Article

LANGUAGE TESTING

Language Testing
1–26

DOI: 10.1177/0265532211398110
ltj.sagepub.com
SAGE

Effects of test-taker characteristics and the number of participants in group oral tests

Fumiyo Nakatsuhara
University of Bedfordshire, UK

Abstract

This study explores the nature of co-constructed interaction in group oral tests by examining whether a test-taker's own and his or her group members' extraversion levels and oral proficiency levels have different influences on conversational styles between two group sizes: groups of three and groups of four. Data were collected from 269 Japanese upper-secondary school students, who took group oral tests either in groups of three or four. All sessions were video-taped and transcribed following Conversation Analysis (CA) conventions. The data were quantitatively analysed in terms of goal-orientation, interactional contingency and quantitative dominance. Then, CA methodology was used to interpret and elaborate the statistical results. The findings have implications for our understanding of the group oral test construct and for appropriate choices of group size in group oral testing.

Keywords

co-construction, conversational styles, extraversion, group oral tests, group size

Introduction

Over the past few decades, a group oral format, where a candidate interacts with two or more candidates while being assessed, has been widely adopted in both high- and low-stakes tests (e.g. College English Test–Spoken English Test in China, Kanda English Proficiency Test in Japan; see also Table 1). A number of studies have investigated group oral formats from a range of perspectives. Group tests, like paired tests (e.g. Brooks, 2009; ffrench, 2003), are capable of eliciting richer language functions from test-takers than interview tests, providing a great potential to assess communication ability (Nakatsuhara, 2009; Van Moere, 2007). Group formats are also practical in resource-limited pedagogical settings (Ockey, 2001), likely to be positively viewed by test-takers, and may provide a

Corresponding author:
Fumiyo Nakatsuhara, Room 125, CRELLA, Putteridge Bury Campus, University of Bedfordshire, Luton LU2 8LE, UK.
Email: Fumiyo.Nakatsuhara@beds.ac.uk

beneficial washback effect for classrooms (e.g. Folland & Robertson, 1976; Fulcher, 1996; Hilsdon, 1991; Shohamy, Reves, & Bejarano, 1986; Van Moere, 2006). On the other hand, it seems questionable to what extent raters can separate scores of one candidate from those of his or her group members (May, 2007), and it has been suggested that in high-stakes tests, a group oral format should be used only as a part of an oral test battery rather than as a solo assessment tool (Bonk & Ockey, 2003; Van Moere, 2006, 2007).

As evidence for the possible susceptibility of test-taker performance to interlocutor behaviour has been uncovered (e.g. Lazaraton, 1996) and awareness of the co-constructed nature of speaking test performance has grown (McNamara, 1996), research has also investigated the influence of test-takers' own and their paired/group members' characteristics on test-takers' performance in paired/group testing. These characteristics include gender, personality, language proficiency and acquaintanceship (e.g. Berry, 2004; Norton, 2005; O'Sullivan, 2000a, b; Ockey, 2006).

Amongst these variables, extraversion- and proficiency-level variables seem to have attracted the most attention (e.g. Foot, 1999; Underhill, 1987). For instance, more than twenty years ago, Underhill (1987, p. 46) expressed his concern about the danger that paired and group tests without an interviewer's intervention might unfairly reward outgoing personalities. It has also been suggested that level-specific tests such as the Cambridge Main Suite examinations could be more suitable for the paired format than those covering a broad range of proficiency levels like IELTS, as the former could avoid wide disparities in proficiency levels between paired candidates (Lazaraton, 2002, p. 119).

Extraversion-level and proficiency-level factors in paired/group tests

Several studies have demonstrated that a candidate's own extraversion level and that of his or her interlocutor(s) in paired or group settings may affect the test scores awarded. However, the results have been mixed in terms of the direction of the impact (positive or negative). For example, in Berry's (1993, 1997) paired test studies, extraverts were likely to obtain higher scores when paired with another extravert, while introverts' scores did not tend to be influenced by their partners. In contrast, in Berry's (2004) and Ockey's (2006) group test studies, extraverts were likely to get higher scores when participating in more introverted groups. Introverts benefited in more extraverted groups in Berry's (2004) study, but were not affected by group members in Ockey's (2006) study. It was also found that a candidate's own extraversion level was likely to benefit their own performance (Bonk & Van Moere, 2004; Ockey, 2006). A test-taker's own and their interlocutors' characteristics (including extraversion level) also seem to affect the performance in a far more complicated way, as a result of interactions between factors (e.g. candidate gender * partner personality * acquaintanceship) (O'Sullivan, 2000b).

The impact of test-takers' proficiency level on paired/group test scores and discourse has also been extensively explored. Csépes (2002) compared scores when test-takers were paired with higher-, same- and lower-proficiency partners, and no significant difference was found in the scores awarded. The results were echoed by Davies' (2009) study of paired tests and also by Bonk and Van Moere's (2004) study of group tests, where test-takers' scores were neither influenced by the mean scores nor by the variation in proficiency level of their fellow group members. Using Conversation Analysis (CA) methods, Nakatsuhara (2004) found a similar amount of asymmetric interactional

features (e.g. sequential and quantitative dominance) in paired test interactions between matched or unmatched proficiency pairs. On the other hand, other qualitative studies reported that test-takers paired with a higher proficiency partner could benefit from better quality language input (Iwashita, 1998; Norton, 2005). O'Sullivan's multi-variable study (2000b) demonstrated a complex three-way interaction effect involving test-takers' proficiency level, their own gender and their partner's gender.

Task implementation conditions in group oral tests

The studies reviewed above show that the influence of test-taker characteristics in paired and group testing is therefore far from simple, showing mixed results across different studies, and complex interaction effects between different variables. One of the reasons for the mixed results could be that these studies were undertaken with different tasks and under different conditions. Skehan (1998) distinguishes between *task qualities* (e.g. task types, task characteristics) and *task implementation conditions* (e.g. time limit, the number of interlocutors). While recent research has attended to such finer task distinctions in a variety of speaking formats (e.g. Berry, 1997; Iwashita, McNamara, & Elder, 2001; Van Moere, 2007; Wigglesworth, 2001), task implementation conditions in group oral testing have as yet been under-researched.

Despite findings of CA and group dynamics studies that showed the number of interactants is a significant factor in deciding how people talk in turn (e.g. Egbert, 1997; Goodwin & Goodwin, 1990; Lerner, 1993; Sacks, Schegloff, & Jefferson, 1974; Schegloff, 1995; Webb & Palincsar, 1996), it seems that for practical reasons educational and assessment research and practice have usually selected group size. In group oral tests, it ranges from three to seven, and the most common number of participants has been three or four, as shown in Table 1.

Among the studies listed in Table 1, only Liski and Puntanen (1983) and Kobayashi and Van Moere (2004) included the group size factor as part of their investigations. Liski and Puntanen (1983, p. 241) found that the test-takers in bigger groups spoke significantly less than those in smaller groups, although the time allotted to each group took the group size into account. Kobayashi and Van Moere (2004) showed that a group size variable (three or four students) did not have a significant influence on overall test scores.

Group size therefore needs to be addressed, when examining the impact of test-taker characteristics on group oral test performance. It is also necessary to gain insights into how and why the impact of test-taker variables occurs in group tests. Following O'Sullivan's (2004, p. 140) suggestion, we also need to investigate the discourse features related to test-taker characteristics, where there is the potential for an interlocutor effect on performance, 'to discover *why* this is happening and *how* exactly it is manifested in the language of the interaction'. For such investigations, CA seems to be the most appropriate methodology to be employed, for reasons which are presented in the following section.

Use of conversation analysis in speaking testing research

CA is based fundamentally on a model of communication as joint activity (Sacks, 1984), aiming to understand 'how the jointly organized activity of talk-in-interaction is carried out' (Stubbe et al., 2003, p. 354).

