# Investigating the Impact of Opening Gambits on Group Oral Scoring

**Dennis Koyama**

**Eric Setoguchi**

**ABSTRACT**

A major priority of the KEPT research team is to be active in the continual evaluation of the KACE tasks and their value as an institutional proficiency test. One of the major purposes of having tasks like the Group Oral assessment is to serve as a direct supporter to some of our most important curriculum goals here at Kanda, namely communication skills and group conversation management. The Group Oral provides us with information about student abilities in this area, which then enhances our program's ability to make positive outcomes related to pushing these proficiencies, including materials improvement, enhanced student motivation, and placement decisions. Now, these are all great things for a program to have in theory, but there is a danger in assuming that simply having a test allows a program to automatically reap the benefits of that test. As responsible testers we must always be aware that the true usefulness of an assessment to perform any given function is dependent on how strongly an argument can be made that the assessment is measuring what it is intended to, and does so reliably and accurately. This includes an ongoing assessment of specific quality measures of the test as a language measurement tool, including test validity, practicality, and usefulness. The present study investigates the effect on aa test-taker's ratings of making an Opening Gambit, or taking the first turn in a group conversation, and the impact such a finding would have on the validity and usefulness of the test.

Since Long and Porter's (1985) seminal article titled *Group work, interlanguage talk, and second language acquisition*, group work has become commonly used in the communicative language classroom. In their article, Long and Porter identified five arguments in support of group work activities in the language classroom. They explained that group work "increases language practice opportunity, improves the quality of student talk, helps individualize instruction, promotes a positive affective climate, … and … motivates learners" (pp. 208-212). One can easily imagine how the form of group work could vary from class to class, as it is inextricable related to the focus of the class. For example, academic speaking courses may focus on group presentations, writing courses might spotlight peer feedback, and academic reading courses may highlight group discussions on an assigned reading. Indeed, a very long list could be generated in a very short time. One way in which language programs can marshal the benefits of group work is by using placement tests to assist administrative decisions on students' placement into classes.

Since its inception in 1989, the Kanda English Proficiency Test's group oral has been used as a research tool and measure of students' English proficiency increases throughout their university studies at KUIS (Bonk, 2000). Unlike the TOEFL or TOEIC, the tasks featured in the KEPT group oral are closely related to KUIS's communicative curriculum. The group oral was specifically designed to mirror some of the common tasks students perform in the classroom, as a mainstay of the freshman and sophomore English curriculums at KUIS is group discussion activities.

***Assessing speaking: A brief overview***

The enterprise of assessing speaking skills has a long history in language education. For many years, the prevailing method of assessing the speaking ability of language students has been structured interviews. The ubiquity of structured interviews in language programs was based on, as Lazaraton (1992) points out, the assumption that interview tests measure the ability of the interviewee to converse. As research increased in the area of assessing speaking, interview tests were examined to identify if the interview was a valid form of assessment which elicited fundamental aspects of communicative behavior. In a detailed study of oral proficiency interviews, van Lier (1989) pointed out that the "almost inescapably asymmetrical" (P501) nature of the oral interview was a stark contrast to the dynamic nature of conversations in which interlocutors had equality of status. Van Lier went on to detail how interviews are planned and inherently dominated by the interviewer who directs the interaction by asking questions (cf., Talmy, 2011). Furthering van Lier's claims, Young (1995) notes that the predetermined role of the interviewer introduces a power difference in the dynamics of the interview. This power differential, Young claims, is not conducive to producing conversational language.

Clearly participants of a group discussion have more opportunities to meaningfully negotiate with and engage interlocutors in the nomination of topics and to employ conversation maintenance strategies such as turn taking, which Gumperz (1982) labeled "conversational involvement." If the construct of interaction in a conversation is what a language program is trying to measure, programs should remove the influence of the interviewer from the interaction.

This can be achieved with group oral tests.

***Group orals.***

Test tasks, such as the group oral discussion, have been designed with a primary aim of affording test takers control over the direction and content of the language on which they will be scored. The group oral utilizes a relatively unconstrained format in which small groups of test takers discuss a topic without support or interference from test administrators (Bonk & Ockey, 2003; Liski & Puntanen, 1983; Ockey, 2009; Shohamy, Reves & Bejarno, 1986; Van Moere, 2006), who might otherwise have an influence on the interaction as seen in the one-on-one oral interview (Brown, 2003; Johnson & Tyler, 1998; Kormos, 1999; Lazarton, 1996; Ross & Berwick, 1992; van Lier, 1989; Young, 1995). In the group oral, a small group of test takers, typically three or four, are assigned a general topic and expected to sustain a discussion on it for a given period of time. Test administrators quietly sit outside of the group and assign scores based on each test taker's contribution to the discussion.

