BIAS SEQUENCES AND TURN-TAKING IN THE VALIDATION OF THE SMALL GROUP ORAL EXAMINATION

doi.org/10.61425/wplp.2019.13.82.107

Gergely A. Dávid
School of English American Studies, Eötvös Loránd University Budapest
david.gergely@btk.elte.hu

Abstract: This research originated in an interest in the characteristics of communication in a group oral test. The researcher's attention was directed towards its turn-taking mechanisms and how communication is shaped in groups of three. The investigation of the test discourse, with reference to insights and predictions from Conversation Analysis, yielded external validity evidence for the conversational nature of the interaction in this test. The findings support the claim that in Phase two of the group oral test the interaction is sufficiently conversational. True to the nature of conversation, each of the recorded exams contained bias sequences, in which interaction involves only two of the speakers from time to time. The findings also suggest a research-based rationale for the number of speakers desirable in a group oral test in that it may be a happy choice to examine students in threes, rather than twos or larger numbers.

Keywords: validation, group oral tests, conversation, turn-taking, test discourse

1 Introduction

In the author's experience, group orals have often been put down as a clever way to test a large number of students in an efficient and economical way, without having to expend much consideration towards reliable and valid testing. Thus, this research originated in a need to show whether there was more to group orals than expediency and efficiency. More importantly, the goal was to collect validity evidence for (or against) the group format at the author's university.

1.1 The research context

The research context was the MA Language Test (now the OTAK-MA Language Test, formerly the Test of Language Competence), a pre-service filter test at the Department of Language Pedagogy at Eötvös Loránd University, Budapest, originally developed for an earlier teacher training programme. Nearly fifteen years later, the test continues to be used with a summative and gatekeeping function. In this paper, the group test of successive academic programmes will be referred to as the Eötvös University group oral. Below is a short description of key design features of the group oral.

1.2 Important design elements

The Eötvös University group oral comprises two phases. Phase 1 elicits extended chunks of language, dominantly monologic in form, but this phase is not dealt with in this study. This study focuses on Phase 2, which is designed to test conversational skills in numerous shorter turns. The two-phase design was influenced by Ochs' distinction of "relatively planned" and "relatively unplanned" (1979, p. 55) discourse categories. It is relatively planned if what one has to say has been thought out previously, and relatively unplanned if the interaction is spontaneous. Phase 1 is preceded by five minutes' preparation time whereas Phase 2 is initiated by the examiners' spoken prompt, leaving no preparation time.

Three relevant design elements of the Eötvös University group oral need to be highlighted. First, the exam tasks require minimal interlocuting. They are so designed that interaction can move ahead without substantial examiner interlocution, the exceptions being the examiner submitted task for Phase 2 and the rescue question, resorted to only in the case of not getting enough evidence for the rating. Whether the interaction in the group oral is driven by the task or is examiner-led is important because a focus on student-to-student interaction was not a logical necessity in early versions of the group oral. Morrison and Lee (1985), for example, report a teacher-led discussion with one teacher in the interlocutor's role. Day and Shapson (1987) report a group discussion in which the examiners appear to lead the discussion, "intervening as little as possible" (pp. 241-242), but no further information is put forward. Scott (1986) does not even specify the extent of the examiners' involvement as interlocutors. It is surprising to read about such designs now because if allowing unhampered and uninterrupted candidate interaction is important, it seems counterproductive to smuggle interlocuting examiners back into the action.

The second significant feature of the Eötvös University group oral is that students take the test in threes (trios), with no allowances for groups of four or five. In the running of the exam, fellow students called helpers are asked make up the trios, should some students not find their partners. In the early literature, by contrast, the number of test-takers varies greatly. Fulcher's (1993) were groups of three. Swain's (1985) were groups of three or four. Shohamy et al. (1986), Reves (1991) and the Hong Kong Examinations Authority (1995) report four students. Morrison and Lee (1985), Day and Shapson (1987), and Hilsdon (1991) all report five. Six students were reported in Lazaraton (1996) and Scott (1986). Folland and Robertson (1976) and Liski and Puntanen (1983) report group sizes between five and seven, whereas Shectman (1988) reports as many as eight and Hutchinson (1986) does not even state their number! More recently, there has been less of a range in the number of candidates reported, the most often reported setups being three or four (Bonk & Ockey, 2003; He & Dai, 2006; Leaper, 2010; Nakamura, 2003, 2005; Ockey, 2009; Van Moere, 2006, 2007;).

The third important design element is that students form the groups themselves, rather than be combined by the organisers. Used since inception, this technique places about 90% of the students efficiently. Leaving it to the students to form groups had its source in the designers' awareness that the different personal characteristics in planned or truly random combinations might cause problems. An explicit rationale for similar situations was articulated as the *bias for best* principle for communicative language testing (Swain, 1985).

2 Theoretical background

2.1Early work with the group oral

The group oral was put to a variety of educational uses towards the end of the last century. The earliest discussion was by Folland and Robertson (1976) at a university in Finland. In the 1980s, the group oral generated considerable interest in Israel: Reves (1980, 1982, 1991), Berkoff (1985), Shohamy, Reves, and Bejarano (1986) described the design of an oral battery for the English matura. In Shectman (1988), the individual and group interview procedures were compared for their predictive validity in a teacher-training programme. In Morrison and Lee (1985), the group oral predicted academic success at the University of Hong Kong. In Canada, Day and Shapson (1987) tested children in French immersion programmes. Hilsdon's (1991) test was a selection device for further study in Zambia. In contrast with group orals in a foreign language, Hutchinson (1986) indicated a trend in Britain to assess students in L1.

However, as Skehan wrote (1991), the group oral was shunned by testers for a long time in the past because the equal opportunities in testing individually weighed more heavily than either the naturalness or symmetrical power relations that groups seemed to offer. He noted, however, that Gorman and Brooks (1986) and Hutchinson (1986) were successful experimentations with group techniques. Concerns, for example, Pavlou (1997), were raised about the difficulty of simultaneously examining a group as examiners would have their attention divided and equal opportunities would suffer as a result. In addition to the paired format, group testing was an alternative at best to the traditional interview-type test. Despite the doubts, Carroll (1980) and Canale (1984) recommended group techniques early on.

2.2 More recent work

The new millennium saw a surge of studies, many of them from Japanese educational contexts. As a result, it had become a much better researched test format (Bonk & Ockey, 2003; Gan, 2010; He & Dai, 2006; Leaper, 2010; Nakamura, 2003, 2005; Nakatsuhara, 2009, 2011; Nunn, 2000; Ockey, 2001, 2006, 2009; Van Moere, 2006, 2007). Most of these studies may be called outcome-based (Lazaraton, 2002), drawing inferences on the basis of scores (Bonk & Ockey, 2003; Nakamura, 2003, 2005; Ockey, 2001, 2006, 2009; Van Moere, 2006). The group oral, however, was still not well-researched in terms of its discourse, the most relevant studies being He and Dai (2006), Gan (2010) and Leaper (2010). Van Moere (2007), Nakatsuhara (2009, 2011, 2013) are analyses of scores and test discourse.

