鳴門教育大学小学校英語教育センター紀要　創刊号, 21−30, 2010

Lessons From a Communicative Test for Young Learners

Gerard Marchesseau
Naruto University of Education

Abstract

This research presents the results of an Oral Proficiency Interview (OPI), conducted with fifth grade elementary school students. Following the test, learners completed a short questionnaire assessing affective factors, which correlated to test performance. The questionnaire also revealed that students had very positive feelings about the testing process, a finding which was corroborated by comments from homeroom teachers and general observations. Each learner was also rated by four individual raters, allowing us to investigate rater reliability. The assessment process illuminated challenges for communicative testing relating to rater reliability and test practicality, and benefits of communicative testing such as the potential for positive washback effect and improved test validity.

(Keywords: young learners, communicative testing, foreign language activities)

1. Introduction

1.1 Literature on testing and assessment

The Ministry of Education, Culture, Sports, Science and Technology (MEXT) is currently in the process of implementing English as a ‘subject area’ under the name “Foreign Language Activities” in elementary schools across Japan for fifth and sixth grade students. Examining the New Course of Study for “Foreign Language Activities”, it is proposed that the goals and means are closely aligned with current thinking on testing and evaluation, specifically, testing and evaluation of young learners as well as broader trends in language teaching.

Consistent with the recent trend towards communicative language teaching, language testers have called for increasing communicative assessment. Particular stress has been placed on directness (Underhill, 1987; Hughes, 1989; Weir, 1990).

That is, the construct which tests seek to measure should be elicited directly by the test. In order to accurately assess test-takers' ability to *speak* English, for example, learners should be required to produce spoken output. By contrast, traditional tests which isolate discrete abilities, such as multiple choice tests of grammatical knowledge may be quite poor at predicting learners' future communicative performance in a real situation. By extension, Underhill (1987), Weir (1990) and Hughes (1989) have also called for authenticity in test situations. Test tasks should be similar to tasks which learners will actually be required to negotiate in the real world; again, because this is seen as being the best way to predict such future performance. We see this thinking reflecting in real testing situations; not just in low-stakes tests but in some very high-stakes tests such as the National Center Test for University Admissions in Japan, which implemented a listening component a number of years ago in effort to make the test more communicative. The TEOFL test also completely eliminated the old grammar-based multiple-choice component and now requires considerable test-taker output, both written and spoken.

When thinking of communicative assessment, Oral Proficiency Interview (OPI) tests likely spring to mind. OPIs can indeed be an effective way of eliciting output to predict future performance. An OPI would clearly seem to be a *communicative* test, yet teachers and testers should still be cautioned against viewing the OPI in narrowly-defined terms. Consider an OPI in which the test-taker asks a serious of questions which are answered in sequence. We must ask; "Does this really reflect communication in the real world?" This type of test may be appropriate in some situations, but it posits the test-taker in a passive role and leaves little room for any real negotiation of meaning. The test-taker is not free to initiate topics or exert any control over the interaction. In this way, the test is far more reflective of traditional classroom culture than a communicative act in the real world. In order to more accurately simulate real communication, Nunn (2000) suggests testing small groups of test-takers and allowing them to control (in large part at least) the process. In this way, the conversation that results reflects the unpredictable nature and joint-construction of real conversation (Bygate, 1987; Tsui, 1994).

Similar themes emerge in the literature concerning testing for young learners. Pinter (2006) and Cameron (2001), for instance, emphasize the role of interaction between the testers and test-takers and caution against divorcing the assessment process from the classroom experience or real-world interaction. Children learn in class by interacting and negotiating with each other and with teachers as they complete a range of games, activities and tasks. It might only

cause confusion or frustration if we then remove the child from this environment, put them in the metaphoric test-takers chair, and tell them to perform. We certainly cannot expect to elicit the child's best performance under these conditions. Testers should cultivate the test-taker's best performance, providing scaffolding and support similar to what we see in the classroom or in real-life interaction with children. As such, assessment should be congruent with learning.

It must also be noted that although this research deals specifically with a *testing* procedure, assessment of young learners should not be limited to testing. McKay (2006) echoes other research, calling for multiple measures of assessment. These should include self-assessment, portfolio assessment, systematic observation of classroom interaction, OPI-based tests as well as more traditional tests, depending on student needs, local conditions and other factors. Young learners' performance on tests is likely to vary even more than adult learners, who are well-familiar with various kinds of tests and have developed strategies to deliver their best performance. Furthermore, young learners might be especially vulnerable to de-motivation. Assessing young learners is a much more delicate and complicated process than mere tests allow. Consider the flower/bud metaphor which has been often equated with second language learning: Second language learning is not a step by step mechanical process, but rather a complex, organic process, more similar to the development of a flower. This metaphor can also be extended to testing. If we rip a flower out of the ground and merely measure its height, we are not going to get a full or accurate picture of its development and we are also likely to stunt its growth (akin to the notion of negative washback). Getting a more accurate picture of development requires great consideration and a range of assessment techniques.

