# Speak and Write： Assessing Spoken Production in Large Classes

Keith Adams

Testing speaking in any foreign or second language （L2） classroom is one of the more difficult evaluation tasks a teacher may face.　Contrasted with discrete point grammar or reading tests, for example, where class size, time for testing, and reliable and efficient assessment of results are manageable issues, speaking tests are far more challenging, especially so when one is dealing with large numbers of test-takers.

With a group of 25 or more students in a 90-minute class session, which is a common testing situation in Japanese university classes, it is very difficult to allot even five minutes for student pairs to complete meaningful and sufficient exchanges in front of the teacher.　Of course, the teacher/tester is also under extreme pressure to have to make on-the-spot assessments of the students' test performances.

This paper will present a spoken production test that can be used in classes of various sizes but is particularly relevant to testing large groups of students.　Discussion will revolve around an analysis of the examination design and format, methods of scoring and interpreting results and test administration from the perspective of various theoretical and practical issues that are essential in evaluating the effectiveness of a test.

The first part of this paper will describe the design of the test task and procedures of implementation before, during and after the test.　Subsequent sections will examine theoretical and practical issues with reference to the test task and procedures addressed in the first section.

## The Test-Task : 'Speak and Write'

The test task is a paired speaking format with a different twist. Instead of the test-takers engaging in some type of verbal exchange, such as a mock interview or role-play, 'Speak and Write' (S&W) assigns one of two basic roles to the partners – the speaker and the writer.

The speaker is given five minutes to speak freely about a prepared topic while the partner writes every word the speaker says. At the conclusion of the allotted time, the writer counts every word the speaker produced and enters the total on the speaker's paper.

In the next round, the writer becomes the speaker and vice versa. In subsequent rounds, the students continue alternating between the two roles. However, students change partners after each round regardless of what role they are assigned for the round. That is, if a student speaks three times and writes three times, she will be involved with six different partners.

Further considerations behind the rationale of alternating partners after each round will be brought up in a later section of this paper.

Topics are also alternated so that there is a balance in that the first group of speakers does not always speak first for each of the topics, which could potentially give the first group of writers a slight advantage when it is their turn to speak. Table 1 illustrates the sequencing of roles and topics.

Finally, each round takes approximately 12-15 minutes. In addition to the five minutes of speaking, a teacher needs to factor in the time required for

Table 1. Sequencing of roles and topics

| Round | Speaker | Writer | Topic |
|---|---|---|---|
| 1 | A | B | About Me |
| 2 | B | A | Japanese Culture |
| 3 | A | B | Japanese Culture |
| 4 | B | A | About Me |

*Note.* A and B refer to all students assigned to Groups A and B the writers to count the words and for students to move to a new partner and get settled for the following round. Therefore, if the test involves only two topics, as seen in Table 1, it could be completed in 50-60 minutes. Adding a third topic would still enable a teacher to administer the test in a 90-minute timeframe.

## The Teacher's Role

During the test time, the teacher is responsible for seating students and moving them to new partners, assigning roles and speaking topics, keeping time and observing if any problems arise, such as any breakdowns in speaking or writing. Of course, at the conclusion of the test, the papers are collected for scoring and assigning grades.

See the Appendix for the instructions to the speakers and writers and the system of moving students to new partners.

It is clear that a teacher's role is primarily focused on ensuring that the test runs smoothly on the day ; however, the teacher has a more demanding, vital role in the lead-up to test day and assessment of the students' performance afterwards. The first area to examine in this regard concerns the decisions and procedures involved in preparing students for the specific test task.

As stated previously, students are informed of the speaking topics in advance. Quite naturally, these topics emerge from the course content so the teacher should choose appropriate themes that have been covered in the class syllabus and lend themselves to the test format. For example, the students in this writer's classes were enrolled in first and second year English communication classes, so typical themes included giving a self introduction, providing street directions, talking about Japanese culture or giving a tour guide description of one's hometown area.

The next key step is to orient the students to the test format and procedures. To do so, the students are given at least one trial run a week or two before the actual test day. The alternating roles the students will play need to be carefully explained, as well as their responsibilities in each role. Similarly, they must be given firsthand experience with the basic logistics of changing partners and exchanging papers.

