

Promoting Positive Backwash in Conversation Classes through Oral Testing

Paul Walsh*

Introduction

Although many people have fond memories of the time spent in their respective education systems, it is unlikely that those of taking tests are among them. In language education in general, and communicative language learning in particular, testing is something that many instructors also do begrudgingly. However, language testing, defined by Hughes as any structured attempt to measure language ability, will be needed “as long as it is thought appropriate for individuals to be given a statement of what they have achieved in a second or foreign language”⁽¹⁾.

For the purposes of this paper I am specifically concerned with issues related to assessment in low-level Japanese university oral communication (conversation) classes, however, a discussion of these has relevance in all oral communication classes.

A great deal has been written on language testing since the 1960s, however, the testing of oral skills compared to other language skills makes up but a small proportion of it. Underhill accounts for this with his assertion that oral tests are, “qualitatively different from other kinds of tests,” and the fact that they “do not easily fit the conventional assumptions about people and testing”. This means that it is not easy to treat tests in the communication classroom in the same way as conventional tests.⁽²⁾

With entire university courses here in Japan devoted to conversation in a

* Research Assistant, Hiroshima University of Economics, Hiroshima, Japan

foreign language it follows that testing methods appropriate to the assessment of oral ability should be employed. Even in institutions where conversation courses are characterized by an almost beginner level, students of which Hughes says should not undergo any formal testing at all – assessment best being carried out informally – some form of formal assessment is generally required⁽³⁾. So if testing is required, the challenge for instructors is to produce a testing program that not only assesses achievement, but also increases motivation and above all creates beneficial backwash⁽⁴⁾.

Conventional assumptions about language testing

Underhill traces prevalent assumptions about language testing back to the beginning of psychometrics and the advent of intelligence testing, from which language teaching adopted the belief in a single factor of general language competence. Although the belief in a single linear scale of language proficiency against which the proficiency of an individual can be measured by means of an objective test instrument has been discarded⁽⁵⁾, “the criteria we use for evaluating tests still favor the statistical assumptions of the mental testing heritage, and the result is a strong bias towards mechanical tests and against the human face of oral tests”⁽⁶⁾.

Approaches to language testing

Lado’s discrete-point approach⁽⁷⁾ outlined in the early 1960s, which used structural contrastive analysis to break language down into small testable segments, met the demand for objective statistical data, but fell out of favor on the grounds that knowledge of these separate segments does not equate to knowledge of the language. As Oller wrote, “the whole is greater than the sum of its parts”⁽⁸⁾, and it was Oller who championed the use of global integrative testing characterized by cloze tests⁽⁹⁾, which require learners to fill missing gaps in a text, and dictation as better ways of measuring the ability of a learner to combine language skills in a way that more closely resembled real life language use in an aim to produce a unitary trait of ‘general language proficiency’. Cloze testing and dictation have been heavily criticized and the unitary trait hypothesis⁽¹⁰⁾ disconfirmed, a fact admitted by Oller himself⁽¹¹⁾.

In response to this, the theory of communicative language competence⁽¹²⁾, which recognized “language use as a dynamic process, involving the assessment of relevant information in context, and a negotiation of meaning on the part of the language user” was developed. Bachman’s model of communicative language competence, building upon that of Canale and Swain, proposed that communicative language competence is comprised of three main components: Language competence (“a set of specific knowledge components that are utilized in communication via language”), strategic competence (“the mental capacity to for implementing the components of language competence in a contextualized communicative language use... providing the means for relating knowledge competencies to features of the context in which language use takes place and to the language user’s knowledge structures”), and psychophysical mechanisms (“the neurological and psychological processes involved in the actual execution of language as a physical phenomenon” mechanisms that take place⁽¹³⁾)

In the words of Canale:

Just as the shift in emphasis from the language form to language use has placed new demands on language teaching, so too has it placed new demands on language testing. Evaluation within a communicative approach must address... new testing formats to encourage creative, open-ended language use, new test administration procedures to emphasize the interpersonal interaction in authentic situations, and new scoring procedures of a manual and judgmental nature.⁽¹⁴⁾