1.2 The New Course of Study

At the present time "Foreign Language Activities" is not slated to be a full subject and thus, there are no requirements or specific recommendations for systematic testing or assessment. Nonetheless, looking at the broad goals and approaches outlined in the New Course of Study (MEXT, 2009), we find a great deal of common ground with the previously sited literature. Like the course of education for lower and upper secondary school, the focus is clearly on fostering pupils' communicative ability and positive attitude towards English (and other cultures more generally). Even at a cursory glance, the emphasis on communication is duly stressed, with the word "communication" or other derivations of "communicate" occurring 22 times within the roughly 3 page document. This is, of course, consistent with broader trends towards

communicative language teaching and testing. Language is not approached as a set of skill-based, atomistic sets of knowledge, but rather as a holistic, integrated, communicative tool. The teacher's role, within "Foreign Language Activities" is to be a facilitator of experiential learning, rather than to impart knowledge in a traditional teacher-centered manner. Learning, as such, should be implicit and direct; 'learning by doing' in essence, with minimal explanation.

Although assessment measures have yet to be specified, the overall goals of "Foreign Language Activities" are entirely consistent with the current view toward assessment of young learners outlined above. Given the holistic nature of "Foreign Language Activities", performance-based assessment would be much more appropriate than testing students' *skills* via paper tests. Moreover, given the broad goals expressed by MEXT (2009), we would expect assessment to include multiple measures, including student portfolios, systematic observation, self-assessment and perhaps variations of the Oral Proficiency Interview. While acknowledging the importance of various assessment techniques, this research focuses specifically on an OPI and its affect on young learners.

2. The test

The interview test was administered in January and February of 2010 to 74 fifth grade students. The learners completed three test tasks in pairs, with one rater/test administrator. In the first task, the rater asked students a series of general questions, requiring students to give basic personal information, express their likes and dislikes, and so on. The rater led the task and alternated questions between the two test-takers. In the second task, learners were given a picture of a dish (such as curry, stew, and a sandwich) and required to give instructions to make the dish. Again, the tester alternated between learners, presenting 2 or 3 dishes in total, but learners were also encouraged to work together, depending on their ability. In the third task, learners were given a map of the area around the school and asked to give directions to specific locations in the area, again working individually, but also collaborating at times. Similar tasks had been done in students' classroom work, previous to taking the test.

Testing two learners at a time required the raters to be more conscientious in eliciting equal output and maintaining a systematic approach so that performance can be objectively assessed. While this represents a challenge, the benefits of testing two learners at a time outweigh the disadvantages: Allowing students to interact alleviates anxiety; it more closely resembles classroom tasks; and small group interaction more accurately reflects the unpredictable nature and

joint-construction of real conversation.

All interviews were video recorded and subsequently rated by four individual raters. Rating scales were adapted from Nunn (2000), and Nunn and Lingley (2004) in discussion with the raters, for our specific purposes. Scores were averaged and presented on a scale from zero to six. Finally, students were given a short questionnaire to assess affective factors, the contents of which can be found in the results section.

Furthermore, qualitative notes and observations were made concerning test practicality and the attitudes of other stakeholders (the homeroom teachers at the school and the raters). This test design allowed us to investigate learners' affective factors (such as their attitude toward the test) and the washback effect of the test. It also allows limited investigation into rater reliability and illuminates several challenges for communicative testing.

3. Results

The distribution of total scores is presented in Figure 1.

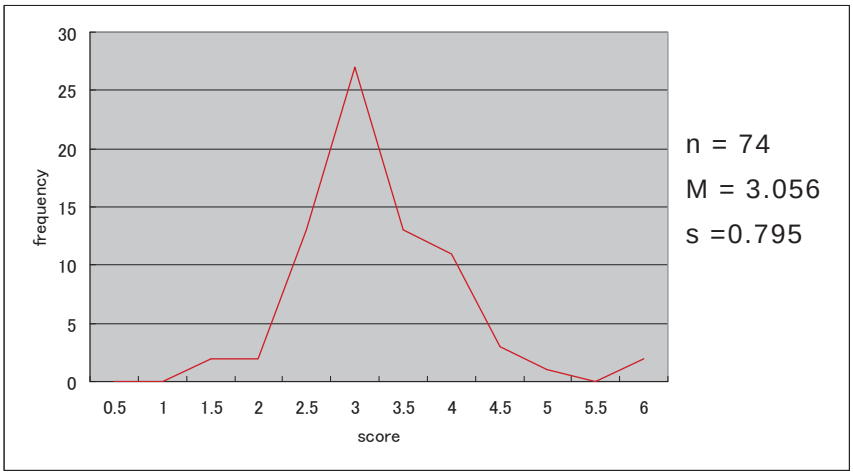

Figure 1. Distribution of total scores

Scores for each of the three rating scales (individual scores for pronunciation, communicative effectiveness and accuracy) closely resembled one another as well as the total score. Thus, no irregularities were found and the data is not presented here.

