Piloting a Listening Test for Placement Purposes

Andrew Gorringe and Sergio Mazzarelli

1. Introduction

The object of this article is to describe how the authors piloted a computer-based listening test to be used to place first-year Kwassui Women's University students into ability groups.[1] It is hoped that such a description may be useful to other EFL/ESL professionals who wish to undertake similar projects. Once completed, the listening test is intended to supplement a vocabulary test that has been used for the same purpose since 2008. Although the vocabulary test was found to be valid and reliable (Gorringe 45-47) and replaced the previously used CASEC test, it was suggested that an additional listening test may help place students with even greater accuracy, especially in classes with an important listening component. It was decided that this new test would be used from April 2011, but that a pilot version would be pre-tested in May 2010. The present article first explains the procedures followed to create the pilot test and then details as well as discusses the results of its pre-testing.

2. Criteria for the New Test

It was felt that the listening test should meet the following criteria:

a. Results should be available immediately after test administration, as students need to be assigned to various groups straight away.

b. The test should be able to discriminate between students of differing language ability. In particular, it should measure ability differences between low-level listeners, who cannot be reliably placed using most commercially available tests.

The first criterion led to the choice of a computer-based test and to the selection of test items that could be scored automatically. The *Moodle* learning

[1] The authors wish to thank their colleagues Richard Bent and Karen Masatsugu for their participation in the creation of the audio recordings used for the test.

management system, already used for the vocabulary test, offered an excellent platform for creating and administering such a test.

The second criterion led to the decision to create a large question bank, since the more items there are on a test the more certain one can be of its reliability. In addition, it was decided to exclude questions that would require skills unlikely to be possessed by low-level listeners. While all language learners encounter problems at the perception or parsing stage of listening comprehension, low-level listeners are very often unable to overcome them by using top-down strategies (Goh 67-68). Developers of testing materials often appear to overlook this issue, perhaps because current trends in teaching listening emphasize top-down processing, and the result is that tests cannot discriminate between low-level listeners.

Since one of the main obstacles to comprehension is that learners quickly forget what they hear (Goh 60), it was decided to subdivide recordings into short sections and to let individual test takers pause and replay them at will. Therefore, the use of room speakers was ruled out. Instead, the recordings would be uploaded to *Moodle*, which allows the embedding of audio files into test pages. Equipped with headphones, test takers would be able to listen to each recording section by clicking on a small audio-player icon placed just above the relevant questions.

3. The Pilot Test

Research also shows that listeners report being unable to recognize known words as another great obstacle to comprehension (Goh 60). In other words, listening vocabulary and reading vocabulary may not always coincide. Therefore, it was thought that it may make sense to measure listening vocabulary separately. It was decided to do this through gap-filling questions.[2] These questions required test takers to complete transcripts of the recordings. The words to be deleted were chosen among the 2,000 most frequently used words in English according to the General Service List. To speed up the process, candidates for deletion were identified by running the texts through Paul Nation's *Range* software, which had already been used to construct the earlier vocabulary test. Because of the time

[2] The term gap-filling questions rather than clozes is used in this paper because strictly speaking listening clozes are recordings from which words have been deleted and replaced by beeps or pauses. See Buck 70.

needed to develop materials, it was also decided that the pilot test would only contain these gap-filling questions, while other items would be added later. Care was taken so that the number of questions included was large enough to guarantee reliability even at this early stage.

With listening tests the choice of texts used is important, since different listening texts have varying speeds, with the fastest being authentic conversation and the slowest a lecture/narration to non-native speakers (Buck 38-41; Hughes 162). In the pilot test, three kinds of text were used. The first type consisted of simple unscripted conversations: only the general topics had been agreed and these were everyday ones. The second type of text was a scripted radio-type monologue. The final type of text used was an unscripted interview. The speakers used in the test were three males and one female, three native speakers and one non-native speaker. The recordings were made using a professional XLR microphone attached to a digital recorder, lightly edited and divided into sections using the *Audacity* software, and exported to mp3 format to limit file size. The shortest section lasted eight seconds and the longest fifty-five seconds, with most lasting about thirty seconds. The number of words deleted in the text for each section varied from one to six, with most sections containing four deleted words for the students to type. The deleted words were never immediately adjacent.

Although the test would require only the basic familiarity with computers that students were from experience known to be likely to possess, it was necessary to make sure every test taker was comfortable with the technology used. Therefore, a sample recording and questions were prepared. The computer lab allowed showing on students' individual screens a live capture of the screen of a computer manipulated by one of the proctors. Moreover, audio from this computer could be played through room speakers. Using these facilities, a demonstration of how to play the audio, fill the gaps, and submit the completed test was shown to test takers immediately before they took the test.

4. The Results

The 76-item test was administered to 39 first-year students at Kwassui on Friday May 14, 2010. There was a 45-minute time limit to the test, and the students were allowed to leave the testing room as soon as they had finished the test (the quickest took 27 minutes 36 seconds and the slowest took 42 minutes 58

seconds). The reliability of the test was determined using the internal measure of reliability Kuder-Richardson 21 (KR-21).