The weakness of the discrete point approach and integrative testing was that they measured a learner’s competence rather than a learner’s performance. Knowledge of discrete elements does not necessarily mean that a learner can apply this knowledge and use it to communicate in a particular situation. Nor did either method require any spontaneous production on the part of the learner with all language input coming from the examiner.⁽¹⁵⁾

The theory of communicative language competence suggested that learners should not only be tested on their knowledge of language, but also on their ability to use it communicatively in a given situation, or as Underhill writes,

“When we test a person’s ability to perform in a foreign language, we want to know how well they can communicate with other people, not with an artificially-constructed object called a language test”⁽¹⁶⁾.

Backwash

Backwash is defined by Hughes as, “the effect of testing on teaching and learning”⁽¹⁷⁾. Backwash can be either beneficial or negative depending upon whether the testing methods employed encourage or discourage the learning behaviors the teacher wishes to encourage. A major key to achieving beneficial backwash is, as most students want to do well on their tests, to test the abilities that the teacher wishes to encourage. The key to achieving beneficial backwash is to test the skills whose development one wishes to encourage. As Hughes points out this is a simple matter of content validity, as “the greater a test’s content validity,⁽¹⁸⁾ the more likely it is to be an accurate measure of what it is supposed to measure.” Despite this, “too often the content of tests is determined by what is *easy* to test rather than what is *important* to test”⁽¹⁹⁾. Related to this is the employment of direct testing – that is when the test requires that the learner perform precisely the skill which we wish to measure – using tasks that are as authentic as possible⁽²⁰⁾.

Direct testing of the productive skills of speaking and writing, the acts themselves providing information about a learner’s ability, is less problematic than that of the skills of listening and reading. In spite of this it is the testing of these skills that are seen as most problematic, and concerns about reliability⁽²¹⁾ (as well as the issue of practicality)⁽²²⁾ often result in attempts to assess oral ability by indirect testing methods, which seek to test the factors that underlie the skills that we wish to assess⁽²³⁾.

Further considerations in promoting positive backwash are employing criterion-referenced tests as opposed to norm-referenced testing (tests that tell us what a learner can actually do in the language rather than how they rank compared with the other learners who take the test), and basing achievement tests on course objectives rather than detailed teaching and textbook content. Criterion-referenced tests, “set standards meaningful in terms of what people can *do*, which do not change with different groups of candidates; and they

motivate students attain those standards⁽²⁶⁾". Basing tests on course objectives rather than content provides a more accurate evaluation of both the learning and teaching, with the result that there will be heightened pressure on not only the learner, but also on the teacher in creating effective courses.

Even if all of these criteria are met, the testing method's potential for creating positive backwash will not be fully realized unless the learners (and the teachers) fully understand what the test entails, especially if a new testing method is being introduced.⁽²⁷⁾

Problems with oral testing

Some of the main objections to oral testing concern their reliability and accountability. Oral tests generally require subjective judgment by the assessor of the test subject, and leave no paper trail in the event of challenges to the grades assigned.

Hughes points out that the distinction between objective testing and subjective testing "is between methods of *scoring*, and nothing else"⁽²⁸⁾. In tests that require no judgment on the part of the scorer, the scoring is objective, and conversely if judgment is called for the scoring is known as subjective. Although objectivity in scoring is often seen as primary goal in the construction of many language tests, Underhill writes that although he, "recognises that oral tests, because they involve a subjective judgment by one of another, are likely to be less reliable; but [he] suggests that the human aspect of that judgment is precisely what makes them valuable and desirable"⁽²⁹⁾.