The questionnaire data is presented in Table 1 below. Learners' test performance correlated positively with each of the affective factors indicated in the questionnaire. Students who answered that they liked English "very much" (*n*=37) performed better than those who responded that they "kind of" like it or

don't like it; Students who enjoyed the test "very much" performed better than the remaining population (t=3.13, sig. at p<.05); students who said they were good at English or "kind of" good at English performed better than those who said they were not good at English (t=3.49, sig. at p<.05); and students who had studied English at "juku" performed better than those who had not (t=3.95, sig. at p<.05). We find no surprises here, of course. With a larger sample size, a full factor analysis would be possible, which is a possible area for further research.

Table 1: *Questionnaire results*

Do you like English?	Frequency	Are you good at English?	Frequency
A/ Yes, very much (とても好き)	37	A/ Yes, I am (得意)	4
B/ Yes, 'kind of' (やや好き)	32	B/ 'Kind of'' (やや得意)	34
C/ No （好きではない）	5	C/ Not really. （あまり得意ではない）	31
D/ No (全然好きではない)	0	D/ Not at all. (苦手)	5

Did you enjoy this test?		Do you study English at juku?
A/ I enjoyed it very much (とても楽しかった)	47	Yes: 43
		No: 31
B/ I 'kind of' enjoyed it (やや楽しかった)	26	
C/ Not really (やや楽しくなかった)	1	
D/ Not at all (全然楽しくなかった)	0	

Pearson's r was used as a measure of rater reliability. This measure was deemed sufficient for the present study. However, a more thorough Rasch analysis would be more appropriate for further study wherein the sole focus is on rater reliability. Scores for each pair of raters were assessed for correlation and the results are presented in Table 2 below.

Table 2: *Rater correlation*

Rater pairs	Pearson's r
A-B	0.6202
A-C	0.6265
A-D	0.7796
B-C	0.6730
B-D	0.7122
C-D	0.7370

4. Discussion

4.1 Advantages of communicative tests as seen in this procedure

Communicative tests have strength in their validity compared with traditional knowledge-based tests (Hughes, 1989). The best way to predict future performance in the 'real world' (assuming that is the goal of tests) is to elicit similar performance, or output in the test. Such tests are called "direct" since they elicit the very performance that they seek to predict, measuring the construct in a direct way. As is discussed below, direct tests tend to sacrifice reliability since it is harder to rate the output. This is of greater concern in high-stakes tests. However, even some large high-stakes tests such as the TOEFL test have been redesigned to directly elicit learner output, reflecting the trend towards a more communicative approach, not just language teaching, but also testing. Certainly, if the aim of a program is specifically to foster communicative ability as it is with "Foreign Language Activities" in elementary schools, one would expect assessment to be in line with this goal. This test procedure sought to recreate the type of situations in which students might actually find themselves at some point in the future.

Communicative tests have also a high potential for positive washback in several ways:

1. Tests can inform and influence teaching: Teachers can be expected to give more communicative lessons when tests are also communicative, requiring learner output.
2. Tests can influence studying: Where tests elicit communicative output, we can expect students to participate and communicate more actively both in and out of class to prepare for the test.
3. The process of taking the test may be a learning experience in itself. In large classes where there is little opportunity for one-on-one discourse with teachers, being required to speak directly with raters (or teachers, as the case may be) under the constraints and pressures of a testing situation might improve students' practical ability.
4. If the process of taking the test is positive, this may increase students' confidence and motivation following the test.

It is impossible to assess the effect that this test had on teaching and studying (the first and second points above) since the present researcher was overseeing the curriculum and deliberately 'taught to the test' at times. However, we can definitively say that the test had a positive washback effect with respect to the third and fourth points. The results indicated that students enjoyed taking the

test in this study, with 47/74 students reporting that they enjoyed it “very much”, and 26/47 students reporting that they enjoyed it somewhat (“kind of”). Only one student seemed to not enjoy the test. This result is very positive, especially in the context of “Foreign Language Activities” at elementary school, where motivation and enjoyment figure highly into the overall program. Homeroom teachers at the school also indicated that the test-taking process is a learning experience in itself, based on their observation and participation in the process, and from talking with students after the test. This was affirmed in informal but systematic discussions with the teachers.