The ambition of these recent studies was to establish whether the group oral might be a viable alternative to the interview test. In general, the score-based studies found that the dominant source of variance in the scores is candidate oral proficiency. Minor sources of variance were those of task, the rating (raters, rating scales) and the testing occasion. Many-facet Rasch Measurement (MFRM) allowed compensating for the effect of these typical performance variables. Ockey (2001) found that an interview test may not be a better alternative. Bond and Ockey (2003) concluded that the group oral "may be a viable short-cut means of estimating the speaking ability of large numbers of examinees more quickly and efficiently than by using interview or other methods" (p. 103), but they were worried about the

effects of further variables (gender and age differences, anxiety, reaction to strangers, shyness, social power, turn-taking, and an array of proficiency levels present in the group) not investigated in their study. Nakamura (2005) concluded that foreign language proficiency is best measured in different discourse modes; through monologues, dialogues as well as multilogues. Van Moere's (2006) findings were encouraging for the group format since variation due to the rating was low. Variation due to the testing occasion facet was higher but still acceptable. However, "there was something as yet unexplained [he concluded] in the testing situation that causes persons to elicit different scores from raters in different test occasions" (p. 435). Suspecting social factors related to candidate interlocutors or group dynamics, Van Moere called for further investigation of unpredictable interaction dynamics and text analysis. This agenda was largely taken up by Ockey (2009), who investigated the variability of scores as a result of differences between assertive and non-assertive personalities in the groups, concluding that the personal characteristics of the members of a group (group membership) can affect a test taker's score.

The discourse-based studies were naturally much smaller scale and have brought mixed results so far. He and Dai (2006) investigated transcripts from a random sample of groups and were disappointed by the low level of interaction. Leaper (2010) revealed that for the most part features of conversation were represented, but where they were not, interaction was not conversation-like at all. Gan (2010) studied only two groups where genuine but very different interactions took place. In the higher group, interaction was constructive and contingent on each other's input. In the lower group, interaction, surprisingly, was also effective due to the negotiation of meaning over linguistic impasses.

2.3 The group oral: past its prime?

Even more recently, the group oral has come under scrutiny and criticism, not in itself but as one of the assessment techniques where there is a contradiction between the co-constructed nature of interaction and the need to award individual scores for each candidate's performance (Wigglesworth, May, Galaczi, Nakatsuhara, & Van Moere, 2010). In this respect, all other oral assessment formats (one-to one interviews, the Oral Proficiency Interview (OPI), the paired oral and the group oral) may receive criticism because due to the co-constructed interaction, the performance rated in the test is not entirely that of the candidate's, but also the interlocutor's. The brunt of the criticism, however, logically goes against the group oral. In this light, those who advocated the group oral over, say, the 2OPI, should have merely substituted the variance generated by another candidate for the variance generated by the OPI examiner (Brooks, 2009; Brown, 2003; May, 2009; McNamara, 1997).

Recent research has attempted to deal with the problem of interlocutor (speaker) variability (Ockey, 2009), the term meaning both examiners in an OPI and/or other test-taking members of a group. Interlocutor variability concerns whether and how speakers in the necessarily co-constructed interaction affect each other's output and scores. Public language examinations seem to take the approach to regulate the interlocutor's techniques so that such variability (and the resulting variance) is minimised. In the future, the debate is likely to focus on whether such score variance is related to the construct of speaking or not, that is, whether it constitutes construct-relevant variance (Messick, 1995) and if so, whether interlocutors can discount it in their scores.

2.3 Validation

Mainstream test validation still follows the groundwork laid down by Messick (1989, 1995), followed up by its adaptation more recently by Kane (2006) stipulating an argument-based approach. Claims of validity, however, if they are not well-supported by empirical evidence, on grounds of the task type, for example, would essentially be claims for face validity (Fulcher, 1996). Clearly, a more thorough approach is needed to go beyond face validity. Whereas a validity argument may be constructed from analyses of the scores, another source of evidence is what actually happens in the test, the analysis of its discourse, which may or may not support the interpretation that the test is a test of communicative competence, oral communication skills in our case.

A degree of internal validation had always been achieved with the Eötvös University group oral through regular (classical and Rasch) analyses as part of the standard quality control procedures. These increased the generalizability of the results by eliminating construct-irrelevant variance (rater and item (task) effects), but could not be counted as validation with respect to an external criterion. The purpose of this study was to collect discoursal evidence from the group oral with a view to creating a validity argument for its external validity by investigating whether turn-taking in the group interaction was sufficiently like or similar to that of conversations in non-testing situations.

2.3.1 The choice of conversation

Earlier validation studies, for example, Johnson (2001) concluded that the OPI was decidedly non-conversational, given the strong interviewer drive to obtain a maximum of information from the candidate, in exchange for their own minimised but purposeful interlocuting. Of course, the inadequacy of the OPI does not guarantee that the group oral will elicit conversational language. In this section, an explanation will be put forward about why conversation was selected in this study to be the external criterion for validation.

First and foremost, conversation is a very basic and widespread kind of language use according to most views: Typical adjectives in the literature include “prototypical”, “central” and “most basic” (Levinson, 1983, p. 284). Typical phrases include “predominant medium” and “primary form [of interaction]” (Drew & Heritage, 1992, p. 19). It is also described as having a “bedrock status” (Atkinson & Heritage, 1984, p. 12) and, more recently, as the “default version of talk” (Gardner, 2006). According to both Levinson (1983) and Drew & Heritage (1992), natural conversation should be a point of reference because that is what a child is first exposed to. As Bachman and Palmer (1996) state, it is the target language use (TLU) of non-test situations that language test performance ought to represent. For a study about the Eötvös University group oral, it was perhaps natural to choose conversation as the external criterion. The designers of the test had chosen conversational English as the focus of testing for Phase 2 at inception.

There may be a problem of labelling various types of interaction as ‘conversation’. Fowler et al. (1979), for example, state that communicative relationships are generally asymmetrical, saying that in conversation “any appearance of intimacy, solidarity and cooperation is generally illusory. Speakers act out their socially ascribed roles” (p. 63). This leaves the reader wondering what Fowler et al. might have understood to be conversations. In

a study essentially very similar to this one, Leaper (2010) labels their own interaction as group discussion, casting some doubt at the same time on conversation being the suitable criterion for the group oral.

Conversation, of course, does not seem the only potentially suitable criterion for such a validation exercise. There is a large variety of institutional interaction types, but it is clear, even intuitively, that these 'talks' are different from conversations. Research showed important differences between conversation and talk in institutional contexts (Atkinson & Heritage, 1984; Drew & Heritage, 1992; Levinson, 1983). Various types of interactions in institutional settings may be seen – at best – as special conversations where turn-taking is "strongly constrained within quite sharply defined procedures" (Drew & Heritage, 1992, p. 580). Most importantly, however, institutional talk did not seem appropriate for this research as an external criterion because it may be construed as a simplification, a culturally and socially restricted form of conversation (Gardner, 2006). By comparison, 'real' conversation does not have a structured order of successive stages, with the exceptions of structured openings and closings. The differences between institutional talk and conversation are accounted for by the differences in the turn-taking systems and the task orientation of institutional interaction. In sum, it may be stated that conversation as the external validity criterion is not unanimously approved, but it is still sufficiently central to warrant its position as the external criterion here.

2.3.2 The theoretical basis for this study

For theoretical underpinnings, the Conversation Analysts Sacks et al. (1974) were turned to, rather than Sinclair and Coulthard (1975) and their followers, for example, Tsui (1994). The latter school of thought was originally couched in the institutional context of schools, which could not be expected to generate typically conversational language. Interestingly, few of the reports on the group oral here (Leaper, 2010; Nakatsuhara, 2009, 2011; Nunn, 2000; Van Moere, 2007) refer to the findings of Sacks et al., who, having transcribed vast amounts of audio recordings of native speaker talk, identified 14 generic features of conversation (pp. 700-701, reproduced below in a simplified form):

1. Speaker change recurs.
2. Overwhelmingly, one party talks at a time.
3. Overlaps in speech are common, but brief.
4. The vast majority of transitions (from one turn to a next) are with no gap or overlap or with a slight gap or overlap.
5. In interaction for more than two speakers turn order is not fixed but varies.
6. Turn size varies.
7. Length of conversation is not specified in advance.
8. What parties say is not specified in advance.
9. In interaction for more than two speakers the distribution of turns is not specified in advance.
10. Number of parties can vary.
11. Talk can be continuous or discontinuous.
12. Turn allocation techniques are obviously used.

