Three Sources of Measurement Error in a Speaking Test

Melvin Shaw, Kent Lavoie and Cynthia Cunningham※

The purpose of an oral proficiency test in English is to measure the ability of each student to speak English. How well can such a test succeed in measuring this ability? To answer this question, we must acknowledge that other factors besides speaking ability often affect a student's score on an oral test. As a first example, students might somehow obtain important information about the interview questions in advance, perhaps through information leaking out of the test interview room, and thus be able to boost their scores. Secondly, the human rater or interviewer might be biased for or against the student by irrelevant factors. Thirdly, assuming that the test has multiple sets of tasks or questions, a student might be asked only the easiest or the hardest questions, the result in either case being that the score would probably not reflect the student's true ability. These three examples illustrate sources of what Brown has called "*measurement error, or error variance*" (1996: 186). A primary goal of the language tester is to reduce such error to the minimum. While there are many sources of error variance, in this report we will investigate only the three exemplified above: variance caused by leaked information, variance caused by rater error, and variance caused by unequal forms of the test. Each of these will be investigated through an examination of an English speaking test given to 130 students at Aomori Public College on June 4, 2003.

In the report that follows, we give an overall description of our test first. Secondly, we describe its method of administration. After that, we give an account of the alternate forms that were used. Finally, the results of the test are analyzed for the three forms of error variance mentioned above.

Overall Description of the Test

The test was part of an ongoing effort of several teachers who are interested in promoting and measuring English speaking proficiency. For a description of an earlier version of this test, please see Schneider et al. (2001). The test of June 4 was a slightly longer and (hopefully) improved version of that earlier test. Like the 2001 test, the test of June 4, 2003 was not related to the students' textbook, and the precise contents of the test were not announced or taught to students in advance. It was felt that by not announcing details of the

※Aomori Public College

test contents, we could better measure the students' true speaking ability (rather than their overnight cramming ability).

The total score for this test was 20 points. This total was calculated by summing the subscores of the test's four parts, as follows:

Part I, Pronunciation = 3 points,
Part II, Small Pictures Discrimination = 3 points,
Part III, Sentence Repetition = 7 points,
Part IV, Big Picture Description = 7 points.

Details of these four parts are given below, along with references to sources which have been helpful for understanding the type of speaking task associated with each part.

Part I, Pronunciation. (Dobbyn 1976, as cited in Nation 1998: 78-79). The student read sentences from a prompt paper which the rater could not see. (This "blindness" factor was critical.) Each sentence contained a target word which the rater listened for (e.g., sip or ship). The rater listened and marked the word heard. Then the rater's marks were compared with a key to yield a subscore. This part of the test evaluated the student's pronunciation of several segmental phonemes that have proven troublesome for Japanese learners of English.

Part II, Small Pictures Discrimination. (Upshur 1973; Oller 1979: 317-319). As in Part I, there was an important blindness factor in Part II. Here the student looked at small pictures which differed only in one or two details. The student had twenty seconds to describe a marked picture accurately enough for the rater to select the same picture on his/her paper, which was unmarked. The rater's marks were compared with a key to yield a subscore. This part of the test evaluated certain aspects of the student's oral communicative ability.

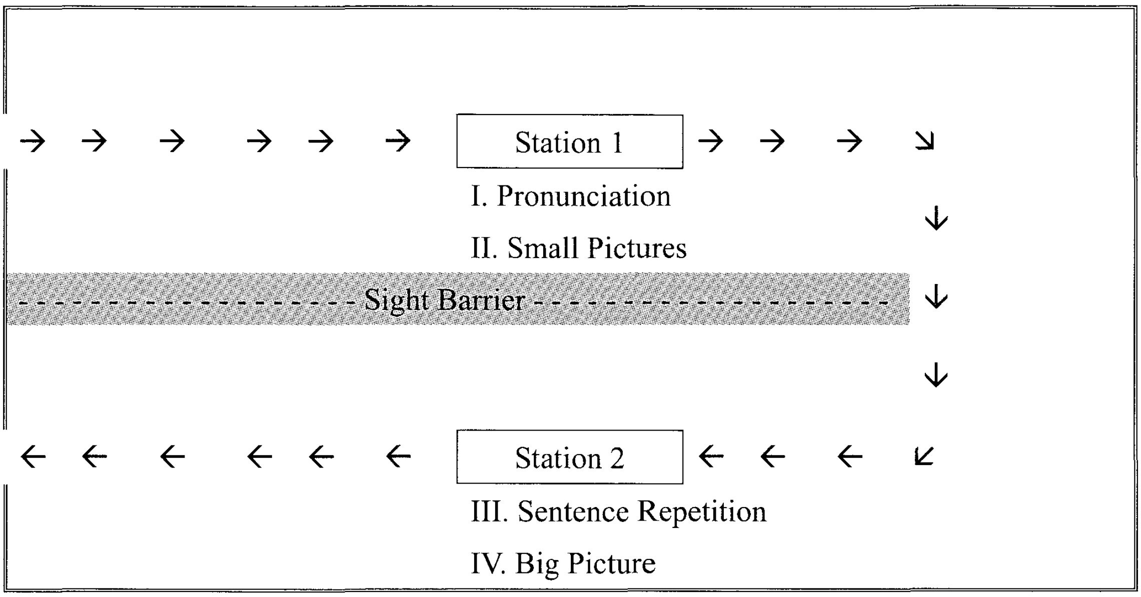
Part III, Sentence Repetition. (Underhill 1987: 86-87; Larsen-Freeman and Long 1991:28). The students repeated sentences that they heard the rater say. The rater awarded one point for each repetition that was completely correct. Although short-term memory was one factor in this task, the real target of evaluation was the student's sense of grammatical accuracy.