Table 1. Tasks and the number of participants in group oral tests

Author (year)	[Name of the test, if operationalized] Tasks	Participant numbers
Folland & Robertson (1976)	- listening to a tape (on a report) - free discussion on the main point of the report that they have listened to	Up to 7
Liski & Puntanen (1983)	- listening to a tape (on a report) - free discussion on the main point of the report that they have listened to	6 (minimum 5 and maximum 7)
Shohamy, Reves, & Bejarano (1986)	- free discussion on a given topic	4
Hilsdon (1991)	[Secondary School Examination, Zambia] - free discussion on a given topic	5
Pavlou (1995)	- free discussion on a given topic	3
Fulcher (1996)	- free discussion on a given topic	(not mentioned)
Nunn (2000)	(not mentioned)	3
Ockey (2001)	- free discussion on a given topic	3
Masubuchi (2003)	[Interactive English Forum] - free discussion on a given topic	3
Nakamura (2003)	- free discussion on a given topic	3 or 4
Bonk & Ockey (2003); Kobayashi & Van Moere (2004); Van Moere (2006, 2007); Ockey (2006)	[The Kanda English Proficiency Test] - free discussion on a given topic (only for Van Moere (2007), free discussion, consensus-reaching and picture difference tasks)	3 or 4
Berry (2004)	[Hong Kong University in-house test] - read a short text - replicate the setting of a small academic seminar on the read text	5 (with occasional exception of groups of 4 or 6)
Hong Kong Examination and Assessment Authority (2005)	[Hong Kong Advanced/ Supplementary Level Examination] - individual presentation based on a text - group discussion in a university-like setting such as a small academic seminar	4 (a minimum of 3)
Gan, Davison, & Hamp-Lyons (2008); Gan (2010)	[School-based group oral assessment] - free discussion based on a film that the candidates watched	3 or 4
He & Dai (2006)	[College English Test–Spoken English Test] - free discussion on a given topic	3 or 4

Over the last two decades, a number of studies have analysed the discourse of various speaking test formats using CA, and CA has been recognized as an invaluable methodology to describe and validate oral proficiency tests (e.g. Brown, 2003; Lazaraton, 1992, 1996, 2002; Ross & Berwick, 1992; Young & He, 1998). CA has informed our understanding of paired and group oral testing discourse. In her research on the interactional features observed

in 30 dyads performing the paired discussion part of the FCE, Galaczi (2004) identified four distinct patterns of interactions, namely *collaborative, parallel, asymmetric* and *blended* patterns, based on Storch's (2002) model of dyadic interaction in an ESL classroom context. She also showed that high scores on the 'interactive communication' scale generally oriented to a *collaborative* pattern of interaction, whilst the low scores to a *parallel* pattern. More recently, Lazaraton and Davies (2008), Galaczi (2010) and Gan (2010) have carried out CA on already scored paired and group test transcripts and identified discourse features salient at different oral proficiency levels. In order to seek to build on such contributions of CA in language assessment, this study will employ CA to obtain a more precise picture of group oral test discourse than has hitherto been available.

Methods and data analyses

Research questions

The present study addresses the following two research questions to compare the impact of test-takers' extraversion levels and oral proficiency levels on conversational styles between two group sizes. Because groups of three or four predominate in group oral tests, this study will focus on these two group sizes.

- Do test-takers' extraversion levels and oral proficiency levels have different influences on conversational styles in groups-of-three participants as against groups-of-four?
- If so, how and why are they different?

Test-takers

The data collection was undertaken in Japan, involving 269 students from five upper-secondary schools. All participants were native speakers of Japanese, aged between 16 and 18 years old, and approximately four-fifths of the participants were females. Therefore, it should be noted that male students are underrepresented.

Participants were asked to take a test in groups of three or four with their classmates as they wished. This was to minimize the possible impact of an acquaintanceship variable on their performance (Norton, 2005; O'Sullivan, 2002). They formed only single-sex groups, and hence, in the case of this research population, by means of controlling the acquaintanceship variable, the gender variable was automatically controlled for. However, this leaves the possibility that a gender effect might have been discovered, if the variable had been tested with the given participants. Nevertheless, in the present research, the participants' first language, educational background and age were comparable, and acquaintanceship and gender variables were controlled.

Measuring extraversion and oral proficiency levels

Following its successful use in the pilot study, a personality questionnaire consisting only of the extraversion scale and the neuroticism scale of the Japanese Eysenck

Personality Questionnaire (EPQ, Iwawaki, Eysenck, & Eysenck, 1980) was administered. The EPQ, based on Eysenck's (1981) theory that personality has a biological foundation, has been recognized as the 'gold standard' for the measurement of extraversion (Costa & McCrae, 1986, p. 66), and the robustness of its construct validity for use with high school students has also been confirmed (Drummond, 1990, p. 236). In fact, a great amount of educational and testing research has been conducted with the EPQ, including Berry's (1993, 1997, 2004) series of studies on the influence of extraversion on scores in paired and group testing. Cross-cultural validation studies have demonstrated the validity of utilizing the translated Japanese EPQ, as a comparable measure to the original English version (Iwawaki, Eysenck, & Eysenck, 1980).

The 20 items of the E-scale were used to measure the level of extraversion, which is the focus of this study (example items include 'Do you like telling jokes and funny stories to your friends?', 'Do you prefer reading to meeting people?'; to be answered dichotomously (yes/no)). The 22 items of the N-scale were included as distractors, since providing only the E-scale questions might make the research intention too obvious to the participants and this could affect the response. Thus, the 22 items were not scored. Out of the 20-point E-scale, the mean score was 11.60 (groups of 3: 11.58, groups of 4: 11.62) and the standard deviation was 4.61 (groups of 3: 4.44, groups of 4: 4.82). The mean and standard deviation of the E-scores obtained in the present study were similar to those of the previous studies employing the Japanese EPQ (Iwawaki, Eysenck, & Eysenck (1980), mean = 11.40, SD = 4.47; Berry (1993), mean = 11.97, SD = 3.97), and the reliability coefficient for the 20 items was .831 (Cronbach's Alpha). This suggests that the questionnaire appropriately measured the degree of participants' extraversion, and that the participants in this study respond within the range expected of Japanese adults.

To measure participants' oral proficiency level, an operational rating scale was developed, based on the experience of the pilot study, which had indicated the lack of appropriate speaking rating scales for Japanese upper-secondary school students. The development included revision of existing speaking rating scales from outside and inside Japan, examination of the guidelines for secondary school education in Japan, drafting a rating scale based on the existing rating scales, while obtaining expert judgements from eight experienced upper-secondary Japanese EFL school teachers, and pilot testing of the scale using multi-faceted Rasch analysis (see Nakatsuhara, 2009 for details of the development and a copy of the scale). The resulting rating scale consisted of five analytical categories (i.e. *pronunciation and intonation*, *grammar*, *vocabulary, fluency* and *interactive communication*) and each category had six levels from 0 (no English level) to 5 (approximately B1 level of the CEFR).

This rating scale was utilized to measure students' oral proficiency level. However, despite the original plan, it was not possible to conduct additional formal speaking tests to assess students' speaking proficiency, owing to practical constraints. Therefore, prior to the group oral test administration, English classroom teachers received one hour of standardization training, using the rating scale. They were then asked to give one holistic score, taking each of the five analytical categories equally into account, on the basis of their knowledge of students' recent class/test performances. Out of the 6-point scale from 0 to 5, the mean score was 2.81 (groups of 3: 2.72, groups of 4: 2.91), and the standard deviation was 1.16 (groups of 3: 1.13, groups of 4: 1.19).

Group oral tasks

Three tasks with different degrees of instructional control and different required outcomes were used in this study. The three task types were as follows: *the information gap* (exchange information and then choose one item based on the exchanged information), *the ranking* (select three items out of the shared 10 items) and *the free discussion tasks* (discuss a given topic to reach a consensus). However, since the focus of this paper is only on group size difference, only interactional features common to the three tasks will be discussed in this paper.