13. Various ‘turn constructional units’ are employed; e.g., turns can be projectedly ‘one word long’, or they can be sentential in length.
14. Repair mechanisms exist for dealing with errors and violations.

The conversational features identified by Sacks et al. have largely withstood the test of time (Gardner, 2006) and stand as potential criteria, divorced from their originally emic approach. Although most conversations assume two speakers, Sacks et al. (1974) identified three special features (boxed above by the author): #5 and #9 can only apply to interaction by more than two, while #12 refers to more complex ways of interaction between more than two. If criteria by Sacks et al. are accepted as definitive, it may be argued that the full potential of conversation can only be realized when more than two participate. It should also be noted that all the three special features have to do with the management of turns. Thus, a clear mandate for this study was to investigate how the trios manage their interaction (turn-taking) in Phase 2 of the oral for conversational evidence.

2.3.3 Conversational bias: the effect of the number of speakers

Sacks et al.’s (1974) feature #12 merits special attention in this study. This feature allows three ordered options that govern turn-taking as rules or techniques:

- Rule 1: Current speaker may decide to select the next speaker.
- Rule 2: If Rule 1 does not operate, another speaker may select themselves as next.
- Rule 3: If Rule 2 does not work, the current speaker may decide to continue.

The result of these rules is what Sacks et al. call *bias* for the current speaker to select the previous speaker as next. It follows that if the next speaker takes the floor, that speaker, the current speaker at that time, will likely select the previous speaker as next. The tendency for bias is reinforced by the local management of conversation: at each transition-relevance place the previous speaker is almost invited to be the next speaker. Conversational bias thus predicts that interaction tends to “stabilise” between two speakers from time to time, in what are called bias sequences, when more than two speakers are present.

1.3.4 The number of speakers as reported in the literature

Strikingly, writers simply state the number of candidates, without putting forward a rationale for the size of groups (Bond & Ockey, 2003; Gan, 2010; He & Dai, 2006; Leaper, 2010; Ockey, 2001; Ockey, 2009; Van Moere, 2006). Nunn (2000) articulated a rationale for balanced interaction, but his focus was on designing rating scales. The other theorizer is Nakamura (2005), who developed a tripartite construct for speaking and had the number of speakers involved as a starting point. Nakamura posited that in order to be able to test oral proficiency, one must test it through monologue, dialogue and multilogue, that is, group techniques. Very interestingly for the focus here, Nakamura further divided the construct into multilogues by small groups and large groups.

Specifying four or more students in a group amounts to ignoring what is known about the nature of conversation. Sacks et al. (1974) found a schism might occur when (at least) four

speakers are present, who might carry on parallel conversations, stating that such a schism is a systematic possibility, built into the conversational system because conversation organises only two of the speakers at any particular time. Whereas this sort of schism is not a problem, for example, in casual chit-chat over dinner, it may, however, be a threat to tests with more than three candidates. Similarly, a group of five might break up into a group of three and a dyad, while a group of six into three parallel dyads. Very few sources on group oral tests discuss this threat. Recently, Nakatsuhara (2009, 2011) reported there was no evidence of schisming in her datasets. However, with the possibility of a schism and the progressive increase in the danger of domineering (and withdrawal) at numbers higher than three, a group of three seems a sufficiently "tight" setup that cannot break up and can still be rated reliably, while it is also the lowest number to allow all conversational features to be in play.

Nakatsuhara (2009, 2011) appears to have come to a similar conclusion. In a comparative study of groups of three and four, her conclusions are highly plausible on the basis of the present author's knowledge of the exam. Nakatsuhara concluded that groups of three are altogether more suitable for the testing of interaction because she found more unexplained variance in groups of four, which the present author understands to be less balanced in terms of interaction and extraversion level variables are more influential in groups of four. Further, she also claimed that there was more success involving introverted participants in groups of three than in groups of four and that there was more evidence of avoidance in groups of four than in groups of three. However, the finding that the influence of proficiency level variables was higher in groups of three appears to go against her other findings.

3 Method

As validation is a broad field, the focus of this study had to be narrowed down and some themes (status of speakers, age and gender differences, etc.) that would also be relevant for validation purposes, the "marshaling of evidence" as Bachman (1990, p. 238) puts it, had to be left for discussion elsewhere. The conversational nature of interaction was investigated through a focus on turn-taking, more precisely Sacks et al.'s (1974) three special features, to be pulled together in the analysis of bias sequences as predicted by Sacks et al.

Video recordings were made of the oral tests. They were transcribed and the turn-taking was investigated. All bias sequences were identified in the transcriptions because they could be taken as an indication of the conversational nature of the test interactions. However, in order to properly identify the bias sequences, the utterances that can or cannot count as turns had to be decided first. Deciding what can count as a turn had its problems; therefore, a co-coder colleague was employed to independently recognize (or not) the turns. Their agreement was subsequently statistically tested.

Further validity evidence was the analysis of data from a replication study four years later than the main study. Yet another source of validity evidence was to come forward from a comparison between calibrated abilities of students whose test performances were video recorded and those whose test performance was not. This comparison was made in both the main study and the replication study, which allowed the researcher to evaluate how the video-based findings might generalise to the whole student population taking the exam.

Thus the main research question for this study may be formulated as follows: Can the interaction, as realised by groups of three in Phase 2, be characterized as conversation?

The research question had to be broken down into a number of sub-questions, or pre-questions. Answering them will lead us some way towards answering the main research question.

1. Is there evidence in the data of Sacks et al.'s special features? Is there evidence for bias sequences in the data?
2. Is there a difference between the language ability of videoed and non-videoed students? Can the results be generalized to the entire population (cohort) of students?
3. Can the existence of bias sequences be confirmed statistically?

3.1 Participants and procedures

The participants were students at Eötvös University, who take the same (group) format oral examination at two different stages (levels) of their studies. The first exam is taken at the end of year 1 followed by the second exam typically taken at the end of year 3. The targeted Common European Framework of Reference (CEFR) level of the year 1 exam is B2+, while the third-year exam is targeted at level C1. The examiners were trained and were drawn from Eötvös University staff at the department.

3.1.1 Test administration

The exam roughly takes 25 minutes for each group and the groups are scheduled to take the test at 25-minute intervals. All groups scheduled to take the test at the same time get the same scripted task to do. For reasons of security and economy, the next group in the schedule receives a different task to do. In fact, the order of the tasks is fixed for the examiners.

	Number of all students taking the test	**Videoed**			
		Students	**Groups**	**Tasks**	**Examiners**
Main study	193	44 (23%)	15 (23%)	15 (83%)	6 (60%)
Replication study	185	56 (30%)	19 (31%)	18 (95%)	6 (60%)

Table 1. The student population and the video sample

Limitations of appropriate equipment demanded that the recordings were made only in a selection of exam rooms, typically one exam room per administration. This meant that only a smaller part of the student cohort could be videoed, which in turn meant that only a sample of the students' performances were transcribed and analysed. It was essentially a convenience sample, where "randomness" could be expected from the fact that the division between being videoed or not being videoed was decided at the signup (and time) as places in the schedule filled up. The study was not an experimental study in the sense that students would have been randomly assigned to two equally large groups, one having been videoed, while the other was not. Table 1. is above an overview of the students and tasks/examiners, respectively, providing

an initial impression of how well the results might generalize to the population of all students in the programme.