Part IV, Big Picture Description. (CAL, 1982) Students were shown a picture rich in language stimuli. They were encouraged to talk as much as they could about the picture for up to two minutes. The rater compared their performance with the descriptions of eight levels (0-7) on an oral rating scale. The level which best matched the student's performance was chosen as the subscore. This part of the test evaluated several factors, especially the student's fluency.

Method of Administration

Students were given information about the format of the test in regular classes before the test date. They were also given opportunities to practice sample test items. A schedule showing students' names and interview times was posted on test room doors several days prior to testing. Two test rooms were used, and 65 students were scheduled for each room. The test rooms were regular classrooms which had been set up to accommodate two interview stations. Desks and chairs were piled up in the middle of each room to make sight barriers. Figure 1 shows the basic arrangement.

Figure 1. View of a testing room and movement of a student.



According to the schedule, students were to enter the testing room, one at a time, at four-minute intervals. When it was his/her turn to be interviewed, the student entered and sat at the first station, where he/she encountered the tasks associated with Parts I and II. These were administered and scored immediately by Rater 1. The student then moved to the second station, where he/she encountered Parts III and IV. These were administered and scored on the spot by Rater 2. (In the meantime, the next student had entered the room and was being tested by Rater 1.) This process continued until all 65 students assigned to the room had been tested. We had originally allowed four hours and twenty minutes for testing all of our students, but we actually ended up taking somewhat longer.

The above description reveals that this was a direct test of speaking in which each interviewer also acted as the sole rater for those parts of the test that he/she administered. About eight minutes were required for giving the complete test to one student. After all tests had been given, the total score for each student was calculated by summing the subscores for the four parts.

Alternate Forms of the Test

Because it would take four to five hours to test all of our students, we realized that we had to control for important information leaking out of the testing rooms. We knew that students would be waiting outside the test room doors, and that those students who had just been tested would be questioned by these waiting students. If our test had consisted of just one set of tasks to be repeated each time, its contents would have been well known to the waiting throng by the time the fourth or fifth student finished. All students tested later would have had access to this shared knowledge of the test, something which the first test-takers would not have had. Thus, the order in which the students were tested would have become a source of error variance compromising the validity of the scores.

To combat this leaked information problem, the test we gave on June 4, 2003 had alternate forms for Parts I-II, III, and IV. (For a bit more about alternate, "parallel forms" of an oral test, see ETS 1999:5.) To make administration of our test easier, Parts I and II were placed on the front and back sides of a single page, so that the two parts were effectively linked. There were ten such pages, the salient features of which are shown in Figure 2.

Figure 2. The ten alternate forms of Parts I-II.

Alternate A	Alternate B	Alternate C	Alternate D	Alternate E
ship, vest, wrong	sip, best, long	ship, vest, wrong	sip, best, long	ship, best, wrong
↓$	↓¥	↑$	↑¥	↓$
□ 6	6 □	□ **six**	**six** □	6 □
♡♥♡	♡♥♡	♥♡♡	♥♡♡	♥♡♡

Alternate F	Alternate G	Alternate H	Alternate I	Alternate J
sip, best, long	ship, vest, wrong	sip, best, long	ship, vest, wrong	sip, vest, long
X ☜	X ☞	☞ X	☞ X	☜ X
■□■	■□■	□■□	■□■	■□■
X X $\underline{\text{x}}$	X $\underline{\text{X}}$ x	x $\underline{\text{x}}$ X	x $\underline{\text{x}}$ X	X $\underline{\text{X}}$ x

As for Part III, sentence repetition, there were three alternate forms of seven sentences each, as shown in Figure 3.

Figure 3. The three alternate forms of Part III.

