Is the oral interview superior to the group oral?

Gary J. Ockey
International University of Japan

Abstract

As performance-based speaking tests become more prevalent, there is a growing need to examine different test formats for different testing situations. This study compares the group oral to the oral interview in two separate test administrations of both tests. Data collection included test scores assigned by two trained raters, a classroom rating assigned by a trained rater who was teaching the class, student questionnaires, rater comments, and informal interviews with students. Classical test analysis and many-facet Rasch measurement were utilized to analyze the data. Little difference was found in regard to test reliability and neither the students nor the teachers showed a clear preference for which test they thought was more effective. However, test ratings as compared to classroom ratings favored the oral interview as a more effective test. It is suggested that the latter finding might be because some weaker students can mask their weaknesses on the group oral by controlling the conversation. The paper also provides insights into some of the problems which arise in performance-based speaking tests and makes a case for the utilization of many-facet Rasch measurement to help reduce some of the subjective factors inherent when a performance is assessed by multiple raters.

Key words: performance testing, speaking, many-facet Rasch

INTRODUCTION

Performance-based speaking tests are becoming more prevalent in language testing as efforts to utilize tests which are in line with communicative teaching methodology increase. As a result, there is an increasing need to consider different types of performance tests and to determine which might be more appropriate in a given situation. In this study, two well-known performance-based speaking tests, the oral interview where one examinee is interviewed by an examiner, and the group oral where a group of students discusses a topic without the intervention of an examiner, were compared to see which might be more appropriate for assessing the proficiencies of a group of graduate students in an intensive English program.

While there has been a great deal of research conducted on the oral interview and its variants (e.g., Nevo & Shohamy, 1984; Bachman & Savignon, 1986; Shohamy et al., 1986; Van Lier, 1989; Stansfield & Kenyon, 1992; Ross & Berwick, 1992; Young & Milanovic, 1992; Shohamy, 1994; Bachman, et al., 1995; Lazarton, 1996; McNamara & Lumley, 1997; Lynch & McNamara, 1998; Kormos, 1999), little research has been reported on the group oral (with the notable exceptions of Liski & Puntanen, 1983; Nevo & Shohamy, 1984; Hilsdon, 1995; Fulcher, 1996; and Bonk, Ockey, & Iishi, 1998). Consequently, the amount of research which has compared the two test formats is also very limited.

Research background

As for comparison of the practicality of the two test formats, it is quite clear that the group oral is more practical than the oral interview. While only one student can be tested at a time by an examiner with the oral interview, in the group oral it is possible for an examiner to test three or more students simultaneously. Folland and Robertson (1976), two of the first researchers to advocate the group oral, claim that the group oral is a practical method of testing oral proficiency. They point out that the test is relatively cheap when compared to other forms of oral assessment and that raters do not get tired because they only have one task, rating students; they are not required to ask questions or control the test as in the oral interview (161).

A few researchers have compared the reliabilities of the group oral and oral interview. Folland and Robertson (1976) point out that the group oral has the advantage of consistency in the test situation, meaning that examiners are more consistent in their administration of the group oral. This is not surprising since for the group oral, the examiner is only required to give the students cards with written prompts whereas for the oral interview, the examiner needs to ask questions and talk to the examinee to elicit the discourse. On the other hand, Shohamy, Reves, and Bejarano (1986) found that the oral interview leads to higher inter-rating reliabilities than the group oral, 0.91 as compared to 0.73. Unfortunately they provide no explanations for this outcome.

A few researchers have compared the validity of the two test formats. For instance, Nevo and Shohamy (1986) had 16 language testing experts compare the oral interview and the group oral in regard to accuracy, feasibility, fairness, and utility standards. The oral interview was rated higher in regard to accuracy, feasibility (reliable, objective, secure), and fairness. On the other hand, the group oral was rated higher in regard to utility standards (serves the practical information needs of the audience). Regarding concurrent validity, Nevo and Shohamy (1986) correlated the results of the tests to other test scores. When the oral interview was correlated with the group oral, a role play test, and a reporting test, the relationship was about 0.7 whereas when the group oral was correlated with the other three tests, the correlation was about 0.6, suggesting that the group oral tested something different than the other tests.

In regard to other aspects of validity, the group oral may have an advantage over the group oral. Some researchers point out that the oral interview is not a valid test since it consists of questions and answers and is not a real discussion (Van Lier, 1989, Lazarton, 1996). The group oral, on the other hand, appears to be a legitimate discussion. Another

reason the group oral may be more valid is because it more closely matches the classroom practices of small group discussion which take place in many communicative classrooms (Webb, 1994).

Researchers have also considered student perceptions of the tests' validity. Nevo and Shohamy (1984) found that 84% of high school students felt the oral interview reflected their true abilities on a speaking test, whereas, only 51% felt the same way about the group oral. Scott (1986) discovered similarly negative attitudes toward the group oral, reporting that only 32% of students believed the group oral provided an accurate evaluation of their abilities. On the other hand, Fulcher (1996) reports that test-takers thought that the group oral was a more valid form of testing than the oral interview (33). He also reports that students felt the group oral was a more natural test-like situation than the oral interview (29).