3.2 Identifying what counts as a turn

The focus on turn-taking meant that it was important to be able to reliably identify what counted as a turn. Below are a few examples put forward to illustrate the difficulties in the identification of turns. Interjections during a speaker's turn could not be recognised as turns unless the current speaker responded to them in some way. Most backchannel signals, for example, "yeah", "right", or "hmm" during the current speaker's turn are typically not turns since the current speaker does not respond to them. Nor are cases of 'echoing', when students repeated what their fellows had previously said. More difficult was a decision to make when the backchannels occurred after the end of a current speaker's turn, especially if there was also a pause and a falling tone reinforced by a projectable end of turn. In Figure 1, for example, when A and B have finished their turns, C only utters an "uhm". C had 1.5 seconds to continue, but apparently did not want to. C's utterance may thus be interpreted as an act of turn-passing and thus a turn in itself. Such acts of turn-passing were recognised as turns. (For the transcription notation, refer to Appendix A1.)

1.		A	Well, //I mean all we have to do is just go into a library and look things up in books and ask someone we ((laughs)) // I think that can be arranged. It's not a problem.
2.		C	// (unclear)
3.		B	// Yeah, that's a good that's a good point anyway.
4.	A turn?	C	Uhm. (1.5)
5.		B	Okay, so sofar we had the picture, live music, what else. Picture, live music er
6.		C	Games

Figure 1. A doubtful turn

Even more confusing may be the single word utterances between clear turns. These are typically "yes", "yeah", "right", which could also be construed as backchannels or as minimal turns to state agreement or simply turn-passing. In Figure 2, A addresses C in utterance 3, to which the response is an indication of thinking. This should count as a turn because C only 'bought time'by first passing the turn back to A.

1.	A	A	Norbi?
2.		B	Can't remember which was it.
3.	Address	A	Steve?
4.	Turn?	C	Mmm.
5.	Q	A	Which should be the last one? Last most important thing. (I mean)
6.		C	People in Budapest are a little bit unkind. They are a little bit different from people who // live in the province

Figure 2. Thinking signal as turn?

3.2.1 Identifying bias sequences

Crucially for this study, the researcher determined two adjacency pairs as the operational minimum of a bias sequence. The reason was that a single adjacency pair (e.g., A-B) might still be part of an A-B-C round. Therefore, at least four adjacent turns between only two of the speakers (e.g., A-B-A-B) were needed for a minimum of a bias sequence. In addition, the beginning of a bias sequence would have to be a clear turn, something more substantial than a simple backchannel or agreement. In Figure 3, recognized turns are numbered (1-5), while those rejected as turns are indicated with dots. The sequence C-B-C-B, bolded, is highlighted with a broken arrow. This would seem to form a minimally adequate bias sequence because unrecognized turns can be ignored. However, the initial utterance by C (no. 2) might also be interpreted as an act of turn-passing or a backchannel signal, therefore, the remaining three cannot be accepted as a bias sequence either.

1.		A	So should it be one ingredient? This not to be late thing,
2.	Begin bias?	**C**	(Possibly,) yeah.
3.		**B**	Well, with a test you can't say that I mean you can't involve into into a test that you have to begin in // time. You shouldn't be late. So I, well, I think one of the most important things is to to have a test which measures the knowledge the student should know or can or might know. // I mean your you shouldn't ask about things or we shouldn't give them tasks which they are not able to do. // I mean of course in accordance to with, with you teaching.
●		A	// Okay, then so
●		A	// Uhum.
●		A	// Uhum.
●		A	[Uhum, yes.
4.		**C**	[Yeah, I think you, you possibly expect from the teacher to tell in advance what you should // except at the exam.
●		A	// Yes.
5.		**B**	Yes.

Figure 3. Unacceptable bias sequence

A clearer example of the minimum of a bias sequence is Figure 4. below. After the examiner's task input, C makes a quick initial contribution, followed by same-time utterances by A and B and clarification. We get to the point where C passes the turn to A. So far the round-the-group pattern is in evidence, which then is followed by A making a point. B responds to this by adding an idea, with A agreeing and B giving a reason. All these last four can be recognised as turns, and together they form a minimal bias sequence between A and B, almost as if it was a conversation within the whole of the group interaction. The turns leading up to the bias sequence could not be construed as a bias sequence, nor can the subsequent turns that follow it be construed in a similar way because C takes over and the pattern following cannot be construed as a bias sequence between any two of the speakers either. It should be noted that many bias sequences were longer than the minimum.

1.		Ex	What we'd like to do is to decide what advice would you give an incoming university student here at ELTE. Okay, you are student at ELTE. And we'd like you to try to, to choose and agree on four, the four most important pieces of advice for a new student. Try to put them in order.
2.		C	The first one should be that he has to be patient.
3.		A	[Yes, patient yes, patient.
4.		B	[Yes. Inevitable.
5.		A	Pardon?
6.		B	It's an inevitable feature.
7.	Turn-pass	C	Yeah.
8.		**A**	And a yes. I think the other most important is that he has to get used to the rush // and standing in lines and uhm he has to be prepared maybe have four or five timetables in hand.
9.		B	// Yeah.
10.		**B**	And a good map.
11.		**A**	Yeah, a good map is (unclear), yes.
12.		**B**	Because travelling here is quite confusing.
13.		C	And also has to be flexible.
14.		A	Yeah, flexible.
15.		B	Yes, especially, yeah
16.	QR	C	What else?
17.		B	Well, yes
18.		A	Uhm, maybe he has to, as you said, patient and patient with I think with, with school with, with studying and with, with uhm other people in his class and teachers because I think the first year is very, very difficult and here we have to get used to each other.
19.		C	And I think he also get, has to get used to that he can only count on himself // and not on others.

Figure 4. The minimum of a bias sequence (excerpt)

A better example of interaction in which a bias sequence is initiated by a more substantial turn is Figure 5 below. Taking away interjected attempts by C, the beginning of the interaction is the round-the-group "pattern", C-B-C-A, in turns 1-4. This interaction is followed by our point of interest (turn 5), at which C develops their argument in a substantial turn that qualifies as the beginning of a bias sequence (5-8). B responds to C, to which C provides the most likely response (after some unrecognized interjections), followed, as most likely, by B again. C's next utterance (9) could be considered as part of the bias sequence, but bias sequences often "peter out" in this way as it is no longer B who responds. Therefore, it is more realistic to consider turns 9-11 to be a return of the round-the-group sequence.