Sentence Set A	Sentence Set B	Sentence Set C
1. Hi.	1. Good morning.	1. Good afternoon.
2. Hi, Koji.	2. Good morning, John.	2. Good afternoon, Yoko.
3. How was vacation?	3. What are you doing?	3. How are you doing?
4. Did you go anywhere?	4. How was your weekend?	4. I'm doing fine. And you?
5. Yes, I went to New York.	5. It was good. I went to Tokyo.	5. Okay. I went to work this morning.
6. Oh, really? What did you do there?	6. Really? Why did you go there?	6. Oh really? You have a part-time job?
7. I met my American friend. He lives there.	7. My brother lives there. I wanted to visit him.	7. Yes, I do. I make bread in a bakery.

Part IV had three alternate forms, which we called "office," "bus," and "pub." Each was a large black and white drawing.

Alternate "Office:" Eight adults (four men and four women) were clearly shown pursuing various activities in an office setting. Objects such as a desk, door, clock, telephone, and articles of clothing were clearly visible.
Alternate "Bus:" Fifteen people were shown standing or sitting inside a crowded bus, but some of these individuals were obscured by others and thus not fully visible. Objects such as a bag, radio, umbrella, and headphones, as well as articles of clothing were visible.
Alternate "Pub:" Eleven people were shown in a drinking establishment, but those in the background were not fully visible. Objects such as a glass of beer, ashtray, slot machine, dartboard, and articles of clothing were clearly visible.

Close examination of the alternate forms for each part reveals that they were not completely different. For example, some of the Parts I-II forms had exactly the same pronunciation items, and some of them shared one or two small pictures. Moreover, the large pictures used in Part IV had certain common features, so that the convenient utterance, "There are many people in this picture," held true no matter which of the three alternates was presented. Nevertheless, combining the several parts (10 x 3 x 3), we see that there were 90 *somewhat* different forms of this test. As things turned out, 69 of these were actually used with our 130 students on June 4. No single form of the test was used more than five times.

Analysis of Test Results for Three Sources of Error Variance

1) Variance Caused by Leaked Information

As was described previously, the 130 students were tested in two rooms. The average scores made by students in each room are reported in Table 1. Also reported are the scores of the first one-third and last one-third of the students tested.

Table 1. Average total scores.

	All students	First one-third	Last one-third
Room 1, Cunningham & Schneider	11.45 n = 65	11.23 n = 22	11.55 n = 22
Room 2, Ruuska & Shaw	10.88 n = 65	11.14 n = 22	10.55 n = 22
Rooms 1 & 2	11.16 n = 130	11.18 n = 44	11.05 n = 44

We can see from Table 1 that the average total scores of the students who took the test later were not appreciably higher than the average scores of students who took the test earlier, at least when the data from both rooms are combined. This might indicate that later test takers were not helped significantly by information gathered from earlier test takers. Is it possible that our 69 parallel forms successfully neutralized the problems of test order bias and leaked information? We would like to believe so.

2) Variance Caused by Rater Error

After the test had been given, the taped interviews of eight students from Room 1 were re-rated by the Room 2 raters. Similarly, the interviews of eight students from Room 2 were re-scored by the Room 1 raters. Although this was not a random sampling of the test takers in each room, neither was there any particular selection criterion other than this practical one: the taped interviews chosen were from among those where the speaker's name could be unmistakably ascertained and which had tolerable sound quality. The results obtained are shown in Table 2.

Table 2. Consistency of scores between the two pairs of raters.

	Test Scores		
Sixteen Students	Original Rating (live) June 4, '03	Second Rating (from tape)	Absolute Difference Between Ratings
1	11	11	0
2	12	14	2
3	14	12	2
4	11	14	3
5	12	11	1
6	11	10	1
7	10	9	1
8	14	14	0
9	10	6	4
10	13	13	0
11	10	11	1
12	10	12	2
13	15	15	0
14	8	7	1
15	10	12	2
16	12	12	0
mean =	11.4375	11.4375	1.25

Looking at the right-most column of Table 2, we see considerable inconsistency between some of these ratings. At least for the fourth and the ninth students, we are left in doubt about the true level of their performance. We cannot avoid concluding that there was appreciable error variance caused by the human raters. At this point the words of McNamara (2000:37) seem appropriate:

> Introducing the rater into the assessment process is both necessary and problematic. It is problematic because ratings are necessarily subjective. . . . The assumption in most rating schemes is that if the rating category labels are clear and explicit, and the rater is trained carefully to interpret them in accordance with the intentions of the test designers, and concentrates while doing the rating, then the rating process can be made objective. . . . But the reality is that rating remains intractably subjective.

Faced with the reality of rater error, we are determined to take those measures that are possible and practical to reduce it. Foremost among these is improved rater training. In the meantime, we are warned against using the results our test for high-stakes decisions about students' grades. As it turned out, the scores of the June 4 test contributed only a small amount (perhaps 10-15%) to each student's final course grade.