Data collection procedures

Data collection was conducted by trained classroom teachers and the researcher. Prior to the group tests, English classroom teachers had assessed their own students' oral proficiency level, and students had filled in a personality questionnaire. The group oral tests were conducted during English class periods. Owing to the limited class time available, most students performed two out of the three tasks in a counter-balanced order. The minimum speaking test time per task was set as 4–5 minutes for groups of three and 6–7 minutes for groups of four.

Methods of data analysis

All the test sessions were video-recorded, transcribed following CA conventions (Atkinson & Heritage, 1984) and analysed both quantitatively and qualitatively.

Quantitative analysis of conversational styles. Conversational styles in group oral tests were quantified for a test-taker's *goal orientation, interactional contingency and quantitative dominance.* The measures were developed by modifying a set of measures utilized to quantify conversational styles in oral proficiency interviews (Kormos, 1999; Young, 1995; Young and Milanovic, 1992).

The measure of *goal orientation* was the proportion of topics initiated by a test-taker relative to the other test-takers in the same group. The measure of *interactional contingency* was the proportion of topics initiated by his or her group members that were ratified and became the topics of the subsequent turns by a test-taker. The measure of *quantitative dominance* was the proportion of words spoken by a test-taker relative to his or her group members.

Since topic was the key concept used to examine the degree of test-takers' topic initiation and topic ratification, a framework to identify 'goal-oriented topic shift' was first established. Based on a review of conversation analysis and language testing literatures and an examination of actual interactions in small-scale pilot data (27 group test sessions), the following nine criteria were devised to identify topic boundaries. The first six are the main criteria and can be seen as stand-alone, while the latter three criteria are used as sub-criteria which should co-occur together with the main criteria.

Main Criterion 1: Unspecific sequence openers, such as 'What do you think?' and 'Any suggestions?'; this is especially used to elicit opinions from quiet group members by addressing their names, for example 'How about you, Mika?'.
Main Criterion 2: Specific sequence openers, such as 'Shall we look at the opinions of others?' and 'Intelligence, how about intelligence?'.
Main Criterion 3: Using a new lexical item with phonological stress (also sometimes in <slow speed> or in a LOUD VOICE and sometimes with a deep in-breath (.hhh)].
Main Criterion 4: Explicitly announcing the introduction of new information; asking for permission to talk about something new (i.e. presequence; Schegloff, 1980) and pointing out missing information that should be talked about.
Main Criterion 5: Rounding off the previous topic by summarizing what has been discussed; this often occurs with short assessments and/or commentaries, sometimes including repetition of salient lexical items that were used in the topic.
Main Criterion 6: Closing a current topic explicitly by providing reasons for abandoning and terminating a topic (often signalling irrelevance of the current topic and abandoning topics); repair is undertaken and topic-refocus is conducted.
Sub-Criterion 7: Using gestures to direct others' attention to something new.
Sub-Criterion 8: Pre-shift tokens, such as 'alright', 'right' (Gardner, 2007), 'so' (Raymond, 2004), 'okay' (Beach, 1993; Button, 1987), 'yeah', 'then', 'but', 'anyway'.
Sub-Criterion 9: Long unfilled pauses, usually exceeding one second (Jefferson, 1989).

Inter-coder reliability was tested in two steps for 1) the topic boundary criteria and 2) the locations of topic boundaries, using 11% of the transcripts from the pilot test data. A group of 14 researchers was involved in the first phase, where they individually chose matching criteria for discrete topic boundaries and then the outcomes were discussed in plenary. Despite minor differences at the beginning of this session, this group agreed that the set of criteria was an effective instrument. Then, for the second phase, another researcher with expertise in the analysis of spoken interaction individually identified topic boundary locations and the outcome was compared with the author's coding result. An overall agreement level of 91% was recorded. After discussing all disagreements and reaching a consensus on which coding should be adopted, the author coded all the transcripts of the main study (for more details of the measures, see O'Sullivan and Nakatsuhara, in press).

Having therefore quantified conversational styles by means of the above three measures, three sets of standard multiple regression analyses were conducted. The six independent variables (IVs) and three dependent variables (DVs) are shown in Table 2.

As it was impossible for all the candidates to perform all three tasks during the limited classroom time, this practical constraint necessitated 'a between-conditions design'. Care was taken, however, to ensure that participants always performed their two tasks under the same condition, either a group of three or a group of four. Although the given data set did not meet the assumption of independence of residuals, it satisfied all the other assumptions for multiple regression analysis, allowing us to consider the results of statistical significance presented in the following section as at least suggestive.

Table 2. IVs and DVs for multiple regression analysis

IVs	Measured by
1: A test-taker's own extraversion level	A test-taker's self (own) E-score (0–20)
2: His or her group members' extraversion levels	A test-taker's self-excluded E mean score in his or her group
3: The variability of his or her group members' extraversion levels	A test-taker's self-excluded E standard deviation in his or her group
4: A test-taker's own oral proficiency level	A test-taker's self (own) teacher-assessment score (0–5)
5: His or her group members' oral proficiency levels	A test-taker's self-excluded teacher-assessment mean score in his or her group
6: The variability of his or her group members' oral proficiency levels	A test-taker's self-excluded teacher-assessment standard deviation in his or her group
DVs	**Measured by**
1: Goal-orientation	Topic initiation
2: Interactional contingency	Topic ratification
3: Quantitative dominance	The amount of talk

The distribution of students' extraversion level and oral proficiency level in each condition was pre-analysed, since it was crucial to make sure that these variables were similarly distributed across all the conditions. Two sets of two-way between-groups ANOVA (see Tables 3 and 4) and Levene's tests of equality of error variance (Levene's tests: extraversion $p = .479$, oral proficiency $p = .849$) were conducted, and they confirmed that means and spreads of the extraversion and oral proficiency scores of test-takers both in two group sizes and among three task types were not significantly different.

Qualitative analysis of conversational styles. The same transcript data were qualitatively analysed using CA approaches to illuminate the quantitative results. Following CA conventions, the analysis operated on the principle that repeated listening in order to

Table 3. Descriptive statistics of extraversion level and oral proficiency level in different conditions

Task	Group size	Variables	N	Mean	Std.Dev.
Info-gap	3	Extraversion	96	11.42	4.39
		Oral proficiency	94	2.68	1.04
	4	Extraversion	64	12.31	5.19
		Oral proficiency	63	2.79	1.28
Ranking	3	Extraversion	104	11.83	4.26
		Oral proficiency	105	2.82	1.07
	4	Extraversion	68	11.93	4.69
		Oral proficiency	67	2.81	1.21
Free discussion	3	Extraversion	77	10.64	4.63
		Oral proficiency	77	2.71	1.06
	4	Extraversion	67	11.58	4.66
		Oral proficiency	66	2.89	1.20

Table 4. Two sets of two-way between-groups ANOVA

DV	Source	df	F	Sig
Extraversion	Task	2	1.36	.259
	Group size	1	2.28	.132
	Task*Group size	2	.422	.656
Oral proficiency	Task	2	.202	.817
	Group size	1	.768	.381
	Task*Group size	2	.288	.750

produce a transcript was itself an important part of the discovery analysis process (Hutchby & Wooffitt, 1998; Lazaraton, 2002; Levinson, 1983), helping to uncover recurrent, systematic features of group interactions. A list of the transcription notations is provided in the Appendix.

The data were treated in two ways in the result section that follows. In the first part, the coded data were analysed through multiple regression to provide a generalizable picture of the data. However, CA is generally hostile towards explicit coding and quantification (Schegloff, 1993). Therefore, when applying CA approaches in the second part of the results, the analysis focuses on interpretation, avoiding the use of explicit quantitative language. The two perspectives are seen as complementary, the CA providing insights that help to explain the statistical results.

Results and discussion

Collective influence of two test-taker characteristics

In terms of results, the two group sizes showed a similar amount of variance in topic initiation explained by the regression model (16.5% and 14.2% in groups of three and four respectively, $p < .0005$). In the amount of talk, the model explained the variance in groups of four (24.3%, $p < .001$) a little better than that of groups of three (19.6%, $p < .0005$). Yet, the variance in topic ratification was not accounted for by the model for either group size. Therefore, topic ratification will not be considered further. The following section will describe further multiple regression results for topic initiation and the amount of talk.

