Testing tasks: issues in task design and the group oral

Glenn Fulcher *University of Surrey*

This article investigates some of the issues which surround the use of tasks in oral tests, with particular reference to the group discussion. This is done from the perspective of a group of students who were asked to attempt three oral tasks. Questionnaire techniques and retrospective reports were used to collect data from the students. The principle is that test-takers have a great deal to offer to the test researcher in making judgements about the value of the tests which they take (Brown, 1993).

The issues surrounding task design and use are complex, and are currently being debated not only in language-testing circles but also in the fields of second language acquisition and discourse analysis. For this reason, this article will refer to discussions in all three areas to shed light on the selection of tasks for use in oral tests. Information from the statistical analysis of tests will also be presented. All views about tests and tasks used in tests, however much some authors might eschew theory or statistical analysis (Underhill, 1987), spring from inherent theoretical positions. These positions make predictions about test scores under particular conditions, and the results of analysis enable the researcher to assess whether a view can be supported by empirical evidence.

Finally, the article will look at what is possibly one of the most problematic questions in proficiency testing: the generalizability of a test score given on one task to another task or tasks. It is arguably the case that, if this is not possible, there is no justification for proficiency testing.

I The group oral test in use and research

Folland and Robertson (1976) were among the first writers to recommend using the group discussion in oral testing, a task type which had previously been regarded with some suspicion (Wilkinson, 1968: 119–20). Since then there have been reports of group testing being used successfully in Israel (Reves, 1980; Shohamy, Reves and Bejarano, 1986; Reves, 1991) and Zambia (Hilsdon, 1991) with school students, and in Italy with university students (Lombardo, 1984). Berkoff (1985: 95) argued that using groups of students overcomes the problems of 'artificial conversation' between a 'distant examiner' and a 'nervous examinee'. Morrison and Lee (1985) also report successful uses of the group discussion with university students in Hong Kong in simulating academic tutorials.

The reports of successful use, the claim of a reduction in 'test-type' language and reduced anxiety in the literature are not well supported with empirical evidence, although it is clear that the studies conducted in Israel were well designed and the results reliable. It appears to be unfortunate that the group discussion task was not retained in the Israeli test battery because of logistical objections from school principals (Reves, 1991: 182), for its retention would surely have generated a great deal of valuable data. Also, it appears to be generally accepted that by using a variety of task types in oral testing a wider range of language may be elicited. Thus, Upshur (1971: 47), Van Weeren (1981: 57), Shohamy, Reves and Bejarano (1986) and Shohamy (1983; 1988; 1990a) all make the case for multiple-task oral tests. If the group oral discussion task type elicits assessable language from students, and if the task type or the language is, or is perceived by the student to be, qualitatively different from other task types, then an excellent case could be made for its inclusion in oral test batteries.

II Sources of data for this study

A total of 47 students attempted three tasks. The students were all Greek-speaking students registered in EFL programmes in Cyprus. The students were all preparing for entry to tertiary education in English-medium establishments. The average age of the sample was 15 years 7 months, and the majority were expected to apply for tertiary posts by the age of 17. There were 21 female students and 26 male students in the sample.

Of the three tasks used, two were one-to-one interviews and one a group discussion. Each task was completed by each student in a space of no more than one-and-a-half weeks. Students left their classes at an allocated time to take the tasks in a room which had been specially prepared for the purpose.

The students were videotaped attempting each task, and rated by five raters on three rating scales. This quantitative analysis presented in this study is based upon the scores awarded on the six-band fluency rating scale, of which three sample bands are provided in Appendix 4.

Task 1 was designed to be similar to a First Certificate in English oral interview, which primarily requires the description of a pictorial prompt followed by a discussion on a related topic. Task 2 was designed to be similar to an English Language Testing Service (ELTS, now superseded by the International Language Testing Service or IELTS) interview, in which students are asked to discuss a previously encountered text. The text was given to each of the

students two days before the interview, and they were allowed 30 minutes to read it, and answer a small number of multiple-choice questions. These questions were not marked, and no feedback was given. Their purpose was merely to help the students read for gist. The text was the same for all students in the sample, and the topic was 'Poverty in the third world'. In the third task, groups of students were allowed 10 minutes to prepare a discussion on the topic of education in their country. Each student was provided with a task card which gave a number of ideas for content of the discussion, but no other guidance devices or controls were used. After the 10 minutes they were invited to discuss the subject with each other for a further 15 minutes. Each of the tasks could be characterized as 'unfocused' (Nobuyoshi and Ellis, 1993: 204) and, because meaning rather than form is the focus, as 'communicative' in the weakest sense of the word (Nunan, 1989: 10).