3) Variance Caused by Unequal Forms of the Test

In the discussion above of variance caused by leaked information, we surmised that our use of many alternate forms of the test may have neutralized that particular source of error. However, once multiple forms are introduced, a new problem arises: ensuring the equivalency of these forms.

In order to check for unequal forms of the test, we first calculated the mean scores that students made on each of the alternate forms of Parts I-II, III, and IV. Although these means differed from each other, was it possible that these differences were due to chance? For example, perhaps a large number of students with especially weak speaking skills all received a certain picture or pronunciation prompt simply by chance. Would not such a coincidence tend to depress the mean score for that prompt? In order to check for statistical significance among the various mean scores, we used one-way analysis of variance (ANOVA). The results of these analyses are displayed in the appendix. Examination of these results shows that no significant differences in the means were detected when α was set at .05. Therefore, we do not reject the possibility that the observed differences in the mean scores for various forms of our test occurred by chance.

Conclusion

There seems to be a trend toward more emphasis on teaching students to use English for communication, and along with this trend has come increased interest in direct tests of speaking. As classroom teachers make use of such tests, they need to guard against the various threats of measurement error.

This report contains a description of the design and administration of a speaking test given at Aomori Public College in 2003. We have described the parallel forms of this test that were used to counteract error variance caused by information leaking from the test rooms. We have examined these multiple forms to see to what extent they themselves might have been a source of measurement error. We have also faced the reality of rater error and acknowledged that it remains a troublesome problem.

Received : December 15, 2003, Accepted : December 17, 2003

Appendix

Table 3a. Observed mean scores for the ten alternate forms of Parts I-II.

Alternate	N =	Sum	Mean	Variance
A	13	53	4.0769	1.4103
B	15	50	3.3333	1.2381
C	12	44	3.6667	1.5152
D	14	48	3.4286	0.7253
E	12	38	3.1667	0.697
F	13	45	3.4615	1.9359
G	13	36	2.7692	1.1923
H	12	48	4.0000	1.6364
I	15	51	3.4000	1.8286
J	11	33	3.0000	1.8000

Table 3b. Parts I-II mean score differences by ANOVA.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	18.7	9	2.08	1.4932	0.15797	1.9588
Within Groups	167	120	1.393			
Total	186	129				

Table 4a. Observed mean scores for the three alternate forms of Part III.

Alternate	N =	Sum	Mean	Variance
Sentence Set A	40	193	4.8250	1.2250
Sentence Set B	47	215	4.5745	1.1193
Sentence Set C	43	184	4.2791	0.7774

Table 4b. Part III mean score differences by ANOVA.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	6.2076	2	3.1038	2.9881	0.0539	3.0675
Within Groups	131.9155	127	1.0387			
Total	138.1231	129				

Table 5a. Observed mean scores for the three alternate forms of Part IV.

Alternate	N =	Sum	Mean	Variance
office	46	155	3.3696	0.8159
bus	49	153	3.1224	0.9014
pub	35	105	3.0000	0.5294

Table 5b. Part IV mean score differences by ANOVA.

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	2.9481	2	1.4740	1.9106	0.1522	3.0675
Within Groups	97.9827	127	0.7715			
Total	100.9308	129				

References

Brown, J.D. (1996). *Testing in language programs.* Upper Saddle River, NJ: Prentice Hall Regents.

CAL (Center for Applied Linguistics). (1982). *Basic English skills test.* Washington, DC: Center for Applied Linguistics.

Dobbyn, M. (1976). An objective test of pronunciation for large classes. *English language teaching journal.* 30, 3:305-325.

ETS (Educational Testing Service). (1999). *TOEIC language proficiency interview manual.* Princeton, NJ: Educational Testing Service.

Larsen-Freeman, D., Long, M. (1991). *An introduction to second language acquisition research.* London: Longman Group.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

Nation, I.S.P. (1998). *Teaching listening and speaking.* Tokyo: Temple University Japan, Graduate College of Education.

Oller, J.W., Jr. (1979). *Language tests at school.* London: Longman Group.

Schneider, D., Shaw, M., Gregory, J. (2001). Report on a speaking test. *Journal of Aomori Public College.* 7, 1:46-55.

Upshur, J.A. (1973). Productive communication testing: progress report. In J.W. Oller, Jr. and J.C. Richards (Eds.) *Focus on the learner: pragmatic perspectives for the language teacher.* Rowley, MA: Newbury House Publishers.

Underhill, N. (1987). *Testing spoken language: a handbook of oral testing techniques.* Cambridge: Cambridge University Press.

Abstract

Teachers who use direct speaking tests need to be aware of the ways that external and irrelevant factors threaten these tests, and they need to take measures to counteract such contaminating influences. This report describes a speaking test given by several teachers at Aomori Public College. The extent to which three types of error variance affected the results of this test is examined.