During the trial run, it is recommended that the teacher choose two themes to give students sufficient experience with the test procedures. Although one might choose topics that will not appear on the test, the other option is to select two announced topics.

The advantage of the second option is that it makes the test a learning experience as well as a measurement of achievement, that is, the test has a positive 'washback' effect on teaching and learning. Giving the students a chance to work with actual test topics aids them in preparation for the real test. Furthermore, if they are allowed to keep their practice test papers, they can use the papers as a resource to increase their output during the actual test.

One suggestion for building upon the practice session is to encourage the students to think about what they wanted to say, but couldn't due to a lack of knowledge of desired vocabulary or uncertainty about a grammatical pattern. This is particularly valuable for the weaker students in the class and makes a positive contribution to the fairness of the test task, which will be addressed in more detail in a discussion of the issue of fairness in testing.

## Interpreting Results and Determining Scores

### Writers

Perhaps the first point to be clarified here is that the writer is not assessed in any way. The writer's sole responsibility is to record every word the speaker produces. Accuracy in terms of correct spelling or punctuation is not taken into consideration. As long as the writer has been accurate in recognizing and recording the speaker's words legibly, the writer has fulfilled the requirements of the role.

However, several questions relating to the writer naturally come to mind, such as the competency of the writer to record what has been said, and these will be examined later in this paper.

### Speakers

As for the core aim of assessing spoken production, one effective tool to evaluate the students' performances is to calculate the mean and standard deviation of the word count results and determine test grades based on the distributions.

Past experience with various groups has shown that the word counts have consistently resulted in satisfactory distributions for grading purposes. However, it must be stated that the results have sometimes revealed negative skewedness (indicating that many students had high scores on the test) such as with Class B in Figure 2 below, or outliers (scores falling outside of the normal distribution), both of which may create problems in determining cut-off points in for grades.

Nevertheless, these statistical issues have not been overly problematic. In fact, as Brown (1997.) stated, negative skewedness in course achievement tests "may actually be a desired outcome" (p. 21) in that it indicates the students have learned the material well.

In sum, by evaluating word count totals through an analysis of the mean and standard deviation, a teacher can get a quick, relatively accurate picture of how students performed on the test task.

Figures 1 and 2 illustrate the distribution of scores from two different classes of university English majors. The means (with standard deviations in parentheses) for the Classes A and B were 206.89 (40.26) and 219.53 (42.37), respectively.

Although the word count is the primary focus of test task performance, other factors, such as vocabulary range – as measured by the number of different words produced – could also be included. Of course, it is a given that students must be informed of all criteria for assessment in advance of the test.

If vocabulary range is included in the assessment, the procedure for the collection of data involves the students in a quick and efficient post-test task. To obtain the number of unique (different) words, a form with the letters of the alphabet as column headers is distributed to students. The students are then given their own test form to enter the different words they used according to the first letter of the word. Unlike the word count, where the same word can be counted more than once, a word is only entered one time in the

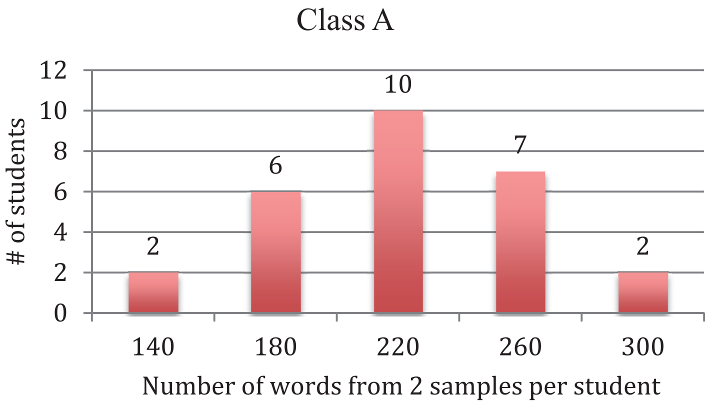

Figure 1 : Distribution of results based on mean and standard deviation.