Mean = 50.92, Median = 52, Minimum = 28, Maximum = 71
Standard Deviation = 9.42
Reliability = 0.83
Standard Error of Measurement = 3.97

From the statistics above we can see that the test has a normal distribution about the mean and has "moderate" reliability.

5. Comparison with the TOEIC IP and the Kwassui Vocabulary Test

Nineteen of the students who took the pilot listening test in May also took the TOEIC IP on June 6, 2010－less than a month after the listening test－so a comparison could be made between those students' scores to determine the concurrent validity of the listening test. The scores from the listening test and the listening section of the TOEIC IP were compared using Pearson's r (correlation coefficient) and Spearman's Rank Correlation Coefficient. With Pearson's r the coefficient was 0.50 and the Spearman's Rank Correlation Coefficient was 0.56. However, these coefficients show little correlation between the two tests.

All the students who took the pilot listening test had previously taken the vocabulary test on April 5, 2010. When comparing the listening test and the vocabulary test Pearson' r Coefficient is 0.69 and Spearman's Rank Correlation 0.66. These figures indicate a reasonable correlation.

6. Discussion

The test was easy to administer and took just over 40 minutes for the slowest of the students to complete. It was found to be moderately reliable and was normally distributed about the mean. However, the test did not correlate well with the listening section of the TOEIC IP listening section. There can be various reasons for this lack of correlation, the most obvious being the small number of students who took both tests. Another reason could be that the TOEIC IP is business-oriented, and the vocabulary as well as the situations used in it may not be familiar to first-year students at a Japanese college, giving those students low,

meaningless scores.

On the other hand, there also could be a problem with the pilot listening test itself and the method of testing used. The gap-filling method used in the test required the students to type their answers, and so correct spelling was needed, even though that in itself is not listening. It must be remembered, however, that there is no way to design tasks that require only listening skills (Buck 128). For example, thirty of the TOEIC IP listening items require test takers to hold a question and three answers in memory before deciding on the right answer. Memory and attention skills of this type would not seem to be part of listening.

Looking at students' individual answers to the pilot test showed that some mistakes indicated grammatical deficiencies that would not hinder comprehension, as if one were to write *"Did he shouted?" instead of "Did he shout?" However, far more errors appeared to reflect problems, such as phoneme discrimination difficulties, that prevented test takers from recognizing words altogether. Conversely, although the exact classification of spelling mistakes is difficult, very few errors were of the type that shows awareness of phonology, as if one were to write *sed* for *said, flud* for *flood,* or *fone* for *phone*.

A final methodological limitation of the method employed could be that items in gap-filling passages are often interdependent, so to answer an item correctly may depend on the test taker's knowledge of other items in the particular section of the test.

The reasonable correlation between listening and vocabulary test may result from the fact that, after all, they both assessed vocabulary.

7. Conclusion

The authors feel that the new listening test will be a worthwhile addition to the vocabulary test that has been successfully used with first-year students at Kwassui over the past few years and look forward to improving it and using it from April 2011. The results of the vocabulary test had already demonstrated the importance of filling gaps in the first-year students' knowledge of high-frequency vocabulary. The data from the pilot listening test, albeit preliminary, appear to confirm the centrality of this issue. In addition, such data can be analysed for patterns in students' word-recognition errors so as to individuate possible perceptual processing problems that can be addressed in the appropriate classes.

On the other hand, the already planned addition of new kinds of questions that will test other listening skills appears essential. Since all types of questions have their drawbacks, relying on a variety of types should reduce the impact of any single drawback and provide more balance. Some of the new recordings will take the form of excerpts from lectures to non-native speakers. As this genre of recordings will be slower than that of recordings made so far, it should be easier to associate them with multiple-answer comprehension questions that would be suitable for low-level listeners.

As the test comes into regular use, further analysis will be undertaken to determine the extent to which it can successfully place students into differing ability groups.

Works Cited

Audacity. Vers. 1.2.6. <http://audacity.sourceforge.net/>.

Buck, Gary. *Assessing Listening*. Cambridge: CUP, 2001.

CASEC (Computerized Assessment System for English Communication). <http://casec.evidus.com/>.

Goh, Christine C. M. "A Cognitive Perspective on Language Learners' Listening Comprehension Problems." *System* 28 (2000): 55-75.

Gorringe, Andrew. "The Design and Implementation of a Computer-Based Placement Test." *The Kwassui Review* 51 (2008): 43-48.

Hughes, Arthur. *Testing for Language Teachers*. 2nd ed. Cambridge: CUP, 1989.

Moodle. Vers. 1.9. <http://www.moodle.org>.

Nation, Paul. *Range*. General Service List Version. <http://www.victoria.ac.nz/lals/staff/paul-nation.aspx>

TOEIC IP (Test of English for International Communication, Institutional Program). <http://www.toeic.or.jp/toeic_en/>.

Received January 31, 2011