The validity of a test is a measure of how well it measures what it purports to measure.⁽³⁰⁾ It's reliability is the degree to which the test scores "obtained on a particular occasion are likely to be *very similar* to those which would have been obtained if it had been administered to the same students with the same ability, but at a different time"⁽³²⁾. Lado states that an unreliable test cannot be valid, "for an unreliable test does not measure". He characterizes oral tests and written essays – tests of productive skills that have "obvious face validity"⁽³³⁾ – as having low reliability (he quantifies the reliability coefficient of oral tests as between 0.70 and 0.79). He goes on to outline three major factors that can influence a test's reliability; time and circumstances, limitations and

imperfections in the test, and scorer fluctuation (which he maintains can be a major factor of the unreliability of production tests.⁽³⁴⁾ While Hughes admits that, “the scoring of oral ability is generally highly subjective”⁽³⁵⁾, he writes that views about subjective language testing have undergone something of a change. While perfect scorer reliability coefficients of 1 that can be achieved by objective testing methods are unobtainable in subjective tests, there are ways of achieving results high enough for the test results to be of value.⁽³⁶⁾

Hughes offers guidelines to help improve reliability of productive tests in terms of both, test-retest reliability and inter-rater reliability. These include taking a sufficient number of samples of learner behavior during the test, with each item being independent and representing a ‘fresh start’ (in the case of an oral interview test he recommends that there be as many ‘fresh starts’ as possible); avoiding giving in to pressures to make the test shorter than is appropriate (he writes that as a general rule, the more important the decisions made based upon the results of a test, the longer it should be.); limiting the freedom of learners in the test to minimize difference in performance over time and to allow direct comparisons of learners; providing clear and explicit instructions to tests with which the learners are familiar with.

In contrast to Lado, Hughes goes on to make the point that one has to take care that tests do not lose their validity, as a result of restricting the scope of what candidates are expected to produce in a test. He proposes that how these two considerations are balanced will depend upon what we are trying to measure in the test – restricting learners in ways that do not compromise this, and also in part on the importance attached to the outcome of the test.⁽³⁷⁾

Underhill outlines four questions that a teacher should consider before administering a test or implementing a testing program, to help guide the teacher in designing the most appropriate test.

1. What is the purpose of the test?

To avoid the pitfall of testing merely because it is accepted or expected practice to give language tests (Proficiency, Placement, Diagnosis, Achievement, A combination of 2 or more of the above)

2. What resources (people, time and equipment and facilities) are available?

3. What does the individual learner stand to gain-or lose – from taking the test?

Do the aims of the institution and the needs of the learner coincide?

4. How learner expectations may affect the outcome of the test?⁽³⁸⁾

Underhill recommends that wherever possible a test should be designed to match the local educational philosophy, as far as this is consistent with the aims of the program. However, if learners come from a traditional foreign language learning background such as in the case of the majority of Japanese university students, there is likely to be a fundamental discrepancy between the cultural expectations of the learners and the objectives of the teaching and oral based testing system. He makes the point that oral tests are particularly sensitive to this, as the personality of the learner is more likely to come out in an oral test than in a written test. “The test is likely to reflect the degree of familiarity with the culture with which the objectives are associated rather than just oral proficiency”⁽³⁹⁾. However, if we accept that after having spent many years studying *about* English, students in Japan need to practice how to use it in ‘real time’ (to use Evans’ analogy, in much the same way as one improves any skill, such as a sport or a musical instrument)⁽⁴⁰⁾ we have also to accept that such a discrepancy is almost inevitable. Any such clash of expectations can be eased, by making learners fully aware of the rationale of the test and ensuring that they fully understand how it works. These are in themselves prerequisites for realizing the maximum potential of positive backwash.

