The validity of the paired interview format in oral performance assessment[1]

Noriko Iwashita
Language Testing Research Centre
The University of Melbourne

Abstract

Studies of oral test discourse have been mainly concerned with interaction between native speaker and non-native speaker (eg Lazaraton 1992; Ross 1992), but there is little research on interaction between non-native speaking candidates in a paired oral interaction test. Not much is yet known about variations in the quality and quantity of language which is produced by non-native candidates interacting with other non-native candidates or about the impact of this variation on test scores and its implications for test fairness.

In relation to the paired oral interaction assessment, the present study reports on the result of a small-scale pilot study addressing the following research questions: 1) Do the test-takers' scores differ in relation to the proficiency of the speaking partner? 2) Does the test-takers' discourse differ according to the proficiency of the speaking partner?

The data is drawn from performances in a task-based oral interaction test. 20 candidates undertook the test twice, once with a partner of the same proficiency level and once with a partner at a different proficiency level. Each interview was rated twice for both candidates. The tapes were transcribed and an analysis of specific discourse features was carried out. Candidates were also asked to complete a questionnaire eliciting their reactions to the two test conditions.

[1]An earlier version of this paper was presented at the annual Language Testing Research Colloquium at Orlando, Florida in March 1997. This research was conducted with the assistance of the MAK Halliday Scholarship. Thanks are due to Cathie Elder, Tim McNamara, Lis Grove, Annie Brown and other colleagues in NLLIA-LTRC for comments on earlier drafts of the paper.

Melbourne Papers in Language Testing 1996 Volume 5.2 pp. 51-66. The Language Testing Research Centre, The University of Melbourne.

The findings of the analysis revealed that while the proficiency of non-native speaking interlocutor has some impact on the amount of the talk, there was little difference in test scores across the two testing formats. However, test-taker feedback suggests that candidates prefer the NNS-NNS interaction mode to the NS-NNS mode as they find it less threatening. The study has implications for testing in general and for classroom assessment of foreign language learners in particular.

1. Background to the study

Communicative teaching methodology has enhanced opportunities for learners to practice speaking skills using communication tasks, and, as a result, speaking skills have generally been included as a part of the assessment of language courses. Compared with the assessment of other skills (eg reading, writing, listening) assessment of the speaking skill is time-consuming as it is usually administered individually by an interviewer/instructor. In many tertiary institutions in Australia, the enrolments in foreign language courses at beginners' level are large and instructors try to seek an assessment procedure which is less time-consuming as well as being fair. Fulcher (1996) reports on a number of cases in which group testing is being used successfully in Israel (Reeves 1991; Shohamy, Reeves and Bejarano 1986), in Zambia with school students (Hilsdon 1991), and in Hong Kong with university students (Morrison and Lee 1985; Berry 1995). Fulcher himself has also investigated test-taker feedback on group orals through questionnaire data and retrospective reports and argues that group orals give students more confidence than having to respond to an examiner. Berkoff (1985) also supports group orals for being a natural mode of communication and for engendering a low anxiety rate. The format of group orals which many of the above studies above report on is the group discussion, but in the present study the format is a paired interview in which a non-native speaking learner speaks with another non-native speaking learner.

The studies above all report on test-takers' positive feedback on the group oral as being natural and non-threatening. For learners in the foreign language situation where the target language is not the main medium of communication, learners mostly practise with other learners and this is therefore a natural mode of communication for

them. It is to be expected that learners' anxiety will be lower in such a situation.

Despite the advantages of the group oral as more natural, less threatening and more economical than individual interviews, there are a number of problems which need to be addressed in the interest of fairness and validity. First of all, it is very hard for assessors to predict the quality and quantity of the language which test-takers are going to produce. In the paired-interview format, the person with whom the candidate is paired may affect their approach to the task. In SLA studies (eg Gass and Varonis 1985, 1986; Iwashita, 1983; Pica, Holliday, Lewis, and Morgenthaler, 1989; Porter, 1986) interlocutor variables (gender, proficiency, ethnicity and native/non-native speaker) have been studied in relation to the amount of interactional modification when communication breaks down. There is substantial evidence that the amount of interactional modification varies according to the type of interlocutor.