Separate influences of two test-taker characteristics

Tables 5 and 6 report the degree to which each variable included in the model contributes to the prediction of the DVs. In general, the more extraverted and more orally proficient test-takers tended to initiate more topics and talk more, especially when they were grouped with less extraverted and less proficient members. However, concerning both topic initiation and the amount of talk, there were two main differences between groups of three and four.

First, extraversion-level variables were more influential in groups of four than in groups of three. For instance, a test-taker's own extraversion level was a statistically

significant variable for topic initiation in groups of four, but not in groups of three. For the amount of talk, while a test-taker's own extraversion level was the only significant variable in groups of three, groups of four showed a significant impact of both a test-taker's own and his or her group members' mean extraversion levels.

Second, whilst there was an influence of the proficiency-level variables in both group sizes, the effect size was larger in groups of three than in groups of four.

Having obtained these two main differences between groups of three and four, it is important to gain insights into possible reasons behind these findings. CA provides the methodology for qualitative perspectives on the data.

Table 5. Multiple regression coefficients [DV: topic initiation]

Group size	Predictors	Standardized coefficients Beta	Sig
Groups of 3	(Constant)		.001
	Ext – self	.107	.077
	Ext – self-excluded group mean	–.082	.195
	Ext – self-excluded group std.dev.	–.050	.424
	Prof – self	**.399**	**.000**
	Prof – self-excluded group mean	**–.344**	**.000**
	Prof – self-excluded group std.dev.	–.001	.988
Groups of 4	(Constant)		.025
	Ext – self	**.225**	**.001**
	Ext – self-excluded group mean	–.141	.067
	Ext – self-excluded group std.dev.	–.056	.472
	Prof – self	**.249**	**.002**
	Prof – self-excluded group mean	**–.187**	**.017**
	Prof – self-excluded group std.dev	.023	.741

Table 6. Multiple regression coefficients [DV: the amount of talk]

Group size	Predictors	Standardized coefficients Beta	Sig
Groups of 3	(Constant)		.000
	Ext – self	**.144**	**.015**
	Ext – self-excluded group mean	–.110	.075
	Ext – self-excluded group std.dev.	–.117	.057
	Prof – self	**.409**	**.000**
	Prof – self-excluded group mean	**–.369**	**.000**
	Prof – self-excluded group std.dev.	–.066	.277
Groups of 4	(Constant)		.000
	Ext – self	**.244**	**.000**
	Ext – self-excluded group mean	**–.183**	**.012**
	Ext – self-excluded group std.dev.	–.087	.231
	Prof – self	**.370**	**.000**
	Prof – self-excluded group mean	**–.277**	**.000**
	Prof – self-excluded group std.dev.	.057	.386

Insights from conversation analysis

The analysis revealed three interactional features unique to the two group sizes that could be related to the above quantitative findings: 1) collaborative atmosphere in groups of three; 2) avoidance behaviour in groups of four; and 3) mechanical turn-taking in groups of four.

Collaborative atmosphere in groups of three. There appeared from the analysis to be a more collaborative atmosphere in groups of three than in groups of four, which might have mitigated the effect of extraversion variables in groups of three. Galaczi (2004, p. 118) describes 'the collaborative interactional pattern' as an interaction exhibiting high mutuality and high equality, where test-takers jointly construct an interaction while sharing and developing each other's idea. Here, I am not claiming that groups of three had more collaborative interaction as an overall pattern of their interaction. Rather, there seemed to be more attempts to make the interaction collaborative in groups of three, and these attempts seemed to be more successful in the smaller group size.

One of the collaborative attempts frequently observed in groups of three was joint utterance completion. The following Excerpt (1) was extracted from the ranking task performed by a group of three (see Table 7). While the proficiency levels of the three test-takers were the same, candidates *A* and *B* were rather introverted with their extraversion level at 6, whereas *C*'s extraversion level is at 12.

Table 7. Test-taker characteristics and performance summary [Task: Ranking]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
A	6	3	33	25	60
B	6	3	33	50	20
C	12	3	33	25	20

Excerpt (1) test-taker: *A, B, C*

1	→	B:	I think en- enthusiasm for teaching is most important point, because (3.0)
2			uh:: (5.0)
3	→	C:	Uh: I agree.
4		A:	Why?
5	→	C:	uh:::: I think enthusiasm is (.) uh:::: (1.0) u(h)h:: huh huh
6	→	B:	*Hai* ((raising a hand)) [Huh
7		A:	[Hah hah hah
8		C:	[Hah huh huh
9		C:	Uh
10		B:	Teacher's enthusiasm makes [us our enthusia(h)sm,
11		C:	[Uh uh
12		B:	so (.5) we study (1.0) very (1.5)
13	→	C:	So ah:[:
14	→	A:	[We can study more work.
15		C:	Uh:::

16 A: I think I agree with enthusiasm for teaching. I think enthu- enthusiasm for
17 teaching is need (.) necessary, because if I (.) I: (.) I don't (.5) if I can't
18 solve the problem, uh: I want to teach (.5) uh I want to: .hh be taught
19 so- taught (.) by him.

This is an example of how candidates attempted jointly to complete an utterance. In line 1, *B* started talking about the importance of *enthusiasm for teaching*. However, she struggled to justify her opinion as shown in her 3- and 5-second intra-utterance silence. At that point, *C* terminated the silence by agreeing with *B* in line 3, which led *A* to ask the reason for agreement. Then, it was *C* who needed to justify her opinion, but the justification was not successful as shown in the partially filled silence in line 5. By that time, *B* came up with the language which she had been searching for in her previous utterance, and helped *C* out by starting to give a reason as if it had been a joint idea of *B* and *C*. Nevertheless, in the middle of her turn, *B* again faced difficulty in completing the utterance in line 12. That was when *C* and *A* jointly provided language to complete *B*'s utterance in lines 13 and 14, followed by *A*'s additional agreement in lines 16–19. In this way, this group helped each other in jointly completing utterances, when others had difficulty in expressing ideas. While joint utterance completion was observed both in groups of three and four, it was noteworthy that such collaborative attempts including all the group members seemed to be unique to groups of three. It appears in other words that groups of three tended to establish a greater degree of solidarity to accomplish the task all together than groups of four. This interesting finding echoes Coulson's (2005) observations about the number of group members in relation to the use of collaborative communication strategies. While he did not present the actual data to compare interactions between different group sizes, Coulson (2005, pp. 128–129) refers to his finding that the recorded students' group performance identified that collaborative peer strategies like 'team-centred repair of an incomplete utterance' could more readily appear in groups of three.

Another source for the collaborative atmosphere in groups of three was related to more successful cases at interactional scaffolding in groups of three. While both groups of three and four showed 'asymmetric expert/novice interactions' (Storch, 2002; van Lier, 1989), where some test-takers gave scaffolding behaviour to involve quiet group members into the interaction, such behaviour tended to result in being more successful in groups of three. In particular, in groups of three, it seemed more successful to involve quiet speakers who were introverted but who had more than a certain proficiency level.

This can be illustrated by means of the interaction described in Excerpt (2), which was clearly orienting to expert/novice interactions in which *D*, who was the most proficient in the group, tried to elicit speech from *E* and *F* (see Table 8). The excerpt comes from

Table 8. Test-taker characteristics and performance summary [Task: Ranking]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
D	15	4	83	13	69
E	0	2	0	38	15
F	16	3	17	50	16

the second phase of the ranking task. In the first phase of the interaction, *D* kept scaffolding the interaction by providing sequence openers, supportive response tokens and effective use of body language, and then they moved on to the second stage of the task where they decided the three most important qualities.