After each of the tasks, all students were invited to complete a questionnaire on their test-taking experience (Appendix 1), and after all three tasks had been completed they were asked to compare the tasks and state which they would prefer to do again if given the choice. Some 45 completed questionnaires were returned. One-third of the students took task 1 first, one-third task 2 first and one-third task 3 first. As the sequence in which the tasks were taken was different, any potential order effect was controlled for.

A group of eight students were asked to attend 'debriefing' interviews in which they were shown the videorecording of one of the tasks, and asked to retrospect on their experiences. A similar procedure was carried out with the raters after they had seen all the videotapes, although this is not reported in this article.

Quantifiable responses to the questionnaire were analysed using iterative principal axis factor analysis in an attempt to identify factors influencing student responses. Scores on each of the tasks were analysed in a G-study, and through the use of a Rasch partial credit model (Linacre and Wright, 1990 Linacre, 1991).

Information collected from both the questionnaires and the debriefing interviews will be referred to under each of the headings below, and statistical data introduced where appropriate. We will look first at the issue of language artificiality in tests, and the student and rater perception of the 'naturalness' of the discourse. Secondly, we will look at task difficulty, an area which has traditionally been very difficult to study and assess. Next we will deal with task type and test-related anxiety in an attempt to discover whether or not certain tasks cause more stress than others. We will then move on to student perceptions of the 'validity' of tasks, their enjoyment of the testing experience, students' preferred tasks, raters' assessments of

task types and the generalizability of test scores obtained on one task to another task.

III Artificiality in oral tests

It has often been claimed that certain oral tests are valid on the grounds of the task type selected. This position is essentially a claim for face validity, defined as '... the degree to which students feel they are performing a real communicative act' (Bartz, 1979). Thus, Wilds (1979: 12) argued that the validity of the Foreign Service Institute (FSI) test was 'unquestionable' because the oral interview was based upon a demonstration of speaking ability in a 'natural context' related to living and working abroad. Claims that test tasks replicate natural contexts and real-life situations which encourage natural language use remain the cornerstone of the claim to validity in oral test design. In this respect, there appears to be little difference between many British and American authors in the field. Proponents of 'communicative' language testing in Britain make the notion of 'real-life tasks' and 'natural language' the touchstone of validity (Morrow, 1982: 56–57; Underhill, 1987). Similarly, Lowe (1983: 235; 1987: 46) claims that the interlanguage round table (ILR) test is valid because it requires the examinee to use the spoken language in natural contexts; Lowe and Liskin-Gasparro (1986: 4) argue that the oral proficiency interview (OPI) is highly face valid because it involves 'real' conversation.

It need not be repeated that the appeal to face validity is neither a necessary nor sufficient condition for the validity of a test (Stevenson, 1985a; 1985b), but the issue of whether or not the task design used in a test is capable of producing a context for 'natural language output' is one which is worthy of investigation. That is, it should be taken out of the arena of face validity and studied in its own right in relationship to reducing potentially confounding affective factors in interpreting oral test scores (Shohamy, Reves and Bejarano, 1986: 213), and in relation to maximizing the amount and quality of assessable data which a task may generate.

Much of the work which has been done on 'interview talk' suggests that the one-to-one oral interview generates a special genre of language different from normal conversational speech (Silverman, 1976; MacPhail, 1985; van Lier, 1989; Perrett, 1990; Lazaraton, 1992). Silverman's work remains the classic statement of the type of language which constitutes an interview, and this deserves quoting at some length. He argued that an 'interview' could be seen as different from a 'chat', 'discussion' or 'seminar' in that interview talk consists of

> (i) a series of questions and answers, in which (ii) answers are taken to stand for underlying patterns relevant to future decisions rather than to present talk . . . while questions will be read as seeking to elicit what 'lies behind' the talk of the respondent in order to settle practical outcomes . . .