1.	C	Well, I don't think it's possible not to, I mean it's possible not to someone who is from abroad all, uhm, the poor people in the country but a I, I just can't get used to the idea that foreigners come to this country and see all of those people begging the in the streets and, erm, and in the underground. And once I felt so terrible because I met someone from Germany and I don't know after ten minutes we met, he said that he had a horrible day because he came to Budapest and in the first hour he was robbed by, by women, by a woman so | ((laughter)) that that I, I wouldn't like to show that situation, I mean to show all of these poorness and poor people.
2.	B	Yes.
●	A	But I think it's uhm//
3.	C	//I don't think it's possible but
4.	A	It is not possible because it, it goes with, with the country. So // so you can't you can't forget about it because otherwise you show a false picture to // to a foreigner.
●	C	// Yes I think so, unfortunately
●	C	//Well, yes.
5.	**C**	But I really would like to show all these beautiful building what we can see in Budapest or in other towns as well because sometimes we Hungarians don't even notice them. I think erm but but when I moved up I mean when I moved to Budapest, I live in Szolnok, I realised that there are so many nice buildings which actually could be nicer if they would be done up but uhm but they, they look very dirty and old but, but they're really nice. // Actually the architecture.
6.	**B**	// I guess I guess Budapest is basically a really beautiful town and there are many things that's worth showing to a foreigner. Well there are the poor areas whi which I definitely wouldn't show him // like the Rákóczi tér and then that area, but there are beautiful historical monuments like the Hősök tere. For example I would, I would take him to a theatre or a, a cinema if can speak speak Hungarian and a, but not only, not only Budapest but I I would show him not of course not all the coun all the country because tha tha tha that takes a lot of time and there are some areas which are really beautiful // so I would show him some areas with with mountains, the the Bakony and there are the
●	C	// Uhm
●	C	// Yeah.
7.	**C**	or Hortobágy which is // very traditional and special.
8.	**B**	Yeah, yes that's the other and the area of the Balaton of course that is really beautiful. Of course it's full full with tourists full of tourists but // you can get over it probably.
●	C	// And yeah.
9.	**C**	Yes, uhm I wouldn't show the trains in Hungary because I travel a lot by trains and are just disgusting. I I guess because I don't know how dirty they could be I mean I know theyt hey are very dirty and I just don't know why they don't keep it clean. So well the exception is the inter-city trains but, but usually Hungarians travel by the uhm normal, the old ones so
10.	A	They haven't got money for // (unclear)
11.	B	// Yes, cleaning up all those trains costs a lot of money and I I think

Figure 5. Sample of a more substantial bias sequence

3.2.2 Checking reliability

The researcher attempted to increase the reliability of the classification (as turns or not) by going over the data many times since turn definitional problems may affect the reliability of such a study. Thus, the language tester's concern with reliability coincided with a characteristic element in CA research methodology (Levinson, 1983; Seedhouse, 2005): the need to engage

with the data seriously. Even so, a colleague was asked to act as co-coder by watching the recordings and identifying the turns in the transcription of the interaction, without any reference to the researcher's coding of the same. Their agreement was statistically evaluated by calculating Cohen's Kappa.

3.2.3 Testing generalizability

First, calibrated student abilities (results, scores) were taken from the item banks of the year-1 and year-3 exams. The use of abilities was an operational decision made on the assumption that if different (much better) students went to take the test in the video room, this would be indicated in their results being different too.

The abilities were based on the item banks. Each of the two item banks comprised five years' data, that is, the abilities were obtained in the context of hundreds of student performances that included the year of the main study and the year of the replication study. The calibrations were Many-facet Rasch measures (Linacre, 2006) that took into account the raters and the tasks as facets (dimensions) of performance. For each study, the ability values for those videoed and those not videoed, that is, for two groups, were then compared using a t-test for the main study and the replication study, both assuming and not assuming equal variances between the video and non-video groups. The alternative of not assuming equal variances was necessary because the non-video group was naturally much larger.

3.2.4 Checking randomness in the data

The presence of bias sequences in the exams was also tested statistically using the Runs test, also known as the Wald-Wolfowitz test. This test investigates whether there are patterns in the sequence of turns (the data), or there are no patterns, and turns follow each other unsystematically. Bias sequences should form a pattern within the sequence of turns.

4 Results

On the basis of both studies, turn-taking in the small group oral can be described according to the three conversational features formulated by Sacks et al. (1974). Below is a description of why that argument may be made.

4.1 Sacks et al.'s three special features

According to feature #5, turn order varies. In the Eötvös University group oral, round-the-group turn order (ABC, BCA, CAB, etc. patterns) alternates with sequences that involve only two of the speakers (ABAB, BCBC or CACA patterns).

According to feature #9, the distribution of turns is not specified in advance. It should be added that the distribution varies across the exams, tasks and student groups.

Feature #12 was the obvious use of turn allocation techniques, which were also in evidence. Sacks et al.'s (1974) turn allocation techniques were of three kinds:

1. Current speaker selects next: The direct nomination of the next speaker, the most obvious kind of selection, occurs only once in the Eötvös University transcriptions. The low frequency of nomination is probably a reflection of the students interacting with their own familiar peers. A less direct form of the current speaker selecting the next speaker is addressing a question to one of the others, typically maintaining eye contact, which is an implicit form of nomination. This occurs more often in the data, for example, when a repair/ clarification sequence is initiated.
2. Other speakers self-select: This is the dominant form of turn allocation. In most transitions, speakers self-select in the data.
3. Current speaker self-selects: This occurs when the speaker decides to initiate a new turn, when the other speakers have not self-selected and the current speaker decides to continue. As might be predicted, this option occurs comparatively more rarely, for example, only 36 times in the main study data.

4.1.1 Further notable characteristics

Overwhelmingly, one party talks at a time. Overlaps are common, but brief. One will typically yield when two self-selecting speakers begin their turns at the same time or when the next speaker, monitoring the current speaker for an end of a turn, believes (wrongly) their turn has come. Transitions from one turn to the next with no gap and no overlap also hold true for the Eötvös University oral. Turn size varies, ranging from single-word turns to a turn of 195 words in the data. The number of turns varied across the 34 recordings, between 12 and 83 per video.

4.2 Chief evidence: bias sequences

The conversational options are jointly responsible for two typically observable turn-taking patterns, which occurred with every trio, and more than once with some groups in the data. One is when the turn conveniently goes round the trio. Following the "turn order is not fixed but varies" feature of Sacks et al. (1974, p. 701), in this study, speakers' turns vary in order (ABC followed by BCA, CAB and perhaps ABC again). The other pattern is short, "bias" sequences involving only two, followed, sooner or later, by the third student's (re)entry into the conversation. Conversational bias is activated here, as predicted by Sacks et al., which restricts interaction to only two speakers for some time, typically ABAB, BCBC, CACA, and so on, before a round-the-group pattern or another bias sequence begins. The sequences in which conversational bias operates are indicated in Appendix A as vertical arrows with a continuous line.

The evidence for the recurrent bias sequences is the most interesting finding from this study. They occur in each of the 34 videos made in the main study and the replication study, in which 63 and 53 sequences were identified, respectively, bringing their number to 116 overall. Bias sequences range between four and nine turns in each video.

Test Value[a]	[a] 1,00
Total Cases	1847
Number of Runs	588
Z	-10,287
Asymp. Sig. (2-tailed)	,000

a. User-specified.

Table 2. Results from the Runs test

The researcher used their own coding for the Runs (Wald-Wolfowitz) test, which showed that there are 'runs', that is, patterns, in the data and that interaction does not simply go round the group in some random sequence, but it is highly patterned (Table 2). It has showed us that bias sequences are at least one of the verifiable patterns in the data.

4.3 Testing reliability

The reliability of the researcher's classification into accepted turns or utterances not accepted as turns was tested against the classifications by the co-coder. As Table 3. shows there was an acceptable level of agreement between the researcher and the co-coder. As can be seen the co-coder accepted fewer utterances as turns that the researcher. Kappa was calculated as agreement index at 0.779, the 'industrial norm' being ≥0.6 in language testing (Fulcher, 2010, p. 83). In terms of a simple percentage agreement, there were 1662 cases (536+1126) of agreement over 1847 cases overall, which equivalent to agreement in 89.9% of all the utterances.

		Cocoder		**Total**
		0	1	
Researcher	0	536	13	549
	1	172	1126	1298
Total		708	1139	1847

Table 3. The crosstabulation of the researcher's and co-coder' judgements

4.4 Statistical tests for generalizability

The fact that the videos were real (not mock) exam videos lends them a certain measure of credibility. However, it has only been possible to make videos of a proportion of the students. Therefore, it was right to ask whether the video sample can be considered a good cross-section of all the students tested. In Table 4. basic student performance data are presented about videoed and non-videoed students in both the main study and the replication study. What is also clearly shown is that while standard deviations are similar, the group means are close and the standard error of the means are similarly large. All this goes to show that these examinations are rather similar.