Excerpt (2) test-taker: *D, E, F*

1	→	**D**:	OK. Let's choose three most important. (.5) So how about choose,
2			you choose one ((pointing at *F*)), you choose one ((pointing at *E*))
3			and I choose one? ((pointing at herself)) OK?
4		F:	OK.
5	→	**D**:	OK, hhh so what's your opinion? ((putting a hand towards *E*))
6		E:	I think love of students is important, because (.5) there is (2.0) not love thu-
7			love students, eh: in school, in school. ((inclining the head showing
8			embarrassment))
9	→	**D**:	OK. So love of students is the one opinion, do you agree with it? ((looking at *F*))
10		F:	Yes, I agree with it. ((smiling at *D*))
11	→	**D**:	Me too. OK, how about you? ((smiling and putting a hand towards *F*))
12		F:	I think sense of humour is the most important. Teacher (1.0) if teacher have
13			sense of humour, uh class be(.)come very fun [huh huh
14	→	**D**:	[Right Huh huh huh yeh:: true
15	→	F:	So, uh (3.0)
16		**D**:	OK.
17		F:	Hu hhuh huh
18	→	**D**:	So,
19	→	F:	How about you?
20		**D**:	Ah OK. Ah:: >I was gonna say< the same thing as you, but ah:: I should uh OK
21			Ah::: (1.0) it's difficult. Ah OK, so I think clear speaking voice is really important.

In line 1, *D* suggested that each of them should select one quality to reach the group decision of the three most important qualities and then invited *E*'s opinion using her hand. *E* chose *love of students*, and attempted to explain the reason why it was important. However, *E* failed to provide a reason and inclined the head showing embarrassment for the incapability. Here in line 9, *D* took over the topic, and nicely rounded it off by getting *F*'s agreement and showing her agreement with the choice. While exchanging smiles between *F* and *D*, *D* asked *F*'s opinion in line 11, which accordingly was answered in lines 12–13. After *D* gave response tokens in line 14, *F* attempted to continue her utterance. However, this attempt failed as shown in the three seconds of silence that followed. Then, *D* helped to terminate the topic by 'OK', which is often used as a device for the closure of a topic (Beach, 1993; Button, 1987), and showed her intention to give her ideas by another pre-shift token 'So' (Raymond, 2004) in line 18. Here, responding to *D*'s intention to move on to another topic, *F* asked *D*'s opinion. This was the first sequence opener directed to *D*, indicating that this group interaction had become a little more collaborative towards the end of the task. While *D*'s scaffolding behaviour resulted in producing the greater amount of talk and in initiating more topics, as a result of these actions, the task was cooperatively completed, considering all group members' opinions equally, including that of *E*'s, a very introverted test-taker.

Table 9. Test-taker characteristics and performance summary [Task: Info-gap]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
G	0	3	17	0	15
H	16	4	67	25	39
I	14	4	8	50	25
J	1	3	8	25	20

Although it was clear that *D* made an effort and she was skilled in scaffolding the interaction, the success of involving quiet speakers might partially rely on the smaller group size. This may not be merely because the smaller group size was more manageable to scaffold. It also seemed that introverted participants, who would prefer to interact with a smaller number of people (Eysenck & Eysenck, 1964), might have encountered a little less difficulty in participating in the smaller group size.

Avoidance behaviour in groups of four. In the above section, we have seen that involving quiet introverted members in groups of three seemed to be rather successful. In contrast, the data showed that quiet members in groups of four tended to remain quiet, even if some scaffolding actions were given. This seemed partly because they could more easily avoid talking in the larger group size. Excerpts (3) and (4) show such examples. Excerpt (3) was taken from the information-gap task performed by a group of four (see Table 9). In this group, *H* and *I* were rather extraverted while *G* and *J* were very introverted. Although *H* managed to involve a quiet test-taker *J* after the second attempt, the initial response of *J* suggested how possible it would have been for quiet members to avoid talking by simply agreeing with others.

Excerpt (3) test-taker: *G, H, I, J*

1		H:	What do you think? ((making deliberate eye contact with *J*))
2		G:	Huh huh uh
3	→	J:	Me too.
4		I:	huh huh huh
5	→	H:	Camera C?
6	→	J:	huh huh And camera B is good for class too, because (.) uh:: this camera, this
7			camera has auto-focus system and (.) flash.
8		I:	Uh huh huh uh
9		J:	Uh uh: Each us, uh? (.5) each each of students, each of students surface is clea(h)r.

Excerpt (3) begins after all the members except for *J* have provided their opinions. In line 1, *H* asked for *J*'s idea, while deliberately making eye contact with her, since *J* had kept her head down watching a prompt card rather than other group members. *J* said only 'Me too' in line 3, just agreeing to the previous speaker's preference. However, in line 5, *H* persistently asked about the camera to which *J* was agreeing. Only after this second attempt, did *J* provide a little longer talk in lines 6–9, which revealed that *J* actually liked Camera B, not Camera C, although she could not fully make herself understood (see line 9).

There was another type of avoidance behaviour that some quiet members in groups of four utilized. In the following group of four where proficiency levels and extraversion

Table 10. Test-taker characteristics and performance summary [Task: Ranking]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
K	3	3	0	17	6
L	5	1	14	17	20
M	12	5	57	33	47
N	12	5	29	33	27

levels are mixed (see Table 10), despite the fact that *M* and *N* (with higher extraversion levels and higher proficiency levels) tried to involve the other two quiet members, it did not seem to be very successful. Excerpt (4) shows how *L* tried to avoid giving her opinion by giving a sequence opener directing another group member.

Excerpt (4) test-taker: *K, L, M, N*

1		N:	Do you have any any ( ) anything else?
2			(8.0)
3		N:	huh [huh
4		M:	[huh huh
5		M:	*Ryoko*?
6			(4.0)
7	→	L:	Do you think about clear speaking voice, *Azumi*?
8	→	M:	Uh:: Yeah, very good point. Uh:: Why do you think so, *Ryoko*?
9	→	L:	uh:: (2.0) That's good point, because we can't hear clearly (1.0) if I ( ) uh
10		M:	Uh:
11			(.5)
12		L:	Uh:: (1.0) but this point is so importance

In line 1, *N* tried to involve two quiet members, *K* and *L,* by providing them with an unspecific sequence opener. However, since it was not successful in eliciting speech from either, in line 5, *M* addressed *L* by name who was sitting in front of *M*. Then, after four seconds of silence, instead of giving her opinion, *L* returned another sequence opener back to *M*, although this time it was a specific sequence opener about *clear speaking voice*. Here, it is interesting to see that, rather than giving her opinion about *clear speaking voice*, *M* asked the reason why *L* had picked up *clear speaking voice* to ask about. Then, finally, *L* started giving her opinion in lines 9–12, although she gave up giving a reason in line 12.

We have so far seen two types of avoidance behaviour. Although such behaviour was observed both in groups of three and four, groups of four seemed to show more of these examples. These actions are akin to 'functional reduction strategies' (Faerch & Kasper, 1983), by which language learners reduce their communicative goal in order to avoid problems. In particular, the above two excerpts seem to show 'adjusted meaning' involving 'sacrifice of part of the optimal meaning, loss of precision or it may lead to a complete shift of the optimal meaning' (Varadi, 1983, p. 83). While the avoidance behaviour was eventually unsuccessful in the above examples, the unsuccessful actions

demonstrated that these test-takers actually tried to avoid the original meaning they wanted to convey, and when forced to provide their opinions, they both faced communication problems.

The above examples seem to suggest that the obligation of contributing to the group interaction might have been felt less in groups of four than in groups of three. This echoes Webb and Palincsar's (1996, p. 860) finding, from observations of group work in the classroom, that students in larger groups could ignore group members' questions or to shirk responsibility to answer questions. The result is also in accord with the assumption that Thorne (1987) made, after examining the mutual impact of extraverts and introverts on L1 dyadic conversational content and style. Thorne (1987) reported that in the verbal report sessions after the conversations, an introvert participant commented that she had been hoping to be put in a larger group so that she could have just sat back and taken it easy! This could be one of the reasons why the impact of the extraversion-level variables was smaller in groups of three than in groups of four. Nevertheless, while the smaller group dynamics seemed to have mitigated the influence of the extraversion-level variables, it did not work the same for the proficiency-level variables. This intuitively makes sense, as the contextual factors (i.e. group size, in this case) cannot force a low-proficiency-level test-taker to talk as much as they would force a low-extraversion-level test-taker to do so, provided he or she had a certain level of oral proficiency.