In addition to giving control to the test takers, group orals have also been noted for aiding institutions with test administration. Group orals have been identified as a possible solution for assessing large numbers of examinees, and have seen praises as an efficient method of judging the oral ability of large numbers of students in a relatively short amount of time (Bonk & Ockey, 2003; Folland & Robertson, 1976, Hilsdon, 1995). Fulcher (1996; He & Dai, 2006) noted how the group oral also provides a situation in which raters can focus on content with the added benefit of having students feel less stressed than in interview style tests.

Research conducted at KUIS has shown the KEPT group orals to be a reliable tool of assessing students' communicative skills as defined at KUIS. In a KUIS-based investigation, Bonk and Ockey (2003) found that prompts do not seem to differ greatly in difficulty across groups, and that the test is reliably separating student into two to three levels of oral proficiency notwithstanding the short testing sessions (15 minute sessions). They also found that raters do not have a stable severity level across years, but that raters tended to perform more consistently as they rated more sessions. Bonk and Ockey also looked at the rubric and found that the scales appeared to be working properly, and no consistent patterns of bias were discernable in the data.

Following Bonk and Ockey (2003), Bonk and Van Moere (2004) investigated the group orals with a focus on threats to validity. They conducted a standard regression with 25 predictors (e.g., gender, shyness/outgoingness score, previous year's oral score, and 22 more). They found more support for the claim that different prompts do not have large effects on oral scores, and that the differences that do exist were easily controlled for. They also found that raters displayed a range in severity, but this was not surprising. McNamara (1996) holds the position that rater variation that is not erratic may be desirable, for if all raters gave the same score on each occasion, there would be a need for only one rater. Bonk and Van Moere also found that interlocutor proficiency and gender did not create a significant difference in scores. While controlling for proficiency and investigating shyness/outgoingness as a possible factor, the researchers found that shyness/outgoingness only slightly affect group oral scores by benefitting outgoing students. Again this is not a surprise, but what is also implied in this

finding is that the outgoingness of one student does not hurt another student who maybe considered shy.

Bonk and Van Moere (2004; Van Moere, 2006) investigated if and how gender and shyness affect group oral assessment outcomes. Bonk and Van Moere's study found that gender had no influence on group scores, and shyness had only minimal affects. Following the 2004 study, Van Moere further investigated shyness and found there was no difference between shy and non-shy students regarding their ability to perform to the best of their ability. In fact, both groups felt that their contribution to the discussion was not inhibited by the other members of their group. In a rigorous study investigating personality constructs on group scores, Ockey (2009) found there was no group effects for non-assertive examinees, and assertive test takers benefitted when grouped with only non-assertive test takers but were disadvantaged when grouped with only assertive test takers.

A microanalysis of the KEPT group orals found that the number of turns taken by group, the interaction between words spoken, and the incoming proficiency level of the student members had no significant impact on student scores as a main effect (Kobayashi, Van Moere & Johnson, 2005). This means that students cannot improve their scores by speaking more and that students are not castigated for speaking less.

***The Group Oral at KUIS***

The group oral test is a well known staple of institutional assessment at the university. As it has been frequently described and reported on in detail in

previous studies, only a brief description will be given here. Test takers view an instructional video immediately prior to taking the group oral. This video is in Japanese and it explained how to complete the task and how test takers would be assessed. The video shows a group of four test takers discussing a topic while a narrator explained what was expected. After viewing the video, groups of four test takers were invited into a testing room and seated in a small circle. Raters introduced themselves by stating their names and then asked test takers to introduce themselves. A prompt, which was written in both Japanese and English, was then distributed. After a minute, one of the raters said, "Would someone like to begin?" Once the test had begun, test takers were expected to sustain a discussion on the assigned topic for eight minutes. For test security reasons, six prompts were used, and groups were randomly assigned a topic. An example prompt is as follows: "Please discuss the following with your group members: Do you prefer Japanese music or foreign music? Why? Have you ever been to a concert or live music event? How did you like it?" Two raters sat outside of the group and assigned ratings for pronunciation, fluency, lexis and grammar, and communication strategies (See Appendix A). After eight minutes, one of the raters ended the test by informing the students that they had completed the test.