Affective factors such as stress may also have an impact on the effectiveness of the group oral and the oral interview. The negative effect of test anxiety is well documented in the English language teaching literature (e.g., Madsen, 1982; Shohamy, 1982; Scott, 1986; Young, 1986). This research, coupled with research which suggests that students feel less stress when taking a group oral than when taking an oral interview (Folland & Robertson, 1976; and Scott, 1986) may be an argument for the group oral.

Taking into account the obvious practicality of the group oral, and that research regarding reliability and validity has produced mixed results, the question arises of whether the prevalence of the oral interview as compared to the group oral is justified.

Research questions

Is the oral interview more reliable than the group oral?

Is the criterion-related validity of the oral interview higher than that of the group oral?

Do students think the oral interview is more valid than the group oral?

Do raters (teachers) favor the oral interview over the group oral?

RESEARCH DESIGN

Setting

Students who enter the Intensive English Program (IEP) at IUJ have been accepted as post-graduate students. Since all content courses are taught in English, the IEP is designed to prepare students to function in English in one of the programs at the University. The IEP provides approximately 175 hours of instruction (which includes about 4 hours of individual tutorial instruction) over a two-month period. The four skills of speaking, writing, listening,

and reading are taught in class sections of between 10 and 13 students. Various teaching styles are employed in the courses, including lectures, small group discussions, role-plays, and student-led activities such as group and individual oral presentations.

Examinees

Students from two IEPs, 1999 and 2000, were considered in the study. There were 45 students in the 1999 IEP, 40 men and 5 women. Twenty-five students were Japanese and 20 were Indonesian. The youngest student was 23 and the oldest student was 42 with most students in their late 20s. TOEFL scores at time of entry ranged from 430 to 603 with most scores between 510 and 590. The overall TOEFL average was 540. (See Ockey 1999 for further details on student proficiency.)

There were 51 students in the 2000 IEP, of whom 13 were women and 38 were men. The students came from the following countries: Japan, 22; Indonesia, 19; Laos, 2; Tanzania, 2; Myanmar, 1; Thailand, 1; Cambodia, 1; Brazil, 1; Malaysia, 1; and Korea, 1. The youngest student was 23 and the oldest student was 36 with an average age of 30.5. The average TOEFL score at the time of entry was 530.5 with a range of 373 to 623; only a few students had scores outside of the target population range of 510 to 590.

Test descriptions

All students take an oral interview as part of the university entrance examination before entering the university and a group oral on the first day of IEP classes. Thus, all students have had experience with both types of tests when they encounter them as part of the course.

The two tests considered in this study took place during the IEP term, one during week four as part of the midterm assessment (the oral interview), and one at the end of the course as part of the final assessment (the group oral). One purpose of the speaking tests is to motivate students to study the content of the course. For this reason, the prompts are based on information which has been read and discussed in class. However, while the prompts for the tests are based on familiar topics (discussed in class), the test is designed to measure proficiency rather than achievement.[i] (See Appendix 1 for the rating scale.) Another purpose of the tests is to act as an assessment tool for grading and placement or exemption. As a result, although the tests may not be considered high-stakes tests, students are highly motivated to do well on them.

Based on the findings of Bonk, Ockey & Iishi (1998) who report that utilizing more than one prompt does not have a significant effect on scores in group speaking tests, multiple prompts were utilized. This prevented students who tested later in the day from gaining

insights into the test (from students who had already taken the test) before it was administered to them.

The oral interview begins with the examiner attempting to relax the examinee by introducing himself and encouraging the examinee to do well on the test. After a couple of simple questions based on content discussed in class to relax the examinee, the examiner asks the examinee a question regarding his opinion about some general issue that was discussed in class. After the examinee responds, the examiner challenges the examinee's opinion by asking questions or disagreeing with him. While the topic is a familiar topic to the students, it is meant to test proficiency (since the student cannot predict the direction of the discussion which takes place). After about 12 minutes of this probing, the examiner asks a relatively simple question before excusing the examinee. The test administration takes about 17 minutes, and a few minutes are needed after the test for the examiner to assign a score to the student. (See Appendix 2 for an example of a set of prompts utilized in the oral interview.)

In the group oral, students are grouped with two other students from their same class. They are not aware of their grouping, however, until an examiner invites them into the testing room. No group contains students who have the same first language and the groupings consist of students with high, medium, and low proficiency (though neither raters nor students are made aware of this grouping strategy) based on classroom ratings given prior to the test. An examiner sits outside of the group, provides ratings while the discussion takes place, and does not participate in the discussion. To begin the test, the examiner introduces himself and asks the examinees to do likewise. Then the examiner gives the students a card which contains a simple prompt based on a topic they have discussed in class. After approximately 5 minutes, the examiner stops the conversation. In similar fashion, the students then have about 12 minutes to discuss a general issue related to a topic discussed in class. It is on this second question that students are assigned proficiency ratings. Again, although the topic is familiar to the students, it is meant to test proficiency. Three students can be tested in about 20 minutes. (See Appendix 3 for an example of a set of group oral questions.)

