash it allows teachers to be less worried about the theoretical aspect of the language as student motivation to complete the functions for the test is paramount.

Ⅺ. The validity and reliability of the implemented test

The author adopted many of the recommendations in literature to improve reliability. Items (the whole semester's syllabus) were included on the test to allow students to be able to answer questions that had been studied, thus establishing a degree of fairness. This was further established by the nature of the test which allowed students the freedom to determine how to answer. The general topic or conversational direction elicited could not exactly be predicted, nor the actual content of students' answers. This type of test therefore appears to parallel the free use of the target language within a real communicative situation. As a result, it could claim to have a higher degree of content and face validity than other techniques (apart from role-play).

The test was very practical in terms of ease of efficiency, administration, scoring, and interpretation of results. The type of questions which were used also directly elicited students' answers to the forms studied in the corresponding lesson, thereby removing any ambiguity and keeping interactions controlled enough to isolate, assess and score. However, as students were rewarded for expansive answers they were not restricted to only produce the targeted forms. The author recognizes that these answers are affected by the test conditions and many students can encounter difficulties due to psychological factors (Hughes, 2002：73). Consequently, students' true abilities may not always be reflected in their test scores, but this is the difficulty of attempting to measure language ability. For these reasons speaking tests are sometimes termed 'unfair', yet if the speaking skill is to be learned there must be further attempts to evaluate it. This is not dissimilar to other test formats, either direct or indirect, where students can also experience difficulties, e.g. students who read and write with difficulty can struggle in written tests. Ultimately these affective factors can only be reduced never eliminated.

Although the main purpose of the test was to let students demonstrate their abilities, it was recognized that the interactions did not allow the author to make fine distinctions between students but merely to group them roughly by ability

levels, thus confirming Hughes' concern that tests should elicit behavior which truly represents the candidate's ability (Hughes, 2003 : 113).

The instructions for the test had been explained and even role-played in the class and familiarized students with the format as well as providing an indication of what type of answers were expected. This was further reinforced with the distribution of a detailed scoring key before the test. These clear and specific directions allowed some degree of consistency and dependability in the scoring despite the fact that the assessment was subjective and impressionistic (Nunan, 1993 : 127). While it is accepted that the score cannot be regarded as an accurate measure in itself, it does allow some distinctions to be made between those who can complete linguistic functions and those who cannot.

Ⅻ. Ways to strengthen the test

As the tasks that the test required students to perform targeted specific linguistic constructs, rather than productive abilities or aspects of general oral proficiency (e.g. turn-taking, back-channeling, topic shifting), it could claim to be reasonably valid in what it measured. However, it is also recognized that the test has limited use for a more general assessment of speaking proficiency, but as a means of assessing students' ability to show they can answer questions it appears to provide a useful and motivating influence. Although there is no empirical evidence to show the effect of the test, the author feels that if a questionnaire were administered it would receive high face validity among the students, although this may be negated because of students' unfamiliarity with the rationale underlying the test. However, that the test lacks any empirical evidence to support it is a concern, therefore it requires further data to support its justifications. This lack of reliable empirical evidence is despite the seriousness of the test which plays a large part in determining if students pass or fail the course.

The author also acknowledges that the grading could be clearer and admits that there is a lack of distinction between students who can answer the questions and those that can answer it well. Although the grading structure was fabricated for this test it does bear some similarities to the grading for certain accepted speaking tests which assess general oral proficiency. Yet despite attempts at objectivity it is reasonable to assume that subconscious factors influenced the scoring. Therefore, if the test used a grading structure that has been validated it may enhance tests accuracy, although due to the nature of the format substantial changes would have to be made. To address this it is proposed that the tests are video recorded and graded at a later time. This would be more time consuming but would offer the possibility to establish scoring validity. These recordings could also allow a second marker to independently grade the interactions to allow comparisons with the author's assessment, thereby establishing inter-rater reliability. This could easily be achieved after explaining the grading structure and providing a video recording of the interactions to the marker.

As the test has not been correlated with a reputable one it lacks a degree of concurrent validity. As a direct comparison would be impracticable, because of the limitations of the test, the author recognizes that without such validation the test will never be able to offer empirical evidence to support its claims. However, if the students' test scores and their final grades were correlated it could offer some justification for the test as well as for a student's final grade. This could allow the author to justify a student's final grade to the university and offer some protection against claims that the test was not an accurate measure of ability. Finally, the tests started almost immediately after the students had entered the room. Due to time considerations the tests did not include a brief period to allow the students to feel at ease. Given more time this initial period would help to ease the students into the test rather than the sudden test format that was used.

Ⅻ. Conclusion

This paper has described an English speaking test and discussed the validity and reliability. While the test cannot make any claims of reliability or validity based on empirical evidence, it does satisfy generally accepted elements that are required for a speaking test. It had a precise purpose of assessing students' ability to appropriately answer questions based on the coursework. It was accepted that these answers had to address the question and show familiarity with the target and related vocabulary.

A major goal of most language programs should be to enable the students to use the new language for communicative purposes meaning students have to acquire the ability to engage in face to face conversation. This should therefore be the basis of evaluating students' linguistic competency. So any doubts that exist about the hidden dangers (Nunan, 1993：127) or weaknesses of oral tests, or the nature of speech, should not be used to replace this valuable and valid means of assessment. While recognizing that accuracy can only be further established by observation and theoretical justification there is no final, absolute, and objective measure of validity so this must serve as a stimulus to provide further research and justification for its use.

References

Brown, H.D. (1987) Principles of Language Learning and Teaching. Englewood Cliffs, NJ: Prentice hall Regents.

Davies, A. (1990) Principles of Language Testing. Oxford: Blackwell.

Gorsuch, G.J. (1997) 'Test Purposes'. The Language Teacher 21.01:1-8.

Henning, G.A. (1987) Guide to Language Testing (Development, Evaluation, Research). Cambridge, Mass: Newbury House.

Hughes, A. (2003) Testing for Language Teachers. Cambridge：Cambridge University Press.

Hughes, R. (2002) Teaching and Researching Speaking. London：Pearson Education Limited.

Madsen, H.S. (1983) Techniques in Testing. New York: Oxford University Press

Nunan, D. (1993) The Learner-Centered Curriculum. Cambridge：Cambridge University Press.

Weir, C.J. (1990) Communicative Language Testing. Hemel Hempstead: Prentice Hall.

Wadden, P. (1993) A Handbook for Teaching English at Japanese Colleges and Universities. Oxford: Oxford University Press.

（平成20年11月27日受理）