Mechanical turn-taking in groups of four. The most salient interactional feature observed more in groups of four than in groups of three was mechanical, unnatural turn-taking. Some groups of four pre-determined the turn-taking order, and followed the order either in a part of their interaction or throughout the whole interaction. This practice in groups of four seemed to be observed, regardless of different combinations of test-takers' extraversion levels and oral proficiency levels in groups. This might have mitigated the impact of both extraversion levels and oral proficiency levels exercised in groups of four, although it seemed that the mitigation of the extraversion-level effect was probably cancelled out by the increased effect of the variable in the manners described above.

In some groups of four, the turn-taking order was specified prior to the interaction by gesture. The following Excerpt (5) shows the very beginning part of the free discussion task performed by a group of four (see Table 11). In this group, *R* had a much lower proficiency level (at 1) than the other three test-takers, and *O* had a much higher extraversion level (at 20) than the other three.

Table 11. Test-taker characteristics and performance summary [Task: Free discussion]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
O	20	3	60	0	15
P	10	3	20	17	20
Q	14	3	0	50	54
R	14	1	20	33	10

Excerpt (5) test-taker: *O, P, Q, R*

1		O:	Have you ever been (.) have you ever going to date, date?
2			(1.0)
3	→	Q:	((drawing a circle with her index finger in the air))
4		O:	*Momoko*? Huh huh huh
5		P:	I have never, but I hear my friend

In this interaction, after *O* asked a question about group members' personal experience, *Q* drew a circle with her index finger in the air, suggesting turns should be taken in a counter-clockwise order. Then, *O* addressed *P*'s name to begin the turn-taking system starting from *P*. Such pre-determination of turn order completely deviates from a model of turn-taking in conversation, where 'ordering of speakers, being locally controlled (i.e. turn by turn), can vary' (Sacks, Schegloff, & Jefferson, 1974, p. 708).

Another gesture which was often utilized to determine the turn-taking order in groups of four is by putting a hand towards the next seated person to speak next. This body language was also utilized to invite a quiet speaker into the interaction, but in some groups of four, the same body language was employed just to pass the turn in the seating order. The following Excerpt (6) shows how the group members described in Table 12 took turns in a mechanical order. In this group, *S* and *V* had a little higher extraversion level than the other two test-takers, and *S* and *T* had a little higher proficiency level than the other two.

Excerpt (6) test-taker: *S, T, U, V*

1		T:	I I think clear, clear speaking voice is very important, because teacher, teacher,
:			
5	→	V:	((putting a hand towards *S* to speak up.))
6		S:	ah I think love of student is good way, because uh love, if do you have communicate with
:			
10	→	S:	((putting a hand towards *U* to speak up))
11		U:	I think uh clear writing is important, because if teacher write writes English, it (.) is not it
:			
15		V:	Ah:: (.5) ah:: I think the most important is ah:: ab- ability to organise class. Maybe I think
:			
19	→		(1.5) ((all of the students watching the interlocutor, as if they have finished the task))

Table 12. Test-taker characteristics and performance summary [Task: Ranking]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
S	12	4	0	25	17
T	8	4	17	50	17
U	10	3	17	25	21
V	13	3	67	0	45

20 Interlocutor: So which three things do you think are the most important?
21 V: Clear %*nandakke*% [speaking voice, ((*V* putting a hand towards *T*))

By pointing out each other (in lines 5 and 10), all participants said which item they thought was the most important in a counter-clockwise way. Nobody ratified or commented on the others' idea, and thus this is the parallel pattern of interaction as defined in Galaczi (2004). After everybody finished their turns, they looked at the interlocutor in silence in line 19, as if they had completed the task. At that point (in line 20), the interlocutor facilitated them to move on to the next decision making phase. Then, the group started sharing their ideas again in exactly the same counter-clockwise turn-taking order by using their hands.

The discourse in the groups of four was also marked by their frequent, mechanical use of 'How about you?' and 'What do you think?' to nominate the next speaker. While these sequence openers were often utilized both in groups of three and four, their mechanical use seemed more frequently observed in groups of four. For instance, in the following Excerpt (7), a group of four (see Table 13) passed the turns in a clockwise manner, by always using 'How about you?'. In this group, *Z* had a little less extraversion level (at 7) than the other three group members, and *Y* had a little less proficiency level (at 2) than the other three.

Excerpt (7) test-taker: *W, X, Y, Z*

1 Y: so (.5) uh:: teacher's word is uh:: important, so clear writing is (.3) most
2 importa(h)nt. Huh huh
3 (1.5)
4 → Y: How (.) about you? ((smiling at *X*))
5 (.5)
6 X: I think clear speaking voice is important, because (1.0) .hhh if teacher speak (.)
7 ah: if teacher doesn't speak clear, we can't listen what she said, so clear
8 → speaking voice is important. huh How about yo(h)u? huh ((smiling at *W*))
9 W: I think love of student is the mo- most important, because if if the teacher loves us,
:
12 → W: So I think love of student is the most important. How about you? ((looking at *Z*))

In this example of the relatively mechanical use of turn giving patterns, it is understandable that the first speaker *Y* utilized 'How about you?' in line 4, as she did not receive any feedback or continuation after giving her opinion, as shown in the 1.5 seconds of

Table 13. Test-taker characteristics and performance summary [Task: Ranking]

ID	Extraversion	Oral proficiency	Topic initiation (%)	Topic ratification (%)	The amount of talk (%)
W	14	3	25	33	29
X	12	3	25	33	21
Y	14	2	50	0	20
Z	7	3	0	33	20

silence. Therefore, this 'How about you?' seemed a genuine invitation for the next speaker. However, the following two 'How about you?' expressions in lines 8 and 12 sounded more unnatural. They did not even give any space for others to come in, and immediately on the completion of giving ideas, the speakers said 'How about you?' and smiled or looked at the next person seated, as if it had signalled the end of their responsibility in talking.

The series of questionings using 'How about you?' to the next person seated without giving any time for others' self-selection of turns generates unnaturalness. This is not only because turn-taking should normally be locally managed at each Transition Relevance Place (by providing options for different ways of current speaker's selection of the next speaker, or self-selection), but this is also because 'question' is one example of the first part of a sequential unit, 'adjacency pairs' (Sacks, Schegloff, & Jefferson, 1974), which binds what could come in the second pair part, and therefore, the interaction ends up with markedly repeated patterns of the same sequences.

The same use of 'How about you?' was also observed in Van Moere's (2007) examination of group discussion tasks. As for this issue, he describes it as a unique feature of the free discussion task, unlike the other consensus reaching or picture tasks. While this feature in the present study was also observed more in the free discussion tasks, this method of giving turns could also be related to group size. In fact, the example that Van Moere (2007, p. 325) reported is also from the interaction in a group of four (while the study involved both groups of three and four).

In addition, the unnaturalness of turn-taking in groups of four was especially highlighted by the lack of *schisming* (Egbert, 1997; Sacks et al., 1974; Schegloff, 1995). Schisming is a transformation where one conversation splits up into two or more simultaneous conversations. This systematically happens in a conversation with at least four participants, and this is one of the most salient phenomena differentiating interactions between groups of three and groups of four (Sacks et al., 1974, p. 713). Nevertheless, among all the recordings of groups of four in the present study, there was no group which exercised schisming, which was surprising considering the number of the recordings. The lack of schisming might have been caused by the incompatibility between testing the group interactions of four candidates, and the opportunity to create a natural turn-taking system among them. That is, as Simpson (2006) articulates, test-takers clearly know that in oral testing they have to display their language abilities so that the examiners can rate the language, and therefore their ultimate target audience is the examiners rather than the other candidates in the group (He & Dai, 2006, p. 389). Thus, they may have unconsciously avoided schisming, as simultaneous talk could make the rating of individual speakers' language more difficult. This incompatibility in talking naturally in groups of four in oral testing occasions could be a partial reason for the unnatural way of turn-taking observed and described above.