He further argues that

> (iii) interview talk is known to be on-the-record; ie the comments of one party will be read as a display of qualities and this reading will be reported upon to other persons with a legitimate 'right to know' and will eventually produce certain future decisions . . . (iv) questions are provided by one person . . . and the talk of some other person is seen as answers-to-questions, and where (v) one person is alone legitimately responsible for the doing of the beginning and the doing of the ending of the interaction, for ending one existing topic and introducing a new topic and for formulating the talk (ie commenting on the talk's context or on the character of what is being said, will be said or should be said). (vi) While, as in all talk, judgements about meanings are made partly on the basis of the sequencing of utterances, in an interview this sequencing is attended to as routinely a managed product of one talker (Silverman, 1976: 141–44; 147).

Perrett (1990) employed Hasan's model of generic structure potential (Halliday and Hasan, 1985: 56) to describe the generic structure of the interview, and characterized the interview genre as one of 'information exchange' in which other text types, including the social uses of language, were not being tested (Halliday and Hasan, 1985: 231). Using the Hallidayan categories of field, mode and tenor to characterize further the interview, Perrett claims on the basis of the analysis of six interviews that the field is constituted by the overt (but secondary) purpose of eliciting factual information, and the covert (but primary) purpose of one participant displaying ability. The mode deals with the part the language plays within discourse, and this is clearly a display of linguistic ability. The tenor of discourse concerns the description of the participants and their relationships, and in this context the power of the interviewer is extremely great, so that '. . . the tenor variables are so strong in these interviews that they ultimately override the other variables of field and mode' (Halliday and Hasan, 1985: 235). It must be noted that in these studies which deal with tenor of the discourse, the interviewer is invariably a native speaker. This is an additional facet of the situation which should be taken into account.

Lazaraton (1992) used conversational analysis to examine the transcripts of 20 oral interviews to test the claim that it represents 'natural conversation'. She concluded that it was responsibility for initiating sequences of talk, and the form of the initiations, which gave the oral interview its special characteristics as a test genre. This research adds to the findings in recent second language acquisition (SLA) studies.

In the field of SLA a number of studies which have looked at native speaker/non-native speaker discourse have suggested that native speakers tend to dominate the discourse (Harder, 1980; Scarcella, 1983) and that non-native speakers tend to make more responding moves than opening moves, and do not introduce new topics (Hatch, 1992). It may also be the case, although this has not been empirically investigated, that oral interviewers, being teachers themselves, use features of teacher talk in the discourse of the interview such as checks, clarification requests, repetitions or expansions of the students' previous utterance (Long, 1983; Ellis, 1992: 34–35).

It would seem from current research that the position of power in which the interviewer is placed within the one-to-one oral interview is so great that the imbalance between interviewer and interviewee control over talk will inevitably lead to the production of unnatural or 'test-type' discourse on the part of the student (Zuengler, 1993).

There are, to my knowledge, no similar studies of the discourse of group oral tests to date, although preliminary descriptions of the discourse of various task types exist in Shohamy, Reves and Bejarano (1986) and Shohamy (1988), while Shohamy (1990b) provides a general review of the use of discourse analysis in testing.

However, responses from students both in the questionnaire and in debriefing interviews did provide a clear indication that certain tasks *were perceived* to encourage more natural conversation than others. From the questionnaire, it was question 6 which elicited the most comment on naturalness. Almost half the students responded that engaging in group discussion with partners gave them more confidence to speak and to say what they wanted, rather than having to respond to an examiner. One student commented that this was much more 'natural' than talking to an examiner in a one-to-one interview. In the debriefing interview there was a strong indication that students who were more anxious about taking oral tests thought that the group oral interview allowed more natural conversation to emerge. One student, who said she believed her marks would be lower than she deserved because she was 'naturally shy', argued clearly that interview talk is artificial, and that being able to discuss a topic of interest with a number of friends relieved her of the stress which she normally felt in these situations. The student was quite adamant that if tests contain one-to-one tasks, interviewers should have some knowledge of psychology to be able to identify students like her so that they would not be penalized for character traits. The following is an extract from the debriefing interview:

Rater: Do you think that [a quiet voice] would influence an examiner?
Student: Yes, definitely, because if I don't have fluency that means I won't get any marks.
Rater: So just for an interview you think that you have to alter your character?
Student: Yes. I try to but I don't think I can do anything different, because I don't feel it's natural. When I talk with others I feel it's more natural than in the interview. I just don't like it.