Rating scale

The same discrete point rating scale, an 11-point scale where scores are assigned from 0-10 with half point steps in each of six categories was employed for both tests (see Appendix 1). The categories include comprehensibility, fluency, grammar, vocabulary, and communicative strategies.[ii]

Data collection

One day before each test, classroom teachers rated their own students in an informal classroom setting.[iii] In the test situation, each student was scored by two raters; one score was given in real time and one was provided based on the viewing of a videotape of the test. After each test, students were asked to fill out a questionnaire regarding the test format.[iv] Informal interviews with individual students took place within a few days of the test. In addition, raters were asked to comment on the strengths and weaknesses of test formats after each test.

Raters and rater training

All raters had teaching experience in the English language teaching field and post-graduate training in applied linguistics (or a related field). In the 1999 test administrations, there were four raters, one American man, one British man, and two American women. In the 2000 administrations, the two men and one of the American women were joined by two new raters, both American men.

There were a number of phases in the rater training. The first phase involved the raters watching videos of students taking tests. In this two-hour training session, raters were introduced to the rating scale and given practice rating students. The raters then administered a group oral to the students in week 1, providing them with further practice with the rating scale. After the test, the test coordinator distributed and discussed a rater report which included an explanation of individual rater severity and consistency as measured by many-facet Rasch.[v] Prior to both the oral interview and the group oral, raters were given training on how to administer each test and further practice in rating students by viewing videotapes.

Student instructions

The students were given specific instructions on how to do their best on the test and how to make it fair for all examinees. For example, in the group oral, students were encouraged to try and share time equally in the context of a natural discussion; dominating the conversation was not considered an appropriate communication strategy and would result in a low score. For the oral interview, students were encouraged to clarify statements, ask questions, and challenge the interviewer as they did in regularly scheduled individual tutorials with their classroom instructors.

RESULTS AND DISCUSSION

The general statistics of each test and classroom-rating situation can be seen in Table 1.

Table 1 General statistics of test and classroom ratings

	Number of students	Range of scores	Mean	Standard Deviation
Oral Interview				
1999 Test	45	19.5 – 41.5	31.5	5.8
1999 Class	45	9.0 – 43.0	30.6	7.9
2000 Test	51	19.4 – 45.2	30.6	6.8
2000 Class	51	19.0 – 46.0	30.4	6.2
Group Oral				
1999 Test	45	27.0 – 43.0	35.6	4.1
1999 Class	45	19.0 – 44.0	33.6	6.4
2000 Test	50*	16.0 – 45.0	33.8	5.8
2000 Class	51	20.0 – 47.0	32.4	6.4

Scale is from 0-50

* One student had a family emergency and could not take the test.

There are two findings that stand out in regard to the general test statistics. First, the mean of the group oral is higher than the mean of the oral interview in both test administrations. This is not unexpected, however, since four weeks of intensive English training took place between the test administrations. Second, the standard deviations are quite different from test to test and situation to situation. This is most apparent in the 1999 group oral where the standard deviation of 4.1 is relatively small compared to the other standard deviations. Furthermore, in all cases (except the 2000 oral interview), the standard deviation of class ratings is larger than that of test ratings, possibly because teachers are more confident in assigning ratings to students they teach. It could be that when teachers are not sure about a student's proficiency, they tend to err on the safe side by assigning a score more in the middle of the scale.

Practicality

Not surprisingly, the 1999 and the 2000 tests results showed that the group oral was much more practical. The group oral required about 13 minutes of teacher time per student whereas the oral interview took approximately 40 minutes of teacher time per student—three times as much. In regard to training to administer the tests (not including the time it took to norm raters at the beginning of the term), 15 minutes was spent for the group oral as compared to 2 hours for the oral interview. It should also be pointed out that it was evident from the videos and raters' comments that the two-hour training session for learning to administer the oral interview was not nearly enough to prepare examiners to effectively administer the test.

Reliability

A Pearson correlation and a simple coefficient agreement formula were applied to measure the inter-rating reliability. Pearson was employed because it is widely used to correlate data on an interval scale. A simple coefficient agreement formula was utilized because it is more sensitive to small rating differences and it reveals the percentage of student scores which show tolerable differences. The results can be seen in Table 2.