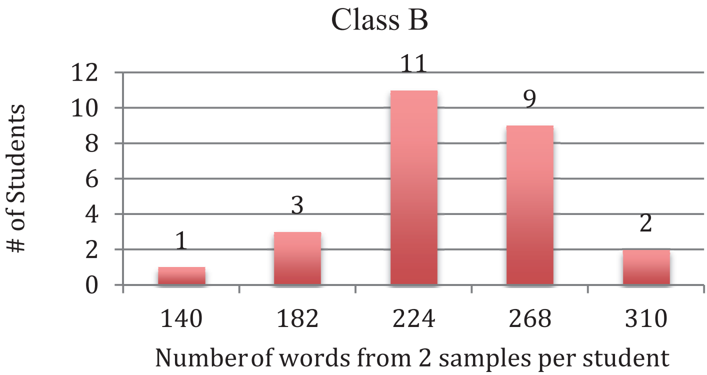

Figure 2 : Distribution of results based on mean and standard deviation.

unique word count.

This method of data collection has the double benefit of saving the tester considerable time and giving the students immediate feedback with a second perspective on their performance.

Once a tester has the unique word data, there are different possible avenues of measurement ; however, the tester does have to make some important decisions that can be seen in the data in Table 2.

In general, those in the top percentile (S1-S4) in total words produced were also at or near the top of the unique word rankings. However, the students who produced between 92-101 total words (S5-S7) had unique word totals that were very similar to the bottom group of students in the total words ranking. On the other hand, S8 was slightly below the group mean in total words (92), but produced the fourth-highest number of unique words.

These results do raise some fundamental issues in assessing spoken production. For example, was S5 more 'fluent' than S8 due to a superior total word count ? Did S8 demonstrate greater complexity of production than S5 given a higher ratio of unique words to total words ? Did S5's word count indicate better mastery of the vocabulary range than S13 ?

There are statistical procedures that can provide partial answers to these questions. One example is a Fluency Index (FI) formula used by Bonzo (2008) to measure fluency in writing samples. However, the formula was designed to distinguish between students who had the same percentage of unique words but who had different word totals. In other words, both S3 and S9 had almost identical unique word percentages (68% and 69% respectively),

Table 2. Total Word and Unique Word Results from one sample.

| Student | WT | UW | Student | WT | UW |
|---|---|---|---|---|---|
| S1 | 118 | 71 | S9 | 89 | 60 |
| S2 | 117 | 73 | S10 | 84 | 57 |
| S3 | 109 | 74 | S11 | 83 | 55 |
| S4 | 109 | 60 | S12 | 79 | 52 |
| S5 | 101 | 53 | S13 | 75 | 54 |
| S6 | 98 | 54 | S14 | 74 | 55 |
| S7 | 92 | 55 | S15 | 64 | 50 |
| S8 | 90 | 68 | | | |

*Note.* WT=total number of words produced in the sample. UW=number of unique (different words) in the total number of words.

but S3's greater output (109 vs. 84 words) resulted in a clearly superior rating on the FI : 4.77 vs. 4.26. This example suggests that perhaps total output should be given more weight than unique words in this test task assessment.

However, for the practicing teacher who may not have the time or expertise for advanced statistical analyzes, the results in Table 2 could be dealt with in a less sophisticated but practical way. First, the unique word total could be analyzed through the same mean/standard deviation measurement used for the total word counts. Then the teacher would have to determine relative weights for the two data sets, based on the teacher's judgments and interpretations of the appropriateness of the two criteria to the teaching and testing context.

Of course, a tester could also choose to focus on selected 'content' vocabulary items rather than basing assessment on the unique word figure, a significant percentage of which will naturally be comprised of grammatical items. In this case, one might narrow the focus by assessing the use of selected key vocabulary and including that factor in the students' grade for the test.

In conclusion, a tester can use a range of vocabulary measurement in various ways. Although its inclusion calls upon the tester to make some important decisions in terms of weighting the two measurements, an assessment of vocabulary range can complement the total words measurement and can also serve as a very valuable learning resource for the students.

This brings an end to the discussion of the task test, procedures and assessment. The following sections of this paper will focus on some of the theoretical and other associated issues relating to this approach to testing oral production.