Furthermore, in the field of language testing there have been a number of empirical studies (eg Ross 1992; Ross and Berwick 1992; Lazaraton 1993) which have investigated test discourse in relation to interview behaviour. Ross and Berwick (1992), for example, examined the extent of interviewer accommodation, and found that interviewers tend to accommodate to the performance of candidates at the lower end of the proficiency continuum more than they do with candidates at the top end. Studies by Berry (eg. 1993) investigate the relationship between learner characteristics and performance and have given some evidence that differences in learner personality could affect performance on paired oral tests.

Though positive reactions to group orals on the part of test-takers have been reported, it is not certain whether the group oral is a fair mode of assessment. To date, empirical studies have used only one source of data to examine the issue, but in order to investigate whether group oral assessment is a fair measure of assessment for foreign language courses, multiple sources of data are required.

Based on a review of the research in language testing and second acquisition research, two research questions are addressed in the present study.

1) Do test-takers' scores differ in relation to the proficiency of the speaking partner?

2) Does the test-takers' discourse differ according to the proficiency of the speaking partner?

2. Research design

2.1 Subjects

The subjects for this study are learners of Japanese at a tertiary institution in Australia who had received approximately 250 hours of formal instruction when the data were collected. They are all female and approximately 20 years old.

The first language of most subjects is English except for a few bilingual speakers of Chinese and English. Some subjects have spent an extensive period of time (eg. more than three months) in Japan, while others have never been to Japan. Subjects were recruited in class by instructors and volunteered to participate in the study. They were paid for their participation in the study, and written feedback on their performance was also given on request.

The subjects were divided into three groups according to their proficiency based on the result of the initial interview with the researcher, the length of their in-country experience, and a verbal report given by the course coordinator. As a result, the task performance of subjects who belong to two groups, those at the top and bottom ends of the proficiency scale, were used in the present study.

The ten subjects at the High proficiency level have all spent an extensive period of time in Japan (more than one year) and are students who were placed in the top 10% of the course. The ten subjects at the Low proficiency level have not spent extensive time in Japan and are rated below average in the course. Initially, data from these 20 subjects were collected, but due to failures in recording, data for only 17 subjects (8 high, 9 low level subjects) were analysed in the present study. Detailed information about each subject is shown in the Appendix.

2.2 Method

All subjects were asked to attend two sessions in which they were required to perform assessment tasks with two different interlocutors (a NNS of the same proficiency and a NNS at a different proficiency level). Subjects who were paired had known each other through the Japanese course they were studying, but were not friends. In each session, subjects did three different tasks (two one-way tasks and one two-way task) with an interlocutor. Versions of tasks and types of interlocutors were counterbalanced. Each session took approximately thirty minutes. At the end of the second session all subjects were asked to fill in a questionnaire eliciting their comments on the assessment tasks and interlocutors.

Interlocutor type	Group A (n=8)	Group B (n=12)
NNS with the same proficiency (NNS-S)	a	b
NNS with different proficiency (NNS-D)	b	a

Table 1. Task administration 1

	Group A (n=8)		Group B (n=12)	
	Group A1 (n=4)	Group A2 (n=4)	Group B1 (n=6)	Group B2 (n=6)
First task	with NNS-S	with NNS-D	with NNS-S	with NNS-D
	(Task a)	(Task b)	(Task b)	(Task a)
Second task	with NNS-D	with NNS-S	with NNS-D	with NNS-S
	(Task b)	(Task a)	(Task a)	(Task b)

Table 2. Task administration 2

2.3 Assessment tasks

The tasks used for assessment purposes in the present study (one-way and two-way tasks) are widely used in many foreign language classrooms. The distinction between one-way and two-way tasks is made according to the direction in which the information flows (Pica et al. 1993). In a two-way task, each participant has information which his partner does not have. In completing the task, both participants are required to convey the information they possess to their partner. Information flows in two directions from both participants in two-way tasks. In contrast, in one-way tasks, one participant holds all the information which is necessary to complete the task. Information flows in one direction from the participant who holds the information to the partner who has no information at all.

The first task used in the study was a two-way picture sequence task. Each participant was given three out of the six pictures from a cartoon story. Each participant took turns describing the essential features of each picture given without showing it to his/her partner. Together they discussed a possible sequence of pictures to build up a story. Until both participants had decided on the sequence of pictures, they were not allowed to see each other's pictures.

The second and third tasks used in the study are one-way map tasks. The information given to the subject in the role of information receiver was a street map. The subject in the role of information provider was asked to explain how to get to a certain place. The interlocutor was asked to draw a simple map while listening to the explanation. After finishing the task, their roles were swapped, and the task repeated with a different street map.