Table 2 Inter-rating reliability of the oral interview and group oral

	Oral Interview	Group Oral
1999 Test (N=45)		
Pearson Correlation	0.64	0.63
Simple Coefficient Agreement	0.67	0.78
2000 Test (N=50)		
Pearson Correlation	0.75	0.47*
Simple Coefficient Agreement	0.76	0.45*

The simple coefficient agreement reliability is based on ratings being within 5 points of each other.
*A number of problems in this administration make a comparison unreasonable. The correlation is based on only 38 scores due to the failure of one video camera.

The inter-rating reliability is disappointingly low for the oral interview and group oral in both years of the study. As for comparison of the two tests, in the 1999 IEP, the Pearson correlation was almost the same for both tests, 0.64 for the oral interview and 0.63 for the group oral. Based on the simple coefficient agreement results, a higher percentage of students were within an acceptable range[vi] on the group oral than the oral interview, 78% compared to 67%.[vii] Thus, the results for the 1999 test reveal little difference in inter-rating consistency.

In the 2000 IEP, the inter-rating reliability could not be compared due to two problems. First, the failure of one video camera resulted in 12 examinees receiving only one score on the group oral. Second, one rater admitted to giving inflated ratings on the group oral because he thought that students should receive higher scores on the final test than on the midterm test.

In regard to the oral interview, there were some obvious inconsistencies in the way the tests were administered. For instance, on the 2000 oral interview, one interviewer cut most tests to six or seven minutes (rather than approximately 17). The question designed to evaluate the students' proficiency was asked with almost no follow-up questions (1-2 minutes as compared to the expected 12 minutes of probing). In other cases, it was clear that some interviewers challenged students with serious probing questions while others appeared to try to help the students perform well by making the follow-up questions easy. In regard to the group oral, on the other hand, there did not appear to be any serious differences in the

way the tests were administered. It is, thus, apparent that the oral interview did not prove to be superior to the group oral in terms of reliability.

Many-facet Rasch measurement

In order to make the test scores more fair for students (especially since one rater admitted to purposefully inflating scores on the 2000 group oral) and to compare the criterion-related validity of the two tests, many-facet Rasch with the program 3.0 (Linacre, 1996) was utilized to analyze the data. The Rasch model is based on probability and allows one to estimate and correct for the effects of different variables in a test by separating and placing them on a common logit scale.[viii] The rating scale model of Facets was applied in this analysis.[ix] Because it was discovered in interviews with raters after the tests that raters did not actually treat the scale categories (pronunciation, fluency, grammar, vocabulary, and communication strategies) separately, the scores in each category were combined before analysis.[x] The model for the analysis was:

$\text{Log}\,(P_{nijk}/P_{nijk-1}) = B_n - S_i - C_j - F_k$

P_{nijk} is the probability of student n being awarded in situation i by rater j a rating of k.

P_{nijk-1} is the probability of student n being given in situation i by rater j a rating of k-1.

B_n = ability of student n

S_i = difficulty of situation i

C_j = severity of rater j

F_k = difficulty of the step from category k-1 to category k

An example of Facets output (for the 1999 oral interview) which provides a visual comparison of the variables can be seen in Figure 1. The first column shows a logit scale which is a common scale calibrated to a mean of 0. The second column represents the examinees (students) with those at the top of the scale (with a logit measure above 0) more proficient than those at the bottom of the scale (with logit measures below 0). The third column represents the test situation, in this case a rating assigned in real time (live) or a rating assigned from viewing a video (video). A rating high in this column (above a logit measure of 0) suggests that the situation makes it more difficult for a student to get a higher rating while a rating below 0 logits suggests that this situation makes it easier for a student to attain a high rating on the test. The fourth column represents raters who have been assigned one of the numbers shown. A rater high in the column (has a logit measure above 0) assigns ratings which are severe whereas a rater at the bottom of the column (has a logit measure

below 0) is a lenient rater. The fifth column represents the scale used in the study. The scale shows the score a student would achieve if his rater was one of 0 logit severity (a rater who is not lenient or severe) and his situation was of 0 logit difficulty (situation was not difficult or easy) (Linacre, 1994: 6). For example, in Figure 1, a student with a logit measure of 0 (an average examinee) would be expected to score 32 on the 50 point rating scale.

Figure 1 All Facets Vertical Ruler for 1999 oral interview

Logit	examinee	situation	rater	Scale
3				(43)
	*			42
				---
	*			41
2	**			40
	*			---
	*			39
	*			38
1	****			37
	**			36
	****		2	35
	****		1	33
0	*******	live video	3	32
	*****			30
	*****			28
			4	26
-1	*			24
				23
	***			22

-2				21
	**			
				---
	*			
-3				(18)