The students interviewed in the debriefing demonstrated their awareness of a particular aspect of 'interview talk', namely, the long gaps between student and interviewer utterances. A cursory glance at transcripts of interviews shows that the length of pauses between turns is larger than would be tolerated in normal conversation and, although it has not yet been empirically investigated, the perception of the students is that the length of these pauses decreases when a group discussion task is used.

A number of students also pointed out, perhaps obviously, that if the interviewer in the one-to-one situation wrote anything down during the interview, this was most disconcerting, while during the group discussion this was not perceived to be a problem. Such inhibitions have been noticed by other researchers (Rodriguez, 1984), and there may be a case to be made for a second marker separate from the interlocutor being used in oral interviews, as is the case with the University of Cambridge Certificate of Advanced English, and many American OPI tests.

It remains to be seen whether these perceptions are borne out by a discourse analysis of the tasks themselves, but it is clear that half of the students who took part in this study very clearly expressed the view that, for them, a group oral task is a much more natural situation in which to engage in conversation than in a one-to-one oral interview.

IV Task difficulty

Very little is known about the difficulty, or the comparative difficulty, of the various task types which we use in testing or, indeed, within the classroom. Kenyon and Stansfield (1991) recommended field testing of tasks and the use of questionnaire data from students and raters to identify good and poor tasks. Stansfield and Kenyon (1992) attempted to scale a number of speaking tasks – or rather what appear to be functions within speaking tasks – described in the American Council on the Teaching of Foreign Languages (ACTFL) guidelines (1986) according to difficulty. Using a Rasch partial credit model, they asked groups of Spanish, French and bilingual education teachers to assess the difficulty of a number of

Table 1 Task difficulty on a fluency rating scale

	Task		
	1	2	3
Difficulty	−0.36	0.45	−0.90
Standard error	0.10	0.10	0.10
Mean square outfit	0.90	1.20	1.10

tasks in a study for the Development of the Texas Oral Proficiency Test (TOPT). Although Stansfield and Kenyon discovered a reasonable alignment between the suggested difficulty level in the ACTFL guidelines and the assessment of difficulty by teachers, this is still a long way from field testing tasks on students and assessing task difficulty from test scores. It should be remembered that expert judgement in assessing item difficulty in more traditional tests is no substitute for the collection of empirical data from pretesting.

In Table 1 the task difficulty of the three tasks used in this study is measured in logits, using a Rasch partial credit model for estimation directly from student scores. The lower the logit, the easier it is to score more highly on that task. It can be seen that task 3 is the easiest of this trio, and task 2 is the most difficult.

As Nunan (1989: 99) has commented, one of the problems with all literature on task design, whether this be in the field of testing, SLA or classroom methodology, is that the contribution of relative design factors to success or failure is largely unknown. This is true of task difficulty, where there is little empirical evidence which suggests that certain tasks are more or less suitable for learners of particular levels of ability. If the results presented here can be replicated, we may be able to hypothesize that naturalness in discourse outcomes could be one factor which helps to make a task easier. This would be worthy of further investigation.

V Task type and test-related anxiety

Galassi Frierson and Siegel (1984) and Madsen and Murray (1984) have shown that a small but significant amount of score variance in tests can be attributed to test anxiety. It has been suggested that test anxiety is reduced in an oral test, including the interview, because of the presence of a human interlocutor (Ingram, 1985; Savignon, 1972; 1985; Shohamy and Stansfield, 1990).

Scott (1986) investigated test anxiety in relationship to the group oral task and other oral tasks, and concluded that there was no essential difference between questionnaire responses of students who had taken one test or the other. Young (1986) suggests that subjects of low ability suffer more from debilitating anxiety than

high-ability students. However, the general consensus in the literature seems to be that '... ability, not anxiety, is the more important variable affecting OPI scores' (Young, 1986: 439).

In this study, students were asked to report their levels of anxiety immediately after each task, and the questionnaire in Appendix 1 is an adaptation of that from Scott (1986). All correlations between anxiety and scores were positive but nonsignificant, and therefore are not worth reporting. However, retrospective reports from students are worthy of further comment. Questions 2, 3 and 5 on the questionnaire appeared to be related to the concept of anxiety, as indicated by an iterative principal axis factor analysis which was conducted on responses (Appendix 2). In Appendix 2, loadings which are printed in **bold** indicate that these are the questions which have been interpreted as characterizing the factor. Thus questions 2, 3 and 5 load on factor 2 on the questionnaire response to task 1, on factor 3 on the questionnaire response to task 2, and factor 2 on the questionnaire response to task 3. A lower loading on a factor on only one of the three questionnaires was not interpreted as a serious problem for interpretation.