2.4 Data and analysis

2.4.1 Data

Data in the present study were collected from three different sources: assessment scores, discourse analyses and questionnaire responses.

2.4.1.1 Assessment scores

Subjects' performances were assessed on a four-point scale by two experienced raters who are also experienced teachers of Japanese, using the rating scale for a proficiency test developed at the Language Testing Research Centre, University of Melbourne. The six assessment criteria are Grammar and expression, Fluency, Pronunciation, Vocabulary, Communication strategies and Task fulfilment. Task fulfilment was assessed twice; once on Task One and once on Task Two or Three. Interrater reliability was calculated using Spearman's rho. The interrater reliability was .931 ($p<.01$)

2.4.1.2 Discourse features

All subjects' performance on assessment tasks were transcribed and coded. The interactional features considered in the study were selected according to the extent to which they help subjects' task performance in terms of comprehension and speech production. The categories were adopted from categories used by Ross and Berwick (1992). They are slowdown, display question, lexical simplification, comprehension check, fronting, clarification request, grammatical simplification, and other expansions. In addition to these interactional features, the total number of C-units, turns and ungrammatical utterance were also calculated. C-units are utterances and sentences, grammatical and ungrammatical, which provide communication value (Rulon and McCreary 1986).

2.4.1.3 Questionnaire response

Subjects were asked to respond to a number of statements about such issues as interlocutor behaviour, task content and difficulty by choosing from options on a 5 point Likert scale. Space was also provided for open-ended comments about each of the assessment components.

The present study is a preliminary study which aims to explore the potential impact of an interlocutor's proficiency on a test-taker's assessment score and discourse. That being so, the number of subjects is small, and no hypotheses are addressed. All the data mentioned above were analysed using descriptive statistics only.

3. Results and discussion

3.1 Assessment score

Subject	Inter-locutor	G&Ex	Flu	Pro	Voc	Com	TF1 (task 1)	TF2 (task 2/3)	Sum
High	High	3.75	3.84	3.87	3.65	3.56	3.75	3.70	26.2
	Low	3.46	3.65	3.53	3.21	3.0	3.06	2.92	22.94
Low	High	1.50	1.30	1.72	1.61	1.72	1.50	1.74	9.306
	Low	1.27	1.19	1.50	1.36	1.489	1.278	1.166	6.139

Table 3. Mean scores for each rating category and for overall score

Table 1 shows the mean scores for each rating category and for the overall assessment score. On the whole, subjects of high proficiency did better when they were paired with a subject of the same proficiency. On the other hand, subjects of low proficiency did better with a subject of different proficiency.

3.2 The amount of talk

		Task 1		Task 2/3 (Inf sender)		Task 2/3 (Inf receiver)	
Subject	Inter-locutor	C-unit	Turn	C-unit	Turn	C-unit	Turn
High	High	33.12	23.87	28.75	14.75	17.5	16.25
	Low	27.75	14.76	14.75	9.37	9.25	7.62
Low	High	43.2	19.56	18.89	14.89	10.22	9.67
	Low	23.44	17.2	14.22	9.89	7.67	4.11

Table 4. The amount of talk (means)

Table 2 shows the amount of talk by subjects on each task. As explained earlier, interactional features such as clarification requests, feedback, slow-down, and ungrammatical utterances, were coded, but the frequency of each feature was very small (only one to

five occurrences in each dyad). It was assumed, therefore, that the very small number of occurrences of these features would be unlikely to affect the subjects' assessment score, and consequently only the amount of talk in terms of c-units and turns was compared.

The trend evident in the assessment score, was alo observed in the amount of talk. High proficiency subjects talked more when they performed tasks with same proficiency subjects. Low proficiency subjects produced more when they were paired with different proficiency subjects.

However, the large standard deviation shows that there are quite a few individual differences which occur regardless of subjects' proficiency. Table 3 shows some examples of individual variation. H5 talked more when she was paired with same proficiency subject, and this trend was consistent across all tasks. In contrast, H2 talked less when she was paired with the same proficiency subject and talked more when was she paired with the lower proficiency subject. Similarly L1 talked more with the higher proficiency subject and less with the same proficiency subject. However, for the last subject listed, L2, the opposite pattern of behaviour was observed.