In addition to providing a visual representation of the comparison of the variables in the data, Facets can provide a *fair score,* referred to as a measure, which takes into account the differences in variables considered in the data. Thus, assuming the rater variable to be the only variable considered by Facets, if an examinee gets a severe rater, the program considers this in the analysis and adjusts the score up to a measure closer to the average rating he would have received if he had been rated by all of the raters. An example of differences in rater severity can be seen in Figure 1. It can be seen that rater 2 is the most severe (since he is higher in the column) and rater 4 is the most lenient (since he is the lowest in the column). Even though each student is judged by two raters,[xi] a student who is judged by raters 1 and 2 is at a disadvantage when compared to a student who is judged by raters 3 and 4. In all four

test administrations, there was a significant difference in rater severity (see Appendix 4). For example, in the 1999 oral interview, the error-corrected standard deviation of the raters (separation) is 4.32 times the root mean-square estimation error. The reliability of this separation in ratings is 0.95.[xii]

In this analysis, FACETS was also utilized to see whether there was a difference in ratings given during the test (live) and ratings given from watching a video of the test (video). No difference was found in any of the test administrations; in all cases, separation was 0.0. This lack of difference (for the 1999 oral interview) can be observed in Figure 1 where both situations line up on 0 logits.

If students are to be given fair scores, the importance of utilizing a program such as FACETS which adjusts for some of the error resulting from differences in rater severity is evident. In this analysis, it was especially important considering the low inter-rating reliability and the fact that one rater admitted to inflating scores on the 2000 group oral.

Criterion-related validity

As one way of assessing the validity of the two tests, classroom ratings were compared to test performance.[xiii] A classroom teacher assigned a rating to the students in his class on a similar task the day before the test. Teachers were instructed to consider what they knew about the students when they rated them; for instance, if a student performed worse than expected, the teacher could consider this when assigning a rating.

In regard to the oral interview, when the group of students is taken as a whole, there proved to be no difference between measures on the test and scores given in the classroom situation. Both the error-corrected standard deviation of measures (separation) and chi-square as can be seen in Table 3 confirm this.

Table 3 Comparison of classroom rating and test rating

Test	Separation	Reliability	Chi-square	Significant Difference
Oral Interview				
1999	0.0	0.00	1.0	No
2000	0.0	0.00	0.4	No
Group Oral				
1999	2.6	0.87	15.5	Yes
2000	1.9	0.78	9.0	Yes

P<.01 with 1 degree of freedom (for chi-square)

In regard to the group oral both in the 1999 and 2000 tests, when the group of students was considered as a whole, there was a significant difference between measures on the test and ratings given in the classroom situation. In both cases, test measures were significantly higher than classroom ratings. This difference is confirmed by both the error-corrected standard deviation of measures (separation) and chi-square as can be seen in Table 3. This information suggests that at least some of the students performed better than expected. When the scores were looked at more closely, it was noticed that some of the low-level students did better than predicted by their classroom ratings while, in general, the other students performed about as expected.

A possible explanation for the reason some low level students achieved better scores on the group oral than expected emerged from individual interviews with these students. These students confided that they thought it was because they were able to impress raters by controlling the conversation, meaning they were able to mask their weaknesses. Their strategy was to encourage others to speak and appear to be actively involved in the discussion without saying much. A review of the videos also suggested that this might have been the case. There are other reasonable explanations for this finding such as possible positive effects of participating with higher level students in a group.[xiv] In any case, this is certainly an area that needs further consideration. If the group oral is not effective at accurately measuring the abilities of low level students, it is certainly a limitation of the test.

Student perceptions

Students were asked to respond to a five point Likert scale with 1 meaning strongly disagree and 5 strongly agree. A positive response was denoted by a student selecting 4 or 5 and a negative response was denoted by a student marking 1 or 2. The results of two of the items on the student questionnaires can be seen in Table 4.

The students reported that they believed that the test formats were almost equally effective at measuring their English proficiency, both in the 1999 tests and in the 2000 tests. In response to the question: "The test was able to provide an accurate measure of my speaking ability," in 1999, 52.2% of the students responded positively about the oral interview as compared to 51.1% of the students for the group oral. Interestingly, the tests compared almost equally again in 2000 with 76.5% of the students rating the oral interview positively as compared to 79.6% for the group oral. While students were biased against the group oral in earlier research (Shohamy, 1984; Scott, 1986), these findings coupled with Fulcher's (1996) more recent findings might suggest a softening of this bias. This might reflect a changing attitude of students toward group work and collaboration as a valid means

of testing (just as it has become considered a valid means of learning). The results suggest (since the tests show similar attitudes both years, but marked differences from year to year) that the effectiveness of the test depends more on the test situation itself than the test format. In this case, it may be the attitudes of students toward the effectiveness of tests in general overshadow their concerns about the fairness of a particular test format.