## Discussion

### The test construct

In this paper, English speaking ability is defined as the ability to talk about a known topic for five minutes. Success in the test task is measured by the clarity of speech (judged by whether thoughts can be communicated clearly enough so that a partner can comprehend and write the speaker's words), and the total number of words the speaker can produce in the time limit.

At its basic level, the test task does not assess complexity or grammatical accuracy, thus it can be characterized as a test of fluency as measured by the number of words spoken and recorded. An additional element — vocabulary range measured by the number of different words used — can be added for assessment of performances.

**Monologues to measure fluency**

Although the speaker and the writer do interact occasionally in the form of requests or offers for clarification, the test task is essentially a monologue. Although monologues naturally have fewer instances of communication management or pragmatic factors than dialogues, Long (2013) has shown that in terms of assessing fluency, monologues are an effective test task to evaluate a speaker's performance.

In Long's study, the spoken output of Japanese university students in four types of speaking tasks — monologues, dialogues, structured interviews and summaries – was investigated. Discourse was analyzed by means of several fluency indicators, including word count, frequency and length of pauses, and rate of articulation (speaking speed).

The results of Long's analysis indicated that monologues compared favorably with the other speaking tasks in fluency indicator ratings. That is, all were appropriate tasks to which fluency measurements could be applied.

Furthermore, the results of comparing monologues and dialogues revealed that fluency indicators were quite consistent for both tasks, though differences were found in pause frequencies and speaking speed. In both cases, there were more pauses in dialogues than in monologues and speakers also spoke at a faster rate in dialogues.

In terms of the Speak and Write task, the difference in pause frequencies is quite interesting. To draw a parallel interpretation, the pauses that occur while a writer is transcribing the speaker's words in S&W could be seen as equivalent pauses by a speaker in a dialogue to deal with the interlocutor's questions or comments. In other words, in terms of disruptions to speaking time, the differences between a monologue and an interactive conversation may not be as great as one might assume.

### Fairness

In most, if not all, L2 classrooms there is an inevitable imbalance in the background knowledge and skills students can draw on in a test situation. Even if most students can be said to fall within the same general level, there will be students who have gained greater prior background knowledge of English, and in particular, have had more opportunities to develop their proficiency in spoken English.

An undesirable consequence of these differences is that some students could succeed in a test task with relatively little preparation and effort, while others with weaker spoken English skills would be at a disadvantage.

So how can the playing field be leveled ? Essentially, total equality cannot be guaranteed ; however, one way to minimize the inequalities is through the content of the test-task that would require even the 'stronger' students to prepare in order to succeed on the test. Conversely, 'weaker' students could compensate and be rewarded by thorough preparation so that they could narrow the gap and approach the production rate of the stronger students.

Lastly, if one takes the option of including unique (different) word counts in the assessment ; this could add an element of fairness for the stronger students. Assuming that they might indeed have a greater range of vocabulary, including a uniqueness assessment may motivate them to go beyond 'resting on their laurels' and try to maximize their test performance as a reward for their previous efforts and improvement in spoken English skills.

### Anxiety

This affective factor is a major concern in testing speaking. Any type of public speaking is something that most people fear due to concerns about lacking the required knowledge and skills to deliver a speech or presentation successfully.

In terms of speaking tests, the format of the test could exacerbate these fears. Even a one-to-one, teacher-student interview, which is one type of speaking assessment format, can be just as intimidating as making a formal speech to a large audience. In the interview format, the power relationship that exists puts a great deal of pressure on the student and could adversely affect the student's spoken performance (Taylor, 2001).

Although a certain degree of tension is inevitable, and possibly even desirable if it adds an element of excitement, the test developer needs to try to minimize the negative impact of anxiety on test performance. This can be accomplished in various ways, such as using student pairs instead of a teacher-student interview format.

Another important ingredient is adequate preparation and practice with the test format, so that students are familiar with their roles and responsibilities in the task. This also has the potential to engender a sense of self-belief and confidence in the test-takers that success can be attained, which in turn may reduce anxiety levels.

It was with these various affective variables in mind that the S&W format and preparation procedures were designed.