Oral tests can be recorded either on audio tape or video tape which provides a record which can be referred to in the event of a challenge to an assigned score, or if two assessors cannot agree on the score and call in a third assessor to make a judgment. Recording tests can also free the assessor from having to mark the performance of the learner at the time of the test (time allowing). This removes a source of stress on the part of the learner, whose nervousness may be increased by seeing the assessor taking notes whenever s/he says something. On the other hand, the very fact that the learner is aware that his/her performance is being recorded may raise stress levels considerably and also compromise the authenticity of the test. Performances on oral tests seem

to be particularly vulnerable to stress on the part of the learner. Underhill recommends against recording tests if at all possible, and proposes that in an effort to demystify the test experience if at all possible it should be carried out somewhere like an ordinary classroom, or even corridor, rather than in a specially prepared and “profoundly silent” examination room.⁽⁴¹⁾

In practice

So, how does all this actually work in the classroom here at Hiroshima University of University (HUE)? Can an oral testing system provide a valid and acceptably reliable achievement test,⁽⁴²⁾ and does it produce a positive backwash effect in practice? As was said at the beginning of this paper, formal testing of communicative ability is not recommended for the kind of beginner level students that make up the majority of the students taking the beginner level conversation classes at HUE. This low level, however, does make it possible to make the kind of judgments about levels of achievement in the short amount of time available for testing that would be impossible at higher levels.

The most common form of oral test is the interview, “a direct face-to-face exchange between learner and interviewer”⁽⁴³⁾. In my classes, however, I have foregone this technique in favor of monitoring and assessing interaction two learners. The reasons for this are that in an interview, much of the initiative is taken by the interviewer, while the focus of my course is to develop the ability of learners to initiate, control and prolong conversation. “The art of conversation” as it were, relies as much on a learner’s strategic competence as on their language competence. Also, the relationship between tester and learner is an unequal one (usually the learner is speaking to a superior) which can further impair initiative.⁽⁴⁴⁾ In addition, as throughout the course the learners are encouraged to speak to as many of their peers as possible (something that has often proved problematic in my experience), therefore it follows that peer interaction in English is what they should be called upon to accomplish in the test. Peer to peer interaction also negates the issue of interviewer fatigue.

So far, the shift in emphasis to oral testing seems to have had a marked impact on the willingness of students to practice the act of speaking in real time in class, so it seems that the backwash effect has been achieved.

The short time allotted to each learner for assessment, however, does present some problems. In a class of 20 learners in a ninety-minute time slot each learner is assigned only 3 minutes. This adds to the stress levels of the learners, knowing as they do that they have to perform in a very short period of time, and there is little time for them to settle into the test and relax. In smaller classes with more available time I have found that a five-minute time slot for each student is far more effective. Students given more time have more chances to make 'fresh starts' and a wider sample of behavior can be observed. In this type of testing great care must also be taken to ensure that students are as carefully matched as possible.

The short time slot has also resulted in more rote learning of questions and responses than I had hoped. This is of great personal concern as the objective of the course is to develop 'conversational' ability by teaching the use of follow up questions (displaying listening ability and interest in what the partner is saying), and encourage appreciation of whether a conversation partner is understanding what is being said and using communication strategies such as clarification, rather than just knowledge of elements of language. The shortcomings of the testing method expose an area that may have been neglected in my teaching. By making it clear to the learners that this will be a criterion against which they will be assessed, makes them not only more likely to focus on it, but also makes me focus on it more in my teaching.

Conclusions

Overall, oral testing is not easy, and the achievement of valid and reliable results takes considerable time and effort. "Nevertheless," Hughes writes, "where backwash is an important consideration, the investment of such time and effort may be considered necessary."⁽⁴⁵⁾

In fact, assuming that appropriate oral tests can be constructed, the central role of testing in language education in Japan actually makes the potential for producing positive backwash in conversation classes great. The more learners are concerned about the outcome of tests they undertake, the more they are willing to do what it takes to do well on them. In conclusion, make speaking in real time the center of a testing system, and students should be keen to speak

in class.