Table 4 Results of two questions on student Questionnaires

Question	Test	1 SD	2 D	3 N	4 A	5 SA	Negative Responses	Positive Responses
	Oral Interview							
The test was able to provide an accurate measure of my speaking ability.	1999	0	7	14	21	1	7 (16.3%)	22 (52.2%)
	2000	0	2	10	33	6	2 (3.9%)	39 (76.5%)
	Total	**0**	**9**	**24**	**54**	7	**9 (9.6%)**	**61 (64.9%)**
	Group oral							
	1999	0	4	18	20	3	4 (8.9%)	23 (51.1%)
	2000	1	3	6	34	5	4 (8.2%)	39 (79.6%)
	Total	**1**	**7**	**24**	**54**	**8**	**8 (8.4%)**	**62 (66.0%)**
	Oral Interview							
The test made me feel nervous while I was taking it.	1999	2	9	13	12	6	11 (25.6%)	18 (41.9%)
	2000	1	8	23	12	7	9 (17.6%)	19 (37.3%)
	Total	**1**	**17**	**36**	**24**	**13**	**20 (22.0%)**	**37 (40.7%)**
	Group oral							
	1999	4	18	13	8	2	22 (48.9%)	10 (22.2%)
	2000	1	14	15	12	6	15 (31.3%)	18 (37.5%)
	Total	**5**	**32**	**28**	**20**	**8**	**37 (39.8%)**	**28 (30.1%)**

The scores are based on the following Likert scale:
1(SD) = Strongly Disagree 2(D) = Disagree 3(N) = Neutral 4(A) = Agree 5(SA) = Strongly Agree

In regard to test anxiety, the results (see Table 4) show that more students felt the oral interview made them feel nervous than the group oral. Considering the negative responses (since this is a question where a negative response suggests that the students did not feel nervous), it can be seen that in the 1999 test 25.6% of the students did not feel nervous when taking the oral interview while 48.9% reported that they did not feel nervous when taking the group oral. The results for the 2000 test were similar with 17.6% claiming the oral interview did not make them feel nervous as compared to 31.3% for the group oral. These results support the findings of Folland & Robertson (1976) and Scott, (1986) who found that students do not feel as nervous when tested in groups as compared to when tested alone. Coupled with previous research which shows that students perform better when they are less nervous (Madsen, 1982; Shohamy, 1982; Scott, 1986; and Young, 1986), these findings indicate that the group oral may have an advantage over the oral interview in this respect.

Rater perceptions

Rater (teacher) comments on the strengths and weaknesses of each test are reported in Table 5.

Table 5 Rater comments on comparison of oral interview and group oral

The Oral Interview:
• makes it easier to test a specific aspect of a student's ability.
• allows for a systematic probing of each student's own strengths and weaknesses.
• creates a situation where all students can show their true ability.
• gives us more (not necessarily better) discourse.
♦ results in teachers administering the test in different ways.
♦ takes a lot of time.
♦ is too easily influenced by the examiner.
• is not as spontaneous as the group oral.
• puts a great deal of pressure on the student.
• requires thorough interviewer training.
The group oral:
♦ gives us a better picture of the students' communication strategies.
♦ takes much less time to conduct.
♦ is probably fairer in that it eliminates the bias when the rater and the interlocutor are the same person.
♦ is good for less confident students who need time to think.
• provides a sample of discourse which is not subjectively influenced by different teachers.
• is much easier to administer.
• requires a wider range of skills.
• is a more natural conversation.
• results in positive instructional washback.
• could result in one student affecting the score of another.
• can be frustrating for more confident students.
• results in students not getting equal time to speak.

♦ Mentioned by more than one rater

While raters provided both positive and negative responses to both test types, their comments seem to be more positive toward the group oral than the oral interview. Most notably, they think that the group oral is less biased by the rater and tells more about a student's communication skills. On the other hand, the oral interview seems to be considered a more accurate test based on such comments as, "Allows for systematic probing of students' strengths and weaknesses" and "Creates a situation where students can show their true ability". In regard to negative comments on the oral interview, three comments stand out. The first refers to the larger workload, an issue of practicality and the second deals with the lack of uniformity of administration, an issue of test reliability; both have already been discussed as weaknesses of the oral interview. The third involves the effect of the examiner on a student's score which is analogous to the only strongly negative comment on the group

oral, the possible effects of one student in the group on the score of another. The issue of an examiner or another student affecting the score of a student is of obvious concern on both tests, but there is no evidence that the concern is greater where another student or the examiner is the factor. Thus, raters seem to favor the group oral over the oral interview in this study.

CONCLUSION

This study confirms earlier studies which show that the group oral is more practical than the oral interview. In regard to inter-rating reliability, little difference was found in the two formats for the 1999 test (and the for the 2000 test inter-rating reliabilities could not be compared) whereas the group oral was administered more consistently. Also, inter-rating reliabilities were quite low which suggests a lack of thorough rater training, a problem that will likely continue in future test administrations. This underscores the importance of the utilization of a program such as FACETS (a many-facet Rasch program) which can help to make scores more fair by minimizing the effects of differences in rater severity.