### The Writer

Although the responsibilities of the writer are limited to recording the speaker's words, the role is demanding and naturally raises several questions, such as the speed of recording, comprehension of the discourse, and matters related to the policy of allowing the writer to perform the word count.

To begin, the physical effort involved in intensive, 'real time' writing should not be under-estimated. One might suggest that the speaker be given more than five minutes for the task ; however, from personal experiences as a writer, five minutes approaches the mental concentration and physical limit of the writer's endurance.

Of course, the other 'physical' consideration is the speed of the writer. Naturally, some will be faster than others, but what causes any differences in writing speed and what are the consequences for the speaker ?

'Slow writers' were perhaps one of the main initial concerns for the success of this testing format. Anticipated problems might have been for purely physical reasons, but factors such as a concern for correctness or penmanship could have emerged as potential stumbling blocks. Of course, the listening comprehension abilities of the writers were critical factors. If a writer had poor comprehension, the speaker would naturally be severely affected.

However, only on very rare occasions have writers failed to live up to expectations. One possible factor contributing to this positive outcome was familiarity with the themes, key vocabulary and structures enabled writers to follow and record the speaker's output competently. Other contributing factors to comprehension might have included familiarity was the students' accents and pronunciation of English words, the manageable rate of delivery and the relative lack of tension in a peer pair-work environment. Finally, opportunities for clarification were possible as the speakers could monitor the transcriptions and repeat or clarify if necessary.

In situations where a writer was not performing satisfactorily, there are various avenues the tester can take to minimize or rectify problems so that a speaker does not suffer from a slow writer.

The first safeguard is the design of the procedures mentioned earlier where partners are changed after every round. In this way, a slow writer would not diminish a speaker's production more than once. Granted, two or three different speakers may suffer if a writer under-performs in every round ; however, the main responsibility for the teacher is to note the problem and take the circumstances into account during the assessment process.

In an extreme case where a student is not capable or unwilling to fulfill the duties of a writer, an extra round with other writers for those speakers affected by the breakdown could be added after the regular rounds have concluded.

Adding an extra round is also an option when there are an odd number of students present for the test, which will naturally result in some students having to sit out a round and miss a speaking turn.

To conclude this thread, if students have had adequate preparation and practice time as writers and partners are changed, potential problems with 'slow writers' can be largely mitigated so that any variations in writing speeds will be minimal.

The final issue involving writers is the policy to allow the writers to perform the word count. This of course brings up the issue of honesty or the potential of writers 'helping' the speaker either in the word count or with suggestions for the speaker during talking time.

Once again, this issue has rarely emerged as a serious concern. It doesn't take long for a teacher to develop the ability to spot questionable word counts when reviewing the transcriptions. Furthermore, unless a word count puts a student on the borderline of a grade for the test, slight discrepancies may not be of great concern. At any rate, the tester always has the option of confirming any word counts if desired.

However, allowing the students to count the words not only saves the tester a considerable amount of time, but also sends a message of trust and respect to the students which may have positive ramifications for the class environment beyond the test day.

## Conclusion

In evaluating any test, various theoretical and practical qualities must be considered to ensure that the test is appropriate and useful. Core questions focus on a test's reliability, construct validity and authenticity, while practical issues include the time and resources available for test development, administration of the test and scoring the test results.

Bachman and Palmer（1996 maintain that a test's "usefulness"（p. 18）is the result of balancing the combined qualities so that a test is developed with a clear purpose, aimed at specific group of test takers and has appropriate tasks for the purpose in mind. The test and procedures reported in this paper were developed with this concept of usefulness as a guiding theme.

The purpose of the test was to evaluate students' spoken production accurately and efficiently in the challenging context of having large numbers of students to assess in a limited time frame.

The chosen test task was not only effective in dealing with the number of students to test, but also appropriate in that a monologue was shown to be a valid discourse task to measure spoken output.

As for assessing the results of the test, a relatively consistent pattern of results emerged with different groups of students based on an analysis of the mean and standard deviation of the total number of words produced.

This method of measurement provides the data to give the teacher a clear picture of the students' results and a sound platform for assigning final grades for the test. An additional option of including a unique word count was also presented to give teachers another tool to assess performances.