	Groups	N	Mean	Std. Deviation	Std. Err. Mean
Main study	Videoed	44	10.5886	10.84310	1.63466
	Non-videoed	149	13.0293	11.63095	.95284
Replication study	Videoed	56	11.9575	8.86631	1.18481
	Non-videoed	129	9.9398	8.59426	.75668

Table 4. Comparison of basic data of videoed and not videoed students

Statistical tests were done to see more precisely whether videoed and non-videoed student samples come from the same population. The Rasch-calibrated oral language proficiency scores, already corrected for task difficulty and rater severity, appeared to be the best readily available data to test for a potential difference between the means in Table 3, the H_0 being that no significant differences existed between videoed and non-videoed students. The results showed that the H_0 could not be rejected in this case (Tables 5-6).

The similarity of the distributions in Table 4 and 5 was confirmed by the Levene's test, indicating small differences between variances for the main study and its replication. This comparability suggested that it was appropriate to use an independent-samples t-test. As the results in Tables 5 and 6 do not allow the H_0 to be rejected, some notion of generalizability may be formed: it will have to be assumed that significantly more (or less) proficient students were *not* examined in the video room, which in turn indicates that what was observable in the video room may have been observable in and representative of all the other testing rooms too.

Equal variances	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
Assumed	.228	.634	-1.241	191	.216
Not assumed			-1.290	74.680	.201

Table 5. Main study comparison of test proficiency in the video and non-video groups

Equal variances	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
Assumed	.000	.985	1.453	183	.148
Not assumed			1.453	101.743	.154

Table 6. Replication study comparison of test proficiency in the video and non-video groups

5 Discussion

The operation of conversational bias, it is believed, underpins the claim of conversational language and a high level of dynamism in a group with three candidates.

5.1 Additional speakers compete

Sacks et al. (1974) also say that the "current speaker selects previous speaker as next bias remains invariant over any increase in the number of parties" (p. 712). As conversation organises only two speakers at a time, additional speakers "will be under constraint to self-select" (p. 712) for a turn. Of course, being "under constraint" might conjure up images of competing in aggressive or domineering ways, but while a speaker's challenge is to secure their own share in the conversation, the challenge for their partners is to sensitively allow this partner into it. This interactional competence, the ability to turn-take in groups, may be posited as an important facet of oral communicative competence, although not at any level.

It cannot be mere accident that group orals have been used predominantly in higher education. That students in higher education must represent a rather higher level in the foreign/second language is only an assumption, however; it is nearly impossible to verify from the literature. It may be more useful, therefore, to formulate a hypothesis in terms of the Common European Framework (CEFR, 2001) scales here. On the basis of research so far, it is reasonable to expect that the group oral can realise its full potential from level B2 upwards. This is the range of ability where interactants should have the skills to realise the potential in this test format. Below level B2, the need to plan and process their message in a linguistically and pragmatically acceptable way may prevent participating in a fluid encounter with more than one other speaker. For these lower level students interview tests with skilled interviewers or even paired-orals might be more suitable, where it is always more obvious whose turn is next. That obviousness offers more security.

The viability of the group idea might also be limited to special contexts where candidates are rather similar, as in higher education. The explanation is most likely that in university settings candidates are often at a similar level of proficiency, are all adults, if not necessarily at a similar age, have the same status as students and many know each other (familiarity). The Eötvös University group oral operates in just such a context, which is very different from a public language test. The specificities of this context call for further research into the group oral in different educational contexts.

5.2 Potential for group tests with three speakers

On the basis of work by Sacks et al., the merit of the group oral with three students may be that schisming is not a possibility. Not surprisingly perhaps, there was no evidence of schisming in this study. It appears that Sacks et al. provide a strong rationale for the group oral with three candidates. Due to the number of participants and the action of conversational bias, turn-taking remains unstable enough to ensure a high degree of dynamism and prevent both the invariance of turn order in paired interactions (individual and paired orals) and the potential break-up of interaction into parallel dyads and/or groups with a larger number of candidates.

Conversational bias should not be seen as a limitation of the group oral; what it creates instead is a situation in which interactants in the third party position need to display the relevant skills, that is, interactional competence, to enter the conversation in order to take the floor, or as the dyad holding the floor, to sensitively allow the third party back in. It is suggested here that conversational bias might be the reason for dynamism in a group oral with three participants.

Of course, the turn-taking system is but one of the organizational systems in Conversation Analysis (Lazaraton, 2002). The Eötvös University group oral might still prove to be something else after an investigation of the system of repair, of the preference organization, of openings, pre-closings and closings and of topic organization, which this study has not specifically dealt with.

There is another threat, from score-based studies, to the validity argument sketched here. Even if the quality of the rating is high in itself, the score variance generated by one speaker might be found to interfere with another speaker's scores. A fundamental conflict might emerge in this way, between the requirement of making test results individual on the one hand, from which oral proficiency is inferred, and the co-constructed nature of test performance on the other. In this way, what is to be construct-relevant variance from one speaker's interactional competence may be confounded with the construct-irrelevant variance generated by another speaker.

6 Conclusion

In terms of its turn-taking, the language used in the context of the test was shown to be related to conversation as an external validity criterion, as described by Sacks et al. (1974). Their three special features are evidenced in the data, too. The data was collected under testing circumstances, which might have resulted in a simplified and restricted institutional type of talk, given that a test at a university is an institution itself. The simplification, however, has not affected key features of conversation, strengthening in this way the validity argument for the group oral with three speakers.

Proofread for the use of English by: Frank Prescott, Department of English, Eötvös Loránd University, Budapest.

Acknowledgements

The author acknowledges the contribution of Angi Malderez, original co-designer of the group test, and the author's former students, Gabriella Bálint, Gertrud Murnyák and Ágnes Heckmann, who were instrumental in transcribing the videos and finally, Éva Mák, who lent her support as co-coder.

References

Atkinson, J. M., & Heritage, J. (Eds.). (1984). *Structures of Social Interaction: Studies in Conversation Analysis.* Cambridge: Cambridge University Press.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

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. https://doi.org/10.1191/0265532203lt245oa

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a Better Performance. *Language testing,* 26(3), 341-366. https://doi.org/10.1177/0265532209104666

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency. *Language testing,* 20(1), 1-25. https://doi.org/10.1191/0265532203lt242oa

Berkoff, N.A. (1985). Testing oral proficiency: A new approach. In Y. P. Lee, A. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 93-99). London: Pergamon Press.

Canale, M. (1984). Testing in a communicative approach. In G. A. Jarvis (Ed.), *The Challenge for Excellence in Foreign Language Education* (pp. 79-92). Middlebury, VT: The Northeast Conference Organisation.

Carroll, J. B. (1980). *Testing Communicative Performance: An Interim Study.* Oxford: Pergamon Press.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Day, E. M., & Shapson, S. M. (1987). Assessment of oral communicative skills in early French immersion programmes. *Journal of Multilingual and Multicultural Development.* 8(3), 237-260. https://doi.org/10.1080/01434632.1987.9994288

Drew, P., & Heritage, J. (Eds.). (1992). *Talk at Work: Interaction in Institutional Settings. Studies in interactional sociolinguistics.* Cambridge: Cambridge University Press.

Folland, D., & Robertson, D. (1976). Towards objectivity in group oral testing. *English Language Teaching Journal, 30*(2), 156-167. https://doi.org/10.1093/elt/xxx.2.156

Fowler, R., Hodge, B., Kress, G., & Trew, T. (1979). *Language and Control.* London: Routledge and Keegan Paul.