		Task 1		Task 2/3 (inf sender)		Task 2/3 (inf receiver)	
Subect ID	Inter-locutor	C-unit	Tum	C-unit	Tum	C-unit	Tum
H5	High	28	17	20	5	12	12
	Low	11	7	21	21	2	2
H2	High	17	10	7	5	6	6
	Low	30	41	52	39	8	15
L1	High	49	47	15	16	20	21
	Low	28	25	12	22	18	14
L2	High	11	6	13	9	2	2
	Low	27	22	13	16	12	3

Table 5. Individual differences in the amount of talk

How, then, are these individual differences in the amount of talk reflected in the assessment score? Table 4 shows the assessment scores of the four subjects mentioned above.

In the case of H5 and L2, scores were better when they talked a lot than when they talked less. On the other hand, for H2 and L1, the amount of talk did not have much impact on their scores. Certainly, it would be easier to assess subjects' performance if a larger language sample were available.

These individual differences in the amount of talk may be also explained by other factors such as anxiety rate, confidence level, and the perception of task difficulty, as well as the proficiency of the interlocutor.

Subject ID	Inter-locutor	G&Ex	Flu	Pro	Voc	Com	TF1 (task 1)	TF2 (task 2/3)	Sum
H5	High	4	4	4	4	3.5	4	4	27.75
	Low	3	3.75	3	2.75	2	2	3	19.5
H2	High	4	3.5	3.75	3.5	4	3.5	3.5	26
	Low	3.75	4	4	3.75	3.5	3.5	3.5	26
L1	High	1	1	1	1	1.5	1	1.5	8
	Low	1	1	1.5	1	1.5	1	1	7.5
L2	High	1	1	1.75	1.75	1.5	1	1	9.5
	Low	1.5	1.5	2	2	1.75	2	1	12.0

Table 6. Individual differences in the assessment scores

For example, H2, who talked more with the different proficiency interlocutor was willing to rephrase and accommodate to help her less proficient partner. L1, who also talked more with the High proficiency interlocutor, was able to make use of the better quality language input produced by her more proficient partner and was willing to modify her own speech whenever interactional modification was requested through clarification requests and confirmation checks. Both H2 and L1 seemed less anxious about the differences revealed in their interlocutors' talk and were willing to offer help to or receive help from the interlocutor in order to complete the tasks.

On the other hand, H5, who talked less with the different proficiency interlocutor, may have felt awkward or not confident enough to lead the conversation by rephrasing the interlocutor's speech and accommodating her own speech to that of her interlocutor, but may have felt more comfortable speaking with the same proficiency interlocutor. L2, who talked less with the higher proficiency interlocutor seemed to lack the confidence to ask her for help, thus allowing the interlocutor to dominate the conversation. It may be that different personalities affect anxiety rate and

confidence level differently, resulting in considerable variability in performance. Studies by Berry (eg 1993, 1995) have offered evidence of such a relationship between personality trait and performance.

The individual differences found in the present study are also reflected in the variation in questionnaire responses. In general, subjects favoured the paired interview format for reasons similar to those found in past studies. It was considered, for example, that performing tasks with non-native rather than native speaking interlocutors created a non-threatening environment and made test-takers feel more relaxed. Some subjects preferred pairing with the same proficiency subject, while others preferred a different proficiency subject. The higher the proficiency of the subject, the less concern was expressed about the proficiency of the interlocutor.

Questionnaire responses also revealed that the perception of task difficulty has some impact on performance. As mentioned earlier, each task has two versions, and these two versions were designed to be of equivalent difficulty, but some subjects found one version of task more difficult than the other, and vice versa. Some subjects talked more in the task which they found difficult, and others talked less. There is clearly considerable individual variation in terms of the relationship between the amount of talk and the subjects' perception of task difficulty.

4. Conclusion

The present study examined whether the proficiency of a non-native speaking interlocutor has any impact on the assessment score assigned and on the nature of the discourse produced during task performance. The results of the study reveal that subjects gained slightly higher scores and talked more when their interlocutor was of high rather than of low proficiency. However, not all subjects talked more when their interlocutors were high proficiency learners. The production of a larger language sample moreover did not necessarily lead subjects to gain higher scores.

The proficiency of interlocutors is often considered the most important factor in determining performance on the paired interview format in oral performance assessment, but the findings of this small scale study show that subjects' anxiety rate and confidence level in relation to the proficiency of interlocutors affect

the assessment scores and the amount of talk differently. In addition, the subjects' perception of task difficulty had some impact on their scores and the amount of talk they produced.