***Motivation for the current study***

In an effort to understand the test and what is was measuring, the following are just a few of the test factors that have been investigated for their impact on Group Oral score: gender, rater reliability, shyness, talkativeness, prompt type, and class familiarity. This study investigates one further factor, the opening gambit, or the first turn taken in the group conversation. A few things about the opening gambit

make it worthy of attention. First, there are several "how to start a group discussion" type activities in KUIS classroom materials, including Freshman English. In such activities, there is often an emphasis on the importance of playing a leadership role, which includes responsibilities like directing the conversation. It is not an unpopular opinion that Japanese students seem to have difficulty doing this. It is reasonable to assume then that something like "starting the conversation" would be characterized as a good communication skill. It could be hypothesized that if making an opening gambit could therefore be linked to a tendency to score higher on the test.

A pilot study of archived group oral scores found evidence to support a finding that students who made an opening gambit of any type, seemed to score higher on the test overall than students who did not. This was an interesting finding for because it opened up two possibilities. (1) students who tend to make the opening gambit also tend to be the higher level speakers, therefore their language abilities themselves are resulting in a higher overall score, or (2) it is the opening gambit itself that is boosting these students scores. This is not an unreasonable possibility given that raters are often have quite a bit on their plate in trying to give 4 ratings to 4 students in a few short minutes. That there is a tendency for association with beginning a conversation and communication proficiency in teachers minds. When rating, someone might hear an opening gambit, and already begin making assumptions about that student's ability. Confirmation bias, or the tendency our minds have to focus on evidence that supports opinions we currently hold, would amplify this effect over the course of the conversation. At this point it became necessary to know which of these two possibilities was actually

happening. If raters were deciding something about a student's language ability because of the opening gambit this would be harmful to the validity of the test. The test is about a student's ability to maintain strong language performance over the course of the entire discussion.

### I*nvestigating the Opening* Gambit Factor

A study was designed to investigate whether making an opening gambit in the group oral test has an impact on rating. Six experienced raters were recruited to watch a series of previously recorded group oral sessions from previous administrations of the test. It was necessary to use recorded sessions rather than live ratings as it would be necessary to manipulate what raters see in ways that would not be possible in a live rating. Selected group oral sessions from 2007 onwards were previewed, from which 10 were selected for quality and clarity. Using a video editing program, the initial opening gambit was edited out from each video, essentially creating a new opening gambit out of what had been the 2nd turn in the conversation. Care was taken to ensure that the new opening gambit did not seem out of place by containing any reference to information from in the original opening gambit. Removal of the opening gambit alone had minimal impact on the duration of the conversation, typically averaging less than 30 seconds of footage.

#### *Rating*

Initially, each of the six raters went through a self-norming process wherein a sample group oral session was watched followed by review of the recommended benchmark scores, followed by a second session which raters scored independently, later comparing their scores to another set of benchmarks. If there

was a deviation in points of larger than 1.5 in any of the rating categories, the raters were instructed to be aware and to adjust their scoring practices as necessary.

Once normed, the raters were divided into two teams (A & B) of three. Each was provided with a set of 10 sessions to rate, five of which were untouched and five of which had been edited to remove the original opening gambit. A crossed study was employed, where each team rated opposing sets of sessions. For example, for each untouched session rated by team A, team B watched the edited version, and vice versa. Both teams watched the 10 sessions in the same order. The raters were only allowed to watch each session once, and rated all students present in all four scoring categories, to closely match authentic testing conditions as best as possible.

***Results***

The only scores of interest in this study are those of the student who made the original opening gambit in each of the 10 sessions. If it is the language of the student over the course of the conversation that is largely determining their score, there should be no significant difference in score between the two teams. If however the opening gambit itself is having an impact, there will be a measurable difference in score, likely with the team that watched the untouched session assigning a higher score. The scores for the 10 students are shown below in Table 1.