Test ratings as compared to classroom ratings, a possible method for measuring criterion-related validity, favored the oral interview. Some students achieved higher scores than predicted by their classroom performances on the group oral, possibly because low-level students are able to mask their weaknesses. This finding warrants further investigation. If it does prove to be true, ways of alleviating this problem, such as thorough rater training and student instructions for how to take the test, would need to be employed if the test is to be utilized.

The students did not indicate that they thought the oral interview was more effective at measuring their proficiencies, nor did raters (teachers) consider the oral interview to be a better tool for rating performance. This may be evidence that the group oral is gaining acceptance from both teachers and students. Moreover, students reported nervousness more on the oral interview, suggesting it might be less effective than the group oral in this regard.

This study provides no clear answer to the question of whether the oral interview format is superior to that of the group oral format. It does, however, point out that in many ways the oral interview may *not* be a more effective test than the group oral. More importantly, the study offers some insights into the strengths and weaknesses of each test format and some of the obstacles to overcome when conducting performance-based speaking tests.

Notes

[i] The test might not be considered a proficiency test in the strictest sense like that of a test where students are asked questions in a number of different topic areas.
[ii] This rating scale was piloted and revised based on a similar population of students.
[iii] The classroom teachers were all trained raters and utilized the same rating scale that was used for test administrations. Each teacher rated all of his students. Test raters did not rate any students that were in their own class.
[iv] The students were not aware of their test results when they filled out the questionnaires. These questionnaires had been piloted and revised based on students' responses in previous programs.
[v] See Weigle (1998) for a discussion of the effects of Rasch feedback on rater training.
[vi] Chosen to be 5 points on the 50-point scale because there are five categories on the scale and it seemed reasonable to accept a 1 point difference in ratings in each of the categories.
[vii] The fact that the simple coefficient agreement is higher than the Pearson correlation may be due to the fact that the standard deviation on the test was quite small. This would tend to increase the number of inter-ratings that are tolerable. This may suggest that the group oral is not effectively separating students by ability.
[viii] For a good explanation of the theory behind the analysis see *McNamara, 1996, Measuring Second Language Performance.*
[ix] It should be mentioned that rather than combining the categories and utilizing the rating scale model, it might be more appropriate to treat the five categories separately and employ the partial credit model. See Bonk, Ockey, & Iishi (1998) for an explanation of the differences between the partial credit model and the rating scale model of analysis and which may be more appropriate
[x] Students received a *fair* total score based on this combining of categories and scores in each category based on each category being analyzed separately.
[xi] Only two raters will result in a great deal of misfit and error as measured by many-facet Rasch because there are only two data points to connect raters. See Bonk, Ockey, & Iishi (1998) for an explanation regarding this issue.
[xii] See Lunz, Wright, & Linacre (1990) for a further explanation.
[xiii] This might be one way to get at criterion-related validity. Students are in their normal learning environment working on "natural" learning tasks. The ratings given by teachers were meant to represent the students' actual abilities (as judged by the teacher who worked with the students a minimum of 12 hours for 8 weeks in class sizes of 10-13 students).
[xiv] See Webb (1994) page 17 for further explanation.

REFERENCES

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*, 238-257.

Bachman, L. F. & Savignon, S. (1986). The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *The Modern Language Journal, 70*, 380-390.

Bonk, W., Ockey, G. & Ishii, D. (1998). An IRT study of a group oral proficiency test. Paper presented at the Japanese Association of Language Teachers, Omiya, Japan.

Folland, D. & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal 30*, 156-167.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing, 13*, 23-51.

Hilsdon, J. (1995). The group oral exam: advantages and limitations. In Alderson, J. and North, B. *Language testing in the 1990s: the communicative legacy*. Hertforshire: Prentice Hall International. 189-197.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessments: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*, 163-188.

Linacre, J. (1994). Many-Facet Rasch Measurement. Chicago: Mesa Press.

Lincare, J. (1996). Facets 3.0 [Computer Program]. Chicago: Mesa Press.

Liski, E. & Puntanen, S. (1983). A study of the statistical foundations of group conversation tests in spoken English. *Language Learning, 33*, 225-246.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing, 13*, 151-172.

Lunz, M., Wright, B., & Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*, 331-345.

Lynch, B. K. & McNamara, T. F. (1998). Using g-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158-180.

Madsen, H. (1982). Determining the debilitative impact of test anxiety. *Language Learning, 32*, 133-143.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Addison Wesley Longman Limited.

McNamara, T. F. & Lumley, T. (1997). The effect of interlocutor and assessment mode variables in overseas assessments of speaking skills in occupational settings. *Language Testing, 14*, 140-156.

Nevo, D. & Shohamy, E. (1984). Applying the joint committee's evaluation standards for the assessment of alternative testing methods. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, L.A. (ED. 243 934).

Ockey, G. (1999). The use of TOEFL to measure a change in proficiency. *Working Papers on Language Acquisition and Education: International University of Japan, 10*, 1-12.

Ross, S & Berwick, R. (1992). The discourse accommodation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 159-176.