The test task and overall implementation also incorporated key global considerations. First was the choice of topics that were clearly linked to course content so that the students would see the test as relevant. In a similar vein, although S&W was first and foremost a means of assessment, decisions such as allowing students to input data for their own unique word counts were made with the desire to integrate testing and learning.

Another major concern was directed to the fairness of the test in classes where the students often have mixed English-speaking skill abilities. Therefore, efforts were made to ensure the students had adequate preparation for the test content and procedures, which also could contribute to keeping test-day stress and anxiety at reasonable levels.

From the tester/teacher's point of view, the S&W format makes the various facets of testing and evaluating large numbers of students quite manageable. Testing 25 or more students can be done in one class period, and with the praiseworthy contributions of the student writers in the dyads, the teacher has the written transcriptions and word counts immediately available for analysis.

To conclude final remarks, evaluating L2 speaking ability accurately is very complex even under ideal conditions (i.e. small numbers of test-takers). For example, even the construct of oral fluency, often used as a measurement of speaking proficiency, is very connotative as there are a multitude of definitions and indicators that could be applied to the construct (Wolfe-Quintero, Inagaki and Kim, 1998).

If a tester is also presented with a situation that has significant logistical challenges, such as large numbers of students to be tested and evaluated by one person, the task can appear almost overwhelming.

Speak and Write is an approach that attempts to simplify the criterion for evaluation based on one or two indicators（word count and unique word count） and suggests a readily available statistical tool for analysis and subsequent scoring of student performances.

Finally, the task format ensures that the test can be completed smoothly on the day, and perhaps even make the test an interesting event for testers and test-takers alike.

## References

Bachman, L.F. & Palmer, A.S.（1996）. *Language Testing in Practice*. Oxford : Oxford University Press.

Bonzo, J.D.（2008）. To assign a topic or not : Observing fluency and complexity in intermediate foreign language writing. *Foreign Language Annals,* 41（4）, 722-735.

Brown, J.D.（1997）. Skewness and kurtosis. *Shiken : JALT Testing & Evaluation SIG Newsletter*, 1, 20-23.

Long, R.W.（2013, October）. *Complexity and Fluency Indicators of Good Speakers.* Paper presented at the 39th Annual International Conference on Language Teaching and Learning of the Japan Association for Language Teaching, Kobe.

Taylor, L.（2001）. The paired speaking test format : recent studies. *Research Notes*, 6, 15-17. Extract retrieved from http://www.cambridgeenglish.org/images/22642-research notes-02.pdf

Wolfe-Quintero, K., Inagaki, S., & Kim, H.Y.（1998）. *Second language development : Measures of fluency, accuracy & complexity*. Second Language Teaching and Curriculum Center : University of Hawaii Press, 1998.

## Appendix

### I. Test paper instructions

Speaker's Name : ____________________________________

**Speaker :** Talk about the topic for 5 minutes.

- Don't worry about mistakes (grammar, pronunciation etc.).
- Communicate as much as you can.

**Writer :** Write every word the speaker says.

- Don't worry about spelling !
- Count the words. Write the number on the line.

**Speaker** : When you speak, give this paper to the writer.

Topic _____ : Writer's Name : _______________________

Total words : ______________

Note : The writers' transcriptions are entered on lines below the instructions.

### II. System for Assigning Roles and Changing Partners : 28 Students

**Assigning Roles**

1. Seat students in 4 rows of 7*.
(*equal numbers when possible ; adjust number of students in rows according to student numbers)
2. Designate rows.
   - Rows 1 and 3 : 'A' partner.
   - Rows 2 and 4 : 'B' partner.
3. Assign roles for Round 1 :

'A' is the speaker ; 'B' is the writer.

- 'A' gives his/her test paper to 'B'.

4. After the five-minute speaking time finishes, the writers count the words and return the paper to the speakers.

**Changing Partners**

1. The first person in Rows 1 and 3 moves to the back of the row.
   - Everyone else moves up one seat.

2. Students in Rows 2 and 4 don't move.

Continue in this way for each subsequent round.