**Table 1. Test-taker scores with and without opening gambit**

| | With opening gambit | | | | Without opening gambit | | | |
|---|---|---|---|---|---|---|---|---|
| Student # | **P** | **F** | **LG** | **CS** | **P** | **F** | **LG** | **CS** |
| 1 | 3.3 | 3.5 | 3.0 | 4.0 | 3.5 | 3.7 | 3.3 | 4.0 |
| 2 | 3.2 | 3.2 | 3.2 | 3.2 | 3.3 | 3.5 | 3.2 | 3.3 |
| 3 | 3.5 | 3.3 | 3.2 | 3.7 | 3.2 | 3.5 | 3.0 | 3.7 |
| 4 | 2.8 | 3.0 | 3.0 | 2.8 | 2.7 | 2.3 | 3.0 | 2.7 |
| 5 | 3.7 | 3.7 | 3.7 | 3.8 | 2.8 | 3.5 | 3.5 | 3.3 |
| 6 | 2.2 | 2.2 | 2.2 | 2.3 | 2.3 | 2.2 | 2.0 | 2.5 |
| 7 | 2.8 | 3.0 | 3.2 | 3.2 | 2.8 | 3.2 | 2.8 | 3.7 |
| 8 | 3.0 | 3.0 | 3.0 | 3.5 | 3.2 | 3.2 | 3.0 | 3.3 |
| 9 | 3.2 | 3.5 | 3.5 | 3.5 | 2.8 | 2.8 | 2.8 | 3.3 |
| 10 | 2.8 | 2.8 | 2.8 | 3.5 | 3.3 | 3.2 | 3.5 | 4.0 |
| Average | 3.1 | 3.1 | 3.1 | 3.4 | 3.0 | 3.1 | 3.0 | 3.4 |

The scores appear to be nearly identical, and an independent pairs t-test ($p > .05$) found no significant difference between scores, supporting the conclusion that the opening gambit alone did not have a noticeable effect on how the raters perceived their communicative abilities in pronunciation, fluency, lexis and grammar, and communication skills.

### *Implications*

The findings of this study provide evidence supporting the validity of the group oral test as a measure of group communication skills over the course of a prolonged conversation. It is not the intent of the test to award a more favorable rating of proficiency solely based on opening a conversation, and the scores observed here strongly indicate that this is indeed not what is occurring.

It should be noted however that this was an isolated study involving a relatively small sample of raters compared to the full population. Six raters were employed here, in contrast with a full administration of the test where typically 30 or more raters may participate in scoring of the group oral. Furthermore, all raters participating in this study were known as experienced raters with demonstrated reliability at scoring the test. Therefore, it is not known if the scoring behavior observed here can be applied to the full rater population, particularly with raters who are new to the test. As such, it will be necessary to ensure in the future that rater training is putting raters focus on the language production that occurs over the course of a group oral conversation without bias to any particular point in time.

**REFERENCES**


Bonk, W. J. and Van Moere, A. (2004). L2 group oral testing: the influence of shyness/outgoingness, match of interlocutors' proficiency level, and gender on individual scores. Paper presented at the Language Testing Research Colloquium.

Folland, D. and 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.

Gumperz, J. J. (1982). Discourse strategies. Cambridge: Cambridge University Press.

He, L. and Dai, Y. (2006). A corpus-based investigation into the validity of the CETSET group discussion. Language Testing *23*, 370-401.

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

Lazaraton. A. (1992). The structural organization of a language interview: A conversation analytic perspective. SYSTEM, 20, 373-386.

Kobayashi, M., Van Moere, A. and Johnson, K. (2005). Is silence 'golden'?: An investigation of group oral tests. Paper presented at 14th World Congress of Applied Linguistics, Madison, Wisconsin.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. Language Testing, *20*(1), 89-110.

Brown, A. (2003) Interviewer variation and the co-construction of speaking proficiency. Language Testing, *20*(1), 1-25.

Johnson, M., & Tyler, A. (1998). Re-analyzing the OPI: How much does it look like natural conversation? In R. Young & W. He (Eds.). Talking and Testing: Discourse approaches to the assessment of oral proficiency (pp. 27-51). Philadelphia, PA: John Benjamins Publishing Company.

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of roleplay and non-scripted interviews in language exams. Language Testing *16*(2), 163-188.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: The case of CASE. Language Testing, *13*(2), 151-172.

Liski, E., & Puntanen, S. (1983). A study of the statistical foundations of group conversation tests in spoken English. Language Learning, *33*(2), 225-246.

Ockey, G. J. (2009). The effects of group members' personalities on a test taker's L2 group oral discussion test score. Language Testing 26(2), 161-186.

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

Shohamy, E., Reves, T., & Bejarno, Y. (1986). Introducing a new comprehensive test of oral proficiency. ELT Journal, *40*(3), 212-220.

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils:

Oral proficiency interviews as conversation. TESOL Quarterly *23*(3), 489-508.

Van Moere, A. (2006). Validity evidence in a group oral test. Language Testing, *23*(4), 411-440.