Conclusion

This study has compared the impact of two test-taker characteristics (extraversion levels and oral proficiency levels) on conversational styles between groups of three and groups of four. The quantitative results showed that these test-taker characteristics had different

degrees of effects on discourse when mediated by the different group sizes. To summarize, extraversion level (conversely, introversion level) mattered more in groups of four than in groups of three, and whilst there was an influence of the proficiency-level variables in both group sizes, the effect size was larger in groups of three than in groups of four. Close examination of the discourse helped elucidate reasons behind the difference, revealing some interactional features that could explain the relationship between these impacts and co-constructed group interactions.

Limitations of the study and future directions

There are two main limitations in this study. First, test-takers' characteristics might have been measured more comprehensively. For extraversion, self-judgement of extraversion using the Japanese EPQ was the only measure of the variable. Teacher judgement was used as an alternative measure of the participants' level of oral proficiency. Especially, for proficiency levels, these judgements may lack reliability, and it may have been preferable to use a more formal instrument in line with the original study plan with fully trained and standardized raters. However, given that the study was based on fairly broad distinctions between proficiency levels, the teacher judgements were probably adequate for this purpose.

Second, as mentioned earlier, the between-condition design utilized in multiple regression analysis made the data set violate one assumption for the analysis, and this has limited the results of statistical significance presented to being at best suggestive. Thus, it may be of value to replicate the present study in the future employing a within-subjects design having the same individuals perform under each of the conditions.

Implications of the findings

Despite the above limitations, the present study has several implications in the areas of language testing research and practice.

This study has strengthened the interdisciplinary link between research into oral assessment and CA, by building on earlier contributions of CA in language assessment (e.g. Lazaraton, 2002). Moreover, while coding and quantification in CA is somewhat controversial (Lazaraton, 2002, pp. 82–87; Schegloff, 1993), by using a large-scale data set for a CA study, this research has exemplified a possible way of combining quantitative and qualitative methods in order to foster a more in-depth understanding of the mechanics of oral assessment.

The present research has clearly shown that a test-taker's characteristics, his or her group member's characteristics, and task implementation conditions affect the resulting test-takers' discourse in group oral tests. This suggests that if group oral tests are to be used, we should adopt the interactionalist view of construct definition (e.g. Brown, 2003). Regarding the impact of group members' characteristics, the results of this study support Brown's (2003) and Chalhoub-Deville's (2003) views that variability caused by group members is important if we are assessing interpersonal communication skills. This is because, as far as the test-taker characteristics examined in this study were concerned, the examples have shown that group members' characteristics played a significant role in

contributing to the creation of a context where some test-takers display their communication skills to fit their conversants, as would actually occur in non-test interaction.

One could manipulate the degree of influence of test-takers' extraversion level on the discourse depending on the group size. If test designers are interested in minimizing the effect of extraversion-level variables on test-takers' performance, groups of three would appear to be more appropriate than groups of four. Furthermore, it was found that interaction in groups of four could be more mechanical and unnatural as a result of introducing a pre-determined turn-taking system. In this sense, grouping test-takers into groups of four might not always provide a suitable environment for test-takers to display their communication ability. This suggests that test designers should consider which group size would be the best to operationalize the test construct. Similarly, the findings could also be useful for classroom teachers to decide the optimal group size for group work, depending on their teaching purposes and what types of turn-taking they want students to exercise.

Acknowledgements

I gratefully acknowledge the insightful comments made by the Language Testing reviewers and the Editors that helped improve this article. I would also like to thank my PhD supervisor, Mr Tony Lilley, and Professor Cyril J. Weir, Dr Stephen Bax and Dr Tony Green for their suggestions and comments on earlier versions of this article.

References

Atkinson, J. M., & Heritage, J. (1984). *Structures of social action.* Cambridge and New York: Cambridge University Press.

Beach, W. A. (1993). Transitional regularities for 'casual' 'Okay' usages. *Journal of Pragmatics, 19*, 325–352.

Berry, V. (1993). Personality characteristics as a potential source of language test bias. In A. Huhta, K. Sajavaara, & S. Takala (Eds.), *Language testing: New openings* (pp. 115–124). Jyvaskyla, Finland: Institute for Educational Research.

Berry, V. (1997). Ethical considerations when assessing oral proficiency in pairs. In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current developments in language testing* (pp. 107–123). Jyvaskyla: Jyvaskyla University Press.

Berry, V. (2004). A study of the interaction between individual personality differences and oral performance test facets. Unpublished PhD thesis, King's College University of London.

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.

Bonk, W. J., & Van Moere, A. (2004). L2 group oral testing: The influence of shyness/outgoingness, much of interlocutors' proficiency level, and gender on individual scores. Oral presentation at The Language Testing Research Colloquium, California, USA.

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26*(3), 341–366.

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

Button, G. (1987). Moving out of closing. In G. Button & J. R. E. Lee (Eds.), *Talk and social organisation* (pp. 101–151). Clevedon and Philadelphia: Multilingual Matters.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20*(4), 369–383.

Costa, P. T., & McCrae, R. R. (1986). Major contributions to psychology of personality. In R. Modgil & C. Modgil (Eds.), *Hans Eysenck: Consensus and controversy* (pp. 63–72). Philadelphia: The Falmer Press.

Coulson, D. (2005). Collaborative tasks for cross-cultural communication. In C. Edwards & J. Willis (Eds.), *Teachers exploring tasks in English language teaching* (pp. 127–138). New York: Palgrave Macmillan.

Csépes, I. (2002). Is testing speaking in pairs disadvantageous for students? A quantitative study of partner effects on oral test scores. *novELTy, 9*(1), 22–45.

Davies, L. (2009). The influence of interlocutor proficiency in a paired oral assessment. *Language Testing, 26*(3), 367–396.

Drummond, R. J. (1990). Eysenck personality inventory. In J. Hogan & R. Hogan (Eds.), *Business and industry testing: Current practices and test reviews* (pp. 228–239). Austin, TX: Pro Ed.

Egbert, M. M. (1997). Schisming: The collaborative transformation from a single conversation to multiple conversation. *Research on Language and Social Interaction, 30*(1), 1–51.

Eysenck, H. J., & Eysenck, S. B. G. (1964). *Manual of the Eysenck personality inventory.* London: Hodder & Stoughton.

Eysenck, M. W. (1981). Learning, memory and personality. In H. J. Eysenck (Ed.), *A model of personality* (pp. 169–209). Berlin: Springer Verlag.

Faerch, C., & Kasper, G. (1983). Plans and strategies in foreign language communication. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 20–60). New York: Longman.

ffrench, A. (2003). The change process at the paper level. Paper 5, Speaking. In C. Weir & M. Milanovic (Eds.), *Continuity and innovation: Revising the Cambridge Proficiency in English examination 1913–2002* (pp. 367–471). Cambridge: Cambridge University Press.

Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *ELT Journal, 30*, 156–167.

Foot, M. C. (1999). Relaxing in pairs. *English Language Teaching Journal, 53*(1), 36–41.

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

Galaczi, E. D. (2004). Peer–peer interaction in a paired speaking test: The case of the First Certificate in English. Unpublished PhD thesis, Columbia University.

Galaczi, E. D. (2010). Interactional competence across proficiency levels. Oral presentation at the 32nd Language Testing Research Colloquium, University of Cambridge, UK. April 2010.

Gan, Z. (2010). Interaction in group oral assessment: A case study of higher- and lower-scoring students. *Language Testing,* 27(4), 585–602.