Fulcher, G. (1993). The construction and validation of rating scales for oral tests in English as a foreign language (PhD thesis). University of Lancaster, England.

Fulcher, G. (1996). Testing Tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51. https://doi.org/10.1177/026553229601300103

Fulcher, G. (2010). *Practical Language Testing*. London: Hodder education.

Gan, Z. (2010). Interaction in group oral assessment: a case study of higher- and lower-scoring students. *Language Testing,* 27(3), 585-602. https://doi.org/10.1177/0265532210364049

Gardner, R. (2006). Conversation Analysis. In A. Davies & C. Elder (Eds.), *The Handbook of Applied Linguistics* (pp. 262-284). Malden, MA: Blackwell Publishing.

Gorman, T., & Brooks, G. (1986). Assessing oracy. In M. Portal (Ed.), *Innovations in Language Testing* (pp. 106-140). Windsor: NFER-Nelson.

He, L., & Dai, Y. (2006). A corpus-based investigation into the validity of the CET-SET group discussion. *Language Testing,* 23(3), 370-401. https://doi.org/10.1191/0265532206lt333oa

Hilsdon, J. (1991). The group oral exam. Advantages and limitations. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s: The Communicative Legacy* (pp. 189-197). London: Macmillan.

Hong Kong Examinations Authority. (1995). *English language Public Examinations in Hong Kong – A Guided Tour*. Hong Kong: Mimeo.

Hutchinson, C. (1986). The classroom assessment of English language skill. *Teaching English, 19*(3), 30-36.

Johnson, M. (2001). *The art of non-conversation. The Re-examination of the validity of the oral proficiency interview.* New Haven and London: Yale University Press.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (pp. 17–64). Westport, CT: American Council on Education and Praeger.

Lazaraton, A. (1996). Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing,* 13(2), 151-172. https://doi.org/10.1177/026553229601300202

Lazaraton, A. (2002). *A Qualitative Approach to the Validation of Oral Language Tests.* Studies in Language Testing 14. Cambridge: Cambridge University Press.

Leaper, D. A. (2010). Group Discussion Tests: Investigating the Construct of Interaction. Paper presented at the 36th International Association for Educational Assessment (IAEA) Annual Conference. Retrieved from http://www.iaea2010.com/fullpaper/209.pdf.

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

Linacre, J. M. (2014). Facets: Rasch Measurement Computer Program. Version 3.78 [Computer software] Chicago: Mesa Press.

Liski, E. P., & Puntanen, S. (1983). A study of the statistical foundations of group conversation tests in spoken English. *Language Learning, 33*(2), 225-246. https://doi.org/10.1111/j.1467-1770.1983.tb00536.x

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language testing*, 26(3), 397-421. https://doi.org/10.1177/0265532209104668

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: American Council on Education/Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist,* 50(9), 741-749.

Morrison, D. M., & Lee, N. (1985). Simulating an academic tutorial: a test validation study. In Y. Lee, A. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 85-92). Oxford: Pergamon Press.

McNamara, T. (1997). "Interaction" in second language performance assessment: Whose performance? *Applied Linguistics, 18*(4), 446-466. https://doi.org/10.1093/applin/18.4.446

Nakamura, Y. (2003). Oral proficiency assessment: Dialogue test and multilogue test. In *Proceedings of the 2nd Annual JALT Pan-SIG Conference May 10-11*(pp. 52-61). Held at the Kyoto Institute of Technology, Kyoto, Japan. Retrieved from http://jalt.org/pansig/2003/HTML/Nakamura.htm

Nakamura, Y. (2005). The construct of speaking for communicative testing. *Journal of Foreign Language Education, 2,* 61-74. Retrieved from http://www.flang.keio.ac.jp/webfile/kiyo/kiyo_second.pdf.

Nakatsuhara, F. (2009). The effects of the number of participants on group oral test performance. Presentation at the Language Testing Forum, at the University of Bedfordshire, November 20-22. Retrieved from http://www.beds.ac.uk/research/bmri/crella/LTF2009/Nakatsuhara.pdf

Nakatsuhara, F. (2011). Effects of test-taker characteristics and the number of participants in group oral tests. *Language testing*, *28*(4), 483-508. https://doi.org/10.1177/0265532211398110

Nakatsuhara, F. (2013). *The Co-construction of Conversation in Group Oral Tests.* Frankfurt am Main: Peter Lang.

Nunn, R. (2000). Designing rating scales for small-group interaction. *English Language Teaching Journal,* 54(2), 169-178. https://doi.org/10.1093/elt/54.2.169

Ochs, E. (1979). Planned and unplanned discourse. In T. Givón (Ed.), *Discourse and Syntax* (pp. 51-80). New York: Academic Press.

Ockey, G. J. (2001). Is the oral interview superior to the group oral? *Working Papers on Language Acquisition and Education*, International University of Japan, *11*, 22–41. Retrieved from http://www.iuj.jp/language/workingpapers/pdf/LP-11-2.pdf

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* (Doctoral dissertation). UCLA, Los Angeles.

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

Pavlou, P. (1997). Do different speech interactions in an oral proficiency test yield different kinds of language? In A. Huhta, V. Kohonen, L. Kurki-Suonio, & S. Luoma (Eds.), *Current Developments and Alternatives in Language Assessment*. *Proceedings of LTRC 96* (pp. 185-201). Jyväskylä: University of Jyväskylä and University of Tampere.

Reves, T. (1980). The group-oral test: an experiment. *English Teacher's Journal*, *24*, 19-21.

Reves, T. (1982). The group oral examination: a field experiment. *World Language English, 1*(4), 259-262.

Reves, T. (1991). From testing research to educational policy: a comprehensive test of oral proficiency. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s: The Communicative Legacy* (pp.178-188). London: Macmillan.

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

Scott, M. L. (1986). Student affective reactions to oral language tests. *Language Testing, 3*(1), 99-118. https://doi.org/10.1177/026553228600300105

Seedhouse, P. (2005). Conversation Analysis and language learning. *Language Teaching,* 38(4), 165-187. https://doi.org/10.1017/s0261444805003010

Shectman, Z. (1988). Selecting candidates for teachers training college: a group assessment procedure as an alternative to individual interviews. *European Journal of Teacher Education,* 11(2-3), 185-193. https://doi.org/10.1080/0261976880110213

Shohamy, E., Reves, T., & Bejarano Y. (1986). Introducing a new comprehensive test of oral proficiency. *English Language Teaching Journal,* 40(3), 212-220. https://doi.org/10.1093/elt/40.3.212

Sinclair, J. Mc. H., & Coulthard, M. (1975). *Towards an Analysis of Discourse: The English Used by Teachers and Pupils.* Oxford: Oxford University Press.

Skehan, P. (1991). Progress in language testing: the 1990s'. In J. C. Alderson & B. North (Eds.), *Language Testing in the 1990s: The Communicative Legacy* (pp. 3-21). London: Macmillan.

Swain, M. (1985). Large-scale communicative language testing: a case study. In Y. Lee, A. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35-46). Oxford: Pergamon Press.

Tsui, A. B. M. (1994). *English Conversation.* Oxford: Oxford University Press.

Van Moere, A. (2006). Validity evidence in a university group oral test. *Language Testing,* 23(4), 411-440. https://doi.org/10.1191/0265532206lt336oa

Van Moere, A. (2007). *Group oral tests: How does task affect candidate performance and test scores?* (Doctoral Dissertation). Lancaster University, England.