Notes

- (1) Hughes, A. *Testing for Language Teachers*. Cambridge, Cambridge University Press, 1989, p. 4.
- (2) Underhill, N. *Testing Spoken Language*. Cambridge, Cambridge University Press, 1987, p. 3.
- (3) Hughes, op. cit., p. 101.
- (4) 「波及効果」テストが学習者の学習に及ぼす影響のこと。現実のコミュニケーション場面を設定した「話すこと」のテストは、学習者に準備の段階で実際に話すことを学習するように仕向けることになり、望ましい backwash effect があると言える。若林・根岸「無責任なテストが「おちこぼれ」を作る」大修館書店, 1993, p. 159.
- (5) 「客観テスト」採点に際して主観的な判断を必要としないテスト。
Ibid., p. 166.
- (6) Underhill, op. cit., pp. 4-5.
- (7) 「観点別テスト」測定の見点を特定の言語項目や言語使用だけに絞って見ようとするテスト。若林・根岸, op. cit., p. 165-166.
- (8) Oller, J. W. *Language tests at school: a pragmatic approach*. London, Longman, 1979, p. 212.
- (9) 「穴埋めテスト」あるテキストの文章にあなを空けて、受験者は前後の分派からその穴を埋めるだけである。若林・根岸, op. cit., p. 165.
- (10) Bachman, L. *Fundamental Considerations in Language Testing*. New York, Oxford University Press, 1999.
- (11) Oller, J. W. (ed) *Issues in language testing research*. Rowley, Mass, Newbury House, 1983 cited in Bachman, 1999.
- (12) 「コミュニケーション能力・伝達能力」言語を性格に理解し、実際の状況の中で適切に使用する能力。文法的能力 (grammatical/language competence)・社会言語的能力 (sociolinguistic competence)・方略的能力 (strategic competence) などに下位分類されています。文法的能力だけでコミュニケーションが適切に行われることはない。白畑・富田・村野・若林「英語教育用語辞典」大修館書店, 1999, p. 64.
- (13) Bachman, op. cit., p. 84.
- (14) Canale, M. "Testing in a communicative approach". In Gilbert A. Jarvis (ed). *The Challenge for Excellence in Foreign Language Education*, Middlebury, The Northeast Conference Organization, 1984. cited in Bachman, 1999.
- (15) Morrow, K. "Communicative language testing: revolution or evolution?" In Brumfit and Johnson *The Communicative approach to language teaching*, Oxford, Oxford University Press, 1979 and Weir, C. J. *Communicative Language Testing*, Englewood Cliffs, NJ, Prentice-Hall Regent, 1990.
- (16) Underhill, op. cit. p. 5.
- (17) Hughes, op. cit. p. 1.
- (18) 「内容（的）妥当性」テストの内容とする妥当性 (validity)。測ろうとする対象を適切な方法および適切な配分で測定する度合のこと。例えば、コミュニケーション能力を測定

しようとする際に、文法能力テストだけを行って、他の社会言語能力や、談話的能力、方略的を測定しないならば、このテストの内容的妥当性は低くなると考えられている。白畑・富田・村野・若林, op. cit., p. 75.