Scott, M. (1986). Student affective reactions to oral language tests. *Language Testing, 3*, 99-118.

Shohamy, E. (1982). Affective considerations in language testing. *Modern Language Journal, 66*, 13-17.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11*, 99-123.

Shohamy, E., Reves, E. & Bejarno, Y. (1986). Introducing a new comprehensive test of oral proficiency. *ELT Journal, 40*, 212-220.

Stansfield, C. & Kenyon, D. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System, 20*, 347-364.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Webb, N. (1994). Group collaboration in assessment: competing objectives, processes, and outcomes. CSE technical report 386. National Center for research on evaluation, standards, and student testing (CRESST). University of California at Los Angeles.

Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing, 15*, 263-287.

Young, D. (1986). The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals, 19*, 439-445.

Young, R & Milanovic, M. (1992). Discourse validation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 403-424.

Appendix 1

Descriptor Bands for Speaking Tests

	Comprehensibility	Fluency	Grammar	Vocabulary usage	Communicative Skills
10	Rarely mispronounces, able to speak with nearly perfect pronunciation, intonation, and rhythm, little or no foreign accent	Fluent speech, speaks confidently and effortlessly, speech is smooth and natural	Uses high level discourse with near perfect accuracy, shows an ability to use the full range of grammatical structures effortlessly and accurately which are needed to achieve the task	Confidently uses wide range of technical and general vocabulary precisely and effectively	Shows confidence and naturalness, shows ability to negotiate meaning, shows how ideas or opinions are related, may initiate conversation, completes task effectively
9 8	Pronunciation is clear, occasionally mispronounces or has non-perfect intonation or rhythm, articulation is clear, has mastered all sounds, accent may sound foreign, but does not interfere with understanding	Speaks with confidence, but has a few unnatural pauses, occasionally gropes for words unnaturally	Shows ability to use nearly the full range of grammatical structures, but may make some errors when using some complex sentence types, errors do not interfere with meaning	Shows range of technical vocabulary which is sufficient for task, but fine shades of meaning are occasionally inappropriate	Generally confident, responds appropriately to an opinion, shows ability to negotiate meaning, shows how ideas are related, completes task effectively
7 6	Pronunciation is not perfect but can be understood without concentrated listening, articulation is generally clear, may not have mastered all sounds	Speech is a little hesitant, has some unnatural rephrasing and groping for words	May not have mastered full range of structures, but uses complex and simple sentences, may make a few global errors, has no trouble completing task	Has sufficient vocabulary to complete task, but may not use it appropriately, may use technical vocabulary, but not always effectively	Somewhat confident, responds appropriately when asked for opinion, completes task somewhat effectively
5 4	Sometimes mispronounces, may require concentrated listening, but is completely understandable, may not articulate clearly, may not have mastered some sounds	Speech is often hesitant, frequent unnatural rephrasing and groping for words,	May use simple (but generally accurate) sentences to express meaning, complex sentences are used but often inaccurate, can express desired meaning, errors may occasionally interfere with meaning, is able to complete task	Vocabulary is adequate for achieving task, but often used inappropriately. Does not accurately use technical terms used in the field	Not confident, shows agreement or disagreement to opinions at the surface level but not at the discourse level, completes task but not effectively
3 2	Frequently mispronounces, accent impedes comprehensibility, requires concentrated listening but is generally comprehensible	Strained speech, often groping for words, some long unnatural pauses (except for routine phrases)	Relies mostly on simple sentences which are often inaccurate, cannot control complex sentences, mistakes often impede meaning, has difficulty completing task	Lacks the necessary vocabulary to discuss the topic with any sophistication	May use simple phrases to show agreement or disagreement, but does not relate ideas at discourse level, task may not be completed
1 0	Frequently mispronounces, heavy accent, even with concentrated listening often incomprehensible	Fragmented speech that is so halting that conversation is virtually impossible	Cannot control even simple sentences, grammar is not sufficient to complete task	Vocabulary is inadequate to achieve the task	May require prompting, produces monologues which are unrelated, does not complete task

A score in the lower part of the box indicates that a student has not completely mastered the level.

Appendix 2

Example of a set of oral interview prompts

First text-based question
Please tell me what happened to Makiko.

Second text-based question
According to the articles, what are some reasons plagiarism is considered okay in some cultures?

Opinion-based question
Do you think plagiarism is an idea which represents western values and has no relevance in Asia?

Appendix 3

Example of a set of group oral prompts

Text-based question
Identify some of the differences and similarities between the Usagi Motor Case and the Fitzburg Tire Company case. Justify your claims.

Opinion-based question
Discuss how you would solve some of the problems in the Security First Bank Case.

Appendix 4

Rater severity as measured by Facets

	Separation	Reliability
Oral Interview		
1999 Test	4.32	0.95
2000 Test	3.22	0.91
Group Oral		
1999 Test	2.65	0.88
2000 Test	4.79	0.96