Gan, Z., Davison, C., & Hamp-Lyons, L. (2008). Topic negotiation in peer group oral assessment situations: A conversation analytic approach. *Applied Linguistics, 30*(3), 315–334.

Gardner, R. (2007). The *Right* connections: Acknowledging epistemic progression in talk. *Language in Society, 36*, 319–341.

Goodwin, C., & Goodwin, M. H. (1990). Interstitial argument. In A. D. Grimshaw (Ed.), *Conflict talk: Sociolinguistic investigations of arguments in conversation* (pp. 85–117). Cambridge: Cambridge University Press.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing, 23*(3), 370–401.

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

HKEAA (2005). *Hong Kong Advanced Level Examination – Subject examination report and question papers (with marking schemes). AS Use of English 2005.* Hong Kong: Hong Kong Examinations and Assessment Authority.

Hutchby, I., & Wooffitt, R. (1998). *Conversation analysis.* Cambridge: Cambridge University Press.

Iwashita, N. (1998). The validity of the paired interview in oral performance assessment. *Melbourne Papers in Language Testing, 5*(2), 51–65.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to test design. *Language Learning, 51*(3), 401–436.

Iwawaki, S., Eysenck, S. B. G., & Eysenck, H. J. (1980). Japanese and English personality structure: A cross-cultural study. *Psychologia, 23*, 195–205.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger & P. Bull (Eds.), *Conversation: An interdisciplinary perspective* (pp. 166–196). Clevedon: Multilingual Matters.

Kobayashi, M., & Van Moere, A. (2004). Group oral testing: Does amount of output affect scores? Paper presented at Language Testing Forum, Lancaster University.

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

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

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

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.* Cambridge: Cambridge University Press.

Lazaraton, A., & Davies, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment. *Language Assessment Quarterly, 5*(4), 313–335.

Lerner, G. (1993). Collectivities in action: Establishing the relevance of conjoined participation in conversation. *Text, 13*, 213–245.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

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

Masubuchi, S. (2003). *Proceedings of the English teacher conference in Ibaraki.* Mito: Kanto Koshin Chiku Chugakko Eigo Kyouiku Kenkyugikai.

May, L. (2007). Interaction in a paired speaking test: The rater's perspective. Unpublished PhD thesis, the University of Melbourne.

McNamara, T. F. (1996). *Measuring second language performance.* Harlow: Longman.

Nakamura, Y. (2003). Oral proficiency assessment: Dialogue test and multilogue test. *2003 JALT Pan-Sig conference proceedings* online, accessed on 24 July 2007 at http://jalt.org/pansig/2003/HTML/Nakamura.htm.

Nakatsuhara, F. (2004). An investigation into conversational style in paired speaking tests. Unpublished MA thesis, the University of Essex.

Nakatsuhara, F. (2009). Conversational styles in group oral tests: How is the conversation co-constructed? Unpublished PhD thesis, the University of Essex.

Norton, J. (2005). The paired format in the Cambridge speaking tests. *English Language Teaching Journal, 59*(4), 287–297.

Nunn, R. (2000). Designing rating scales for small-group interaction. *English Language Teaching Journal, 54*(2), 169–178.

Ockey, G. J. (2001). Is the oral interview superior to the group oral? *Working Papers, International University of Japan, 11*, 22–40.

Ockey, G. J. (2006). Making a case for the group oral discussion test: The effects of personality on the group oral's score-based inferences. Unpublished PhD thesis, the University of California, Los Angeles.

O'Sullivan, B. (2000a). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373–386.

O'Sullivan, B. (2000b). Towards a model of performance in oral language testing. Unpublished PhD thesis, the University of Reading.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*(3), 277–295.

O'Sullivan, B. (2004). Modelling factors affecting oral language test performance: A large scale empirical study. In M. Milanovic & C. Weir (Eds.), *European language testing in a global context: Studies in Language Testing 18* (pp. 129–142). Cambridge: Cambridge University Press.

O'Sullivan, B., & Nakatsuhara, F. (in press). Quantifying conversational styles in group oral test discourse. In B. O'Sullivan (Ed.), *Language testing: Theories and practices.* London: Palgrave.

Pavlou, P. (1995). Assessing different speech interactions in an oral proficiency test. Unpublished PhD thesis, Georgetown University.

Raymond, G. (2004). Prompting action: The stand-alone 'so' in ordinary conversation. *Research on Language and Social Interaction, 37*(2), 185–218.

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

Sacks, H. (1984). Notes on methodology. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 21–27). Cambridge: Cambridge University Press.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

Schegloff, E. A. (1980). Preliminaries to preliminaries: 'Can I ask you a question?' *Sociological Inquiry, 50*, 104–152.

Schegloff, E. A. (1993). Reflections on quantification in the study of conversation. *Research on Language and Social Interaction, 26*(1), 99–128.

Schegloff, E. A. (1995). Parties and talking together: Two ways in which numbers are significant for talk-in-interaction. In P. Ten Have & G. Psathas (Eds.), *Situated order: Studies in the social organization of talk and embodied activities* (pp. 31–42). Washington, DC: University Press of America.

Shohamy, E., Reves, T., & Bejarano, Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal, 40*(3), 212–220.

Simpson, J. (2006). Differing expectations in the assessment of the speaking skills of ESOL learners. *Linguistics and Education, 17*, 40–55.

Skehan, P. (1998). *A cognitive approach to language learning.* Oxford: Oxford University Press.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119–158.

Stubbe, M., Lane, C., Hilder, J., Vine, B., Marra, M., Holmes, J., & Weatherall, A. (2003). Multiple discourse analyses of a workplace interaction. *Discourse Studies, 5*(3), 351–388.

Thorne, A. (1987). The press of personality: A study of conversation between introverts and extraverts. *Journal of Personality and Social Psychology, 53*(4), 718–726.

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

van Lier, L. (1989). Reeling, writhing, drawing, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly, 23*(3), 489–508.
Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing, 23*(4), 411–440.
Van Moere, A. (2007). Group oral test: How does task affect candidate performance and test score? Unpublished PhD thesis, the University of Lancaster.
Varadi, T. (1983). Strategies of target language learner communication: Message adjustment. In C. Faerch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 79–99). New York: Longman.
Webb, N. M., & Palincsar, A. S. (1996). Group processes in the classroom. In D. Berliner & R. Calfee (Eds.), *Handbook of educational psychology* (pp. 841–873). New York: Macmillan.
Wigglesworth, G. (2001). Influences on performance in task-based oral assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 186–209). London: Pearson Education.
Young, R. (1995). Conversational styles in language proficiency interview. *Language Learning, 45*(1), 3–42.
Young, R., & He, A. W. (1998). *Talking and testing: Discourse approaches to the assessment of oral proficiency.* Amsterdam and Philadelphia: John Benjamins.
Young, R., & Milanovic, M. (1992). Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition, 14*, 403–424.

Appendix: Transcription notation (based on Atkinson & Heritage, 1984; Hutchby & Wooffitt, 1998)

Unfilled pauses or gaps	Periods of silence. Micro-pauses (less than 0.2 seconds) are shown as (.); longer pauses appear as a time within parentheses.
Colon (:)	A lengthened sound or syllable; more colons prolong the stretch
Dash (-)	A cut off
.hhh	Inhalation
hhh	Exhalation
hah, huh, heh	Laughter
(h)	Breathiness within a word
Punctuation	Intonation rather than clausal structure; a full stop (.) is falling intonation, a question mark (?) is rising intonation, a comma (,) is continuing intonation
Equal sign (=)	A latched utterance
Open bracket ([ )	Beginning of overlapping utterances
Percent signs (% %)	Quiet talk
Empty parentheses ( )	Words within parentheses are doubtful or uncertain
Double parentheses (( ))	Non-vocal action, details of scene.
Arrows (><)	The talk speeds up
Arrows (<>)	The talk slows down
<u>Underlining</u>	A word or sound is emphasized
Italics	Japanese words
Arrow (→)	A feature of interest to the analyst