4.2 Challenges for communicative tests

This research illuminates several advantages and disadvantages of communicative testing. One important challenge for communicative tests is to ensure sufficient reliability. OPIs require test-takers to produce output, which must be interpreted by raters and then scored with reference to pre-determined criteria or rating scales. Because this is a subjective process, raters might not always agree with each other (leading to inter-rater reliability issues); or individual raters might not be consistent in their application of the rating scales (leading to intra-rater reliability issues). Herein, the values for Pearson’s *r* indicate a high correlation, but are poor as a measure of inter-rater reliability. Since this was not a high-stakes test, and since the focus was not specifically on rater-reliability, this was not a major concern. Establishing positive washback and striving for higher test validity took priority. Nonetheless, several measures are recommended to improve test validity:

1. It is essential that raters be accustomed to the rating scales. This comes from training and experience. In this study, the raters received a only one to two hour training session on how to use the rating scales and practiced watching video-recorded examples of students who had taken the test previously. Further training should lead to improved rater-reliability.
2. The rating scales must be appropriate and easy to use. The scales used herein had been adapted from previous scales for our specific context. Designing rating scales is a perpetual challenge. Rating scales which are too general can be difficult to apply to actual language use. Rating scales which are too specific (such as scales which predict many possible answers), however, can be very difficult and bulky to actually use. Establishing a balance is important, and testing the rating scales out and revising them before use is more important. The latter step was not taken in the present study.
3. OPIs are not all created equal. Validity can be greatly improved if test-takers

output is convergent, that is to say, if certain answers can be deemed correct. If, for example, test-takers are shown a picture of people doing things and asked to describe what the “girl wearing the red shirt is doing”, raters can easily assess whether test-takers answer correctly or not. If test-takers are asked what they did last weekend, however, rating becomes much more subjective. Note that the former test task is not something that learners would often do in the real world, whereas talking about their weekend is a much more common communicative act. While the latter might be harder to rate, it more closely resembles reality and may thus be considered a more valid test item.

Another issue for communicative tests or OPIs specifically, is practicality. Test-takers must take tests in small groups or individually, requiring a great deal of time. Raters also require considerable training on how to use the rating scales, as mentioned above. In the current study, each test-taker was rated by four individual raters, which yielded an average score which approaches the learners’ true score or *z* score. This was very labor intensive and would not be practical in many situations.

5. Conclusion

Communicative tests and OPIs are recommended as an important component of broader learner assessment in the context of “Foreign Language Activities” at elementary schools. There are clear obstacles to communicative testing in general such as rater reliability and practicality issues, but the benefits outweigh the disadvantages, particularly in the low-stakes environment in which we find ourselves at the moment, with “Foreign Language Activities” being introduced as a ‘subject area’ rather than a full subject requiring systematic and regulated testing. Communicative tests can clearly have a positive impact on students’ attitude and motivation. One would also expect such direct assessment to have a positive influence on the teaching side as well. Equally importantly, communicative assessment is consistent with the goals of “Foreign Language Activities” at elementary schools. Broadly speaking, if our stated goal is to foster students’ communicative ability, assessment (as well as material and teaching) should be in line with that goal.

References

Bygate, M. (1987). *Speaking*. Oxford: Oxford University Press.
Cameron, L. (2001). *Teaching languages to young learners*. Cambridge:

Cambridge University Press.
Cheng, L. & Watanabe, Y. (Eds.) (2004). *Washback in language testing: Research contexts and methods* (pp.19-36). London: Lawrence Erlbaum Associates, Inc.
Cook, G. (2000). *Language play, language learning*, Oxford: Oxford University Press.
Gipps, C. (1994). *Beyond testing*. Brighton: The Falmer Press.
Hasselgren, A. (2000). The assessment of the English ability of young learners in Norwegian schools: An innovative approach. *Language Testing*, 2 (17), 261-277.
Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.
McKay, P. (2006). *Assessing young language learners*. Cambridge: Cambridge University Press.
Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2009). Foreign language activities. Retrieved August 1, 2009 from: http:// www. mext. go.jp/a_menu/shotou/new-cs/youryou/eiyaku/gai.pdf.
Nunn, R. (2000). *Designing rating scales for small group interaction. ELT J*, 54 (2), 169-178.
Nunn, R. & Lingley, D. (2004). Formative evaluation and its impact on ELT curriculum. *JACET Bulletin*, 39, 73-86.
Pinter, A. (2006). *Teaching young language learners*. Oxford: Oxford University Press.
Tsui, A. (1994). *English conversation*. Oxford: Oxford University Press.
Underhill, N. (1987). *Testing spoken language*. Cambridge: Cambridge University Press.
Weir, C.J. (1990). *Communicative language testing*. London: Prentice Hall.