In validating the use of a paired interview format for oral performance assessment, how are these individual variations to be considered in relation to issues of fairness? There are many other interlocutor variables other than proficiency, such as gender, ethnicity, age and personality, which potentially raise test-takers' anxiety. In the present study these potential variables are excluded, but further larger scale studies investigating the relative contribution of these variables to performance will allow us to gain more insight into the validity of paired interview format and enable us to evaluate in more depth whether this is a fair mode of assessment.

References

Berkoff, N.A. (1985). Testing oral proficiency: a new approach. In Lee, Y.P. (Ed.) *New Directions in Language Testing.* (pp. 93–100). Oxford: Pergamon Institute of English.

Berry, V. (1994). Personality and the assessment of language performance. Paper presented at the 20th Australian Applied Linguistic Association Congress. University of Melbourne, July.

Berry, V. (1995). A qualitative analysis of factors affecting learner performances in group oral tests. Paper presented at the 17th Annual Language Testing Colloquium, Long Beach March.

Fulcher, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing 13*, (1), 23–49

Gass, S. & Varonis, E. (1985). Task variation and non-native/non-native negotiation of meaning. In Gass, S & Madden, C. (Eds.) *Input in Second Language Acquisition.* (pp.149–161). Rowley, MA: Newbury House.

Gass, S. M. & Varonis, E. M. (1986). Sex differences in NNS/NNS Interactions. In Day, R. (Ed.) *Talking to Learn: Conversation in Second Language Acquisition.* (pp. 327–351). Rowley MA: Newbury House.

Hilsdon, J. (1991). The group oral exam: advantages and limitations. In Alderson, J.C. & North, B. (Eds.). *Language Testing in the 1990s.* (pp.189–197). London: Modern English Publications and the British Council.

Iwashita, N. (1993). Comprehensible output in NNS-NNS interactions in Japanese as a foreign language. Unpublished MA thesis. University of Melbourne.

Lazaraton, A. (1992). The structural organisation of a language interview: a conversation analysis perspective. *System 20,* (3), 373–386.

Morrison, D. M. & Lee, Y. P. (1985). Simulating an academic tutorial: a test validation study. In Lee, Y. P. (Ed.) *New Directions in Language Testing.* (pp. 85–92). Oxford: Pergmon Institute of English.

Pica, T., Kanagy, R. & Falodun, J. (1993). Choosing and Using Communication Tasks for Second Language Instruction and Research. In Crookes, G. & Gass, S. (Eds.). *Tasks & Language Learning.* (pp. 9–34). Multilingual Matters.

Pica, T., L. Holliday, Lewis, N. & Morgenthaler, L. (1989). Comprehensible output as an outcome of linguistic demands on the learner. *Studies in Second Language Acquisition 11,* 63–90.

Porter, P.A. (1986). How learners talk to each other: Input and interaction in task-centered discussions. In Day, R. (Ed.). *Talking to learn: Conversation in second language acquisition.* (pp. 201–222). Rowley, Massachusetts: Newbury House Publishers.

Reves, T. (1991). The group-oral test: an experiment. *English Teachers' Journal 24,* 19–21.

Ross, S. (1992). Accomodative questions in oral proficiency interviews. *Language Testing 9,* (2), 173–186.

Ross, S. & Berwick, R. (1992). The discourse of accomodation in oral proficiency interviews. *Studies in Second Language Acquisition 14,* 2, 159–176.

Rulon, K. A. & McCreary, J. (1986). Negotiation of content: Teacher-fronted and small-group interaction. In R. Day (Ed.). *Talking to learn: Conversation in second language acquisition.* (pp. 182–199). Rowley, Massachussetts: Newbury House.

Shohamy, E., Reves, T. & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal 40,* 212–220.

Appendix

Details of subjects

	Proficiency	Study at Year 12 level	L1	Other L2s	length of in-country experience
H1	H	yes	English	French	12 months
H2	H	no	English	none	12 months
L1	L	no	Cantonese	Mandarin English French	none
H3	H	yes	English Cantonese	French Mandarin	none
L2	L	no	Cantonese	English	none
H4	H	yes	English Cantonese	French	none
H5	H	yes	English	none	none
H6	H	yes	English	French German	12 months
H7	H	yes	English	French	3 weeks
L3	L	no	Mandarin	English	1 month
H8	H	no	English	none	12 months
H9	H	yes	English	none	12 months
L4	L	yes	English	German	none
L5	L	yes	English	none	none
L6	L	yes	English	none	3 weeks
L7	L	yes	English	French	none
L8	L	yes	English	none	none
L9	L	yes	English	French	none