- (19) Hughes, op. cit. pp. 22–23.
- (20) 「現実性」テストが、どの程度現実の言語活動の実態を反映しているか、その程度のこと。若林・根岸, op. cit., p. 159.
- (21) 「信頼性」テストの結果が、受験者の学力等をどの程度正確に表現しているか、その程度のこと。Inter-rater reliability 「採点者間信頼性」複数の採点者が同一の反応（答案）にたいしてどの程度同一の得点を与えているかである。Subjective test 「主観的テスト」では、しばしば、同一の答案に対して、採点者によって異なる点数を与えることが多い。最移転の基準をあらかじめかなり厳密に用意しておく必要がある。Test-retest reliability 「再テスト信頼性」テストをある期間（学習者の能力に変化が起こらない期間）において実施した場合に、どの程度同じ得点となるかを見る。この2回のテストで、得点やその文布が著しく異なる場合は、そのテストの信頼性は低いものとしなければならぬ。若林・根岸, op. cit., p. 163–4.
- (22) 「実用性」テストの実用性。理論的にどのように優れたテストであっても、この実用性がなければテストは広まらない。ただ、逆に実用性があるというだけで、妥当性 (validity) のないテスト。
白畑・富田・村野・若林, op. cit., p. 75.
- (23) Hughes, op. cit. p. 15.
- (24) 「到達基準拠テスト」ある特定の行動目標に到達できたかどうかを測定しようとするテスト。このテストでは受験者の順位づけでなく、その行動目標を遂行する能力があるかどうかの問題となる objectives-referenced test 「目的基準拠テストとも呼ばれる」。白畑・富田・村野・若林, op. cit., p. 165.
- (25) 「集団標準拠テスト」個人的の能力を集団の中での位置により定義しようとするテスト。偏差値や5段階評価も、この集団標準拠テストから得られる。白畑・富田・村野・若林, op. cit., p. 166.
- (26) Hughes, op. cit. p. 18.
- (27) *ibid.*, p. 46.
- (28) *ibid.*, p. 19.
- (29) Underhill, op. cit. p. 5.
- (30) 「妥当性」テストのテストの「妥当性」とは、測定すべき能力・技能等のものを測定しているか、ということである。たとえば「会話」の能力を測定するのに、「紙と鉛筆」によって会話文を完成させても。受験者に会話能力があるかどうかを知ることはできない。このテストは妥当性がない。若林・根岸, op. cit., p. 168.
- (31) Lado, R. *Language Testing*. London, Longman, 1961, p.330.
- (32) Hughes, op. cit. p. 29.
- (33) 「面妥当性」あるテストを見たとき、測定しようとしている能力を実際に測っているようにみえること。表面妥当性ではいと受験者の望ましい参加を得ることが困難である。若林・根岸, op. cit., p. 168.
- (34) Lado, op. cit., pp. 330–2.
- (35) Hughes, op. cit. p. 114.
- (36) *ibid.*, p. 36.

- (37) *ibid.*, pp. 36–41.
- (38) Underhill, op. cit. pp. 11–21.
- (39) *ibid.*, p. 20.
- (40) Evans, D. “Why do oral testing?” In *The ETJ Journal* 4(2), 2003.
- (41) Underhill, op. cit. p. 41.
- (42) 「到達度テスト」ある特定のコースにおいて学習されるべき事柄をどのぐらい身につけたかを測定するテスト。若林・根岸, op. cit., p. 165.
- (43) Underhill, op. cit. p. 54.
- (44) Kormos, J. “Stimulating conversations in oral-proficiency assessment: a conversation analysis of role-plays and non-scripted interviews in language exams”. In *Language Testing* 16(2) pp. 163–188.
- (45) Hughes, op. cit. p. 114.

Bibliography

- Bachman, L. *Fundamental Considerations in Language Testing*. New York, Oxford University Press, 1999.
- Canale, M. “Testing in a communicative approach”. In Gilbert A. Jarvis (ed). *The Challenge for Excellence in Foreign Language Education*, Middlebury, The Northeast Conference Organization, 1984.
- Evans, D. “Why do oral testing?” In *The ETJ Journal* 4(2), 2003.
- Hughes, A. *Testing for Language Teachers*. Cambridge, Cambridge University Press, 1989.
- Kormos, J. “Stimulating conversations in oral-proficiency assessment: a conversation analysis of role-plays and non-scripted interviews in language exams”. In *Language Testing* 16(2) pp. 163–188.
- Lado, R. *Language Testing*. London, Longman, 1961.
- Morrow, K. “Communicative language testing: revolution or evolution?” In Brumfit and Johnson *The Communicative approach to language teaching*, Oxford, Oxford University Press, 1979.
- Oller, J. W. *Language tests at school: a pragmatic approach*. London, Longman, 1979.
- Oller, J. W. (ed) *Issues in language testing research*. Rowley, Mass, Newbury House, 1983.
- 白畑・富田・村野・若林『英語教育用語辞典』大修館書店, 1999.
- Underhill, N. *Testing Spoken Language*. Cambridge, Cambridge University Press, 1987.
- 若林・根岸『無責任なテストが「おちこぼれ」を作る』大修館書店, 1993.
- Weir, C. J. *Communicative Language Testing*, Englewood Cliffs, NJ, Prentice-Hall Regent, 1990.