Wigglesworth, G., May, L., Galaczi, E., Nakatsuhara, F., & Van Moere, A. (2010). Exploring Interactional Competence in Paired and Group Speaking Tests. Symposium at the 32nd Language Testing Research Colloquium, 12-16 April, 2010, Cambridge, UK.

APPENDIX A: Transcript of a sample phase 2 video

APPENDIX A1 Transcription notation

Text // text	(Double oblique) Indicates the point in the current speaker's turn at which the next speaker's talk begins.
//text texttext	If it is located at the beginning of a turn, it marks a turn by the next speaker while the previous speaker was talking. Thus, it may indicate the beginning of an overlap. These are typed in the order in which they occur in the current speaker's turn.
(unclear)	This appears if the message is incomprehensible.
(you've got)	Message which is only guessed because it is not well heard.
(( ))	Noteworthy comments about how the turn was made or about non-verbal features such as laughter.
●	Dots appear in the left hand column to indicate utterances that cannot be recognised as turns.
A,B,C	Student codes which represent different persons in each task, but A was always the student on the left, B in the middle and C on the right. Bolded letters indicate bias sequences.
Ex (1 and 2)	Examiner (first and second)
|	Observed option 3
.	Falling tone
,	Moderately rising tone
?	Strongly rising tone
A	Address, nomination
[text text	Obvious overlap with previous speaker(s)

APPENDIX A2: Sample task

●	Ex	What we'd like you to totototo agree on is three things three things that you would rather not show a foreign friend visiting Author country. And three things that you would like to show. You've got to agree on three things that you would, three things that you wouldn't.
1.	C	Well, I don't think it's possible not to I mean it's possible not to someone who is from abroad all uhm the poor people in the country but a I I just can't get used to the idea that foreigners come to this country and see all of those people begging the in the streets and erm and in the underground. And once I felt so terrible because I met someone from Germany and I don't know after ten minutes we met, he said that he had a horrible day because he came to Author city and in the first hour he was robbed by by women by a woman so | ((laughter)) that that I I wouldn't like to show that situation I mean to show all of these poorness and poor people.
2.	B	Yes.
●	A	But I think it's uhm//
3.	C	//I don't think it's possible but
4.	A	It is not possible because it it goes with with the country. so // so you can't you can't forget about it because otherwise you show a false picture to // to a foreigner.
●	C	// Yes I think so, unfortunately
●	C	//Well, yes.
5.	**C**	But I really would like to show all these beautiful building what we can see in Author city or in other towns as well because sometimes we Hungarians don't even notice them. I think erm but but when I moved up I mean when I moved to Author city, I live in Szolnok, I realised that there are so many nice buildings which actually could be nicer if they would be done up but uhm but they they look very dirty and old but but they're really nice. // Actually the architecture
6.	**B**	// I guess I guess Bu Author city is basically a really beautiful town and there are many things that's worth showing to a foreigner. Well there are the poor areas whi which I definitely wouldn't show him // like the Rákóczitér and then that area, but there are beautiful historical monuments like the Hősöktere. For example I would I would take him to a theatre or a a cinema if can speak speak Hungarian and a but not only not only Author city but I I would show him not of course not all the coun all the country because thathatha that takes a lot of time and there are some areas which are really beautiful // so I would show him some areas with with mountains, the theBakony and there are the
●	C	// Uhm
●	C	// Yeah.
7.	**C**	or Hortobágy which is // very traditional and special.
8.	**B**	Yeah, yes that's the other and the area of the Balaton of course that is really beautiful. Of course it's full full with tourists full of tourists but // you can get over it probably.
●	C	// And yeah.
9.	**C**	Yes, uhm I wouldn't show the trains in Author country because I travel a lot by trains and are just disgusting. I I guess because I don't know how dirty they could be I mean I know they they are very dirty and I just don't know why they don't keep it clean. So well the exception is the inter-city trains but but usually Hungarians travel by the uhm normal, the old ones so

10. A They haven't got money for // (unclear)
11. B // Yes, cleaning up all those trains costs a lot of money and I I think
12. C Oh yeah.
13. A And one thing that I wouldn't show I think it's Balaton because it's not so Hungarian so it's
14. C Yeah, there are all of those // German signs
15. A // It it has changed a lot but what I would show them definitely was the most beautiful thing in Author city and this is the castle. I like it. | // very much and maybe I would take him to concert to Mátyás // templom.
- B // Yes.
- C // Uhm
16. **B** Well, there are many beautiful castles in Author country. A few years ago I saw the castle at Keszthely // it it was extremely beautiful
17. **C** Yes, that that's the most beautiful I think // that's my personal idea.
18. **B** The that's I would definitely show.
19. **C** I would too. So we would show the nice buildings erm // the nice areas
20. **A** // the mountains
21. **C** Yes, the mountains and I would show a very unique animal, which is the well I don't know the name of it in in English. It's the grey bull I guess because that's unique to Hungarians there is no animal like that in other countries.
22. **A** What kind of animal is it?
23. **C** Well, // it's like a erm a huge // bull with with a a
- B // Oh, yes yes
- B // it's like
24. **B** Or an ox maybe
25. **C** Well, yeah ((laughter))
26. **B** It's in the Hortobágy a few // a few one of them
- A // Uhm.
27. **C** And it's grey.
28. A Are you familiar with?
29. B Well, not really we we have learned it in in the biology lessons at high school so ((laughter)) This is the only I know about them.
30. C Uhm they are found in in Hortobágy and I would like national clothes, the folklore clothes and the I think foreigners would enjoy the folk dance of Hungarians. I think so. So I I would show that one too and I would take him to to a cage uhm there you can swim so it's like a a cage where where there is a bath in it. Do you know what I mean?
31. A Yes, I know.
32. C It's a it's found in Tapolca I think so.
33. B I haven't seen anything like that before I think.
34. C Well, you know a cage with with waters in it, and a cage which has uhm these crystals like for example there is one in uhm
- Ex (unclear)
35. C Yes.
36. B Oh yes, all right. Yes, yes, all right. I got it, yes. Well it depends in on the season, I think, so in the summer or in the autumn it would worth taking him or her to a forest maybe even in in Buda, or or in the countryside because the forests are really beautiful in those // periods of time. There are a lot of nice and beautiful places in Author country, I think.

●	C	// Uhm.
37.	C	So what is the third thing third thing we we wouldn't show
●	A	Er
38.	B	Maybe the zoo I think, it it's I think it's just a torture of the animals // to keep to keep them locked up in cages. // Well, yes, yes .And those those safari style styles those are much better, but there are there's no place for for them in Author country I guess. So it's hard to keep the lions on their plain.
●	C	// Well yes
39.	A	You are right but people got used to to going to zoo. They often forget about // about this (unclear)
40.	B	// yes but I think the Hungarian zoo is not not one of the best in in the // world wild
41.	C	// I agree with you
42.	A	I don't remember because I was in the zoo when I was a child and I don't remember but I want to go to the zoo, than I will see.
43.	B	Well, I've been there several times and and it's interesting because there are a hundreds of animals and there are really exotic ones
44.	C	Well, yeah, but if you think about it, actually it's pretty bad that they they are locked up in very small areas.
45.	B	Yes.
46.	A	But this the a case in all zoos.
47.	C	Well actually not really because there are open zoos or whatever they are called, where where they don't have cages. So (2 sec)
48.	A	What we wouldn't show is what you said and I agree with you with the train, trains and the buses as well.
49.	B	Yes, and the mass traffic.
50.	C	And the rude the people. I wouldn't want him or her to meet those rude people I I meet every day. so
51.	A	Maybe the supporters of Ferencváros (laughter)
52.	B	I think there are many of them.
●	Ex	Ok, thank you very much.