Student responses indicated that there was a fair degree of anxiety prior to doing the picture interpretation task, with less anxiety prior to taking the text prompt task and the least amount of anxiety being generated prior to the group discussion task. This could be a function of the increasing amount of time allowed for preparation prior to the task taking place. When actually doing the task, less able students reported suffering from anxiety in task 2, whereas more able students did not. It will be remembered (see above) that the text prompt task was the most difficult of the three tasks, and thus for less able students it would seem that the mismatch between task difficulty and student ability caused this to be an extremely stressful experience for them. This evidence points up the importance of matching task difficulty with ability in reducing test-related anxiety.

Question 17 on the questionnaire attempted to see if the students who reported being nervous during one or more of the oral tasks would be capable of analysing the nature or their anxiety and considering ways in which the oral test might be made less stressful. Over half the students in the sample reported some anxiety before or during taking a test. Their responses may be summarized in the following four categories.

1 Interviewer related

Most of these concerned the person interviewing the students or conducting the oral test, and fell roughly into two subcategories:

comments concerning the interviewer him or herself, and comments concerning the organization of the oral test. Learners expressed the view that the interviewer should encourage them to speak openly by helping them to relax. In the second category, learners suggested that they should meet the interviewer prior to the test, so that he or she would not be completely unfamiliar.

It would appear that part of the stress associated with an oral test does seem to be the foreboding connected with having to talk for a given length of time to a person whom one has never met before.

2 Technology related

Three students reported feeling nervous during the interviews simply because there was a video camera in the room. This point is to be taken seriously, as recording equipment may make some students more nervous than they would otherwise be. However, this cannot be helped when it is necessary to grade after the interview or when recording is used for moderation or research purposes.

3 Task related

Two students strongly believed that students should be provided with information on the format and subject of the oral test prior to the test taking place. Not knowing the subject or precisely what would be expected of them was the prime cause of anxiety for this small number of students.

4 Student related

Four students specifically commented that anxiety had nothing to do with the testing situation or the interviewer, but with their own state of mind. 'Uncertainty in myself' was one reason for nervousness, and another student said that the only way to feel less anxious was to study harder, that 'knowing more English and being more fluent' was the only way to reduce anxiety. Two students said that nothing could be done to reduce anxiety as for them it was a natural part of any testing situation.

It may be suggested (noncontroversially) that test anxiety could be reduced if appropriate interviewers are chosen and trained to be good interlocutors, if there is some form of 'warm-up' prior to the start of the test proper and if students are provided with some information about what is expected of them prior to the oral test taking place. If recording equipment is to be used during the test, its position and proximity to the students must be considered carefully.

However, it is clear from the results that the group discussion is perceived by most students as the one which induces the least anxiety.

VI Perceptions of validity

Questions 1 and 6 were seen to relate to the students' perception of the validity of the use of each task to allow them adequately to demonstrate their oral ability in a way which would allow a valid score to be awarded (see Appendix 2).

There was agreement among most students that task 1 provided them with an opportunity to speak English, but that this in itself would not allow the examiner to collect adequate evidence upon which to assign a valid score. The students who perceived the task to be most valid were those of lower estimated ability levels, and as ability increased the perceived validity of task 1 was seen to decrease. Almost half the students argued that the test was 'too easy' to constitute a test of what they were actually capable of, and typical responses to question 6 included

> no thinking was needed.
> not challenging enough.
> not demanding enough.
> too easy to do well.
> I couldn't demonstrate my knowledge of vocabulary.

Task 2 on the other hand, which was the most difficult of the three tasks, was seen as valid by all students irrespective of the level of ability. There was general agreement that the task was challenging, that there was a need to 'think quickly' and express oneself fluently. Of the few students who did not share this view, the reason given was that the task was 'too difficult', and these students came from the lower ability group. Once again, it may be seen that there is a potential relationship among perceptions of validity, task difficulty and learner ability, which needs to be further investigated.

The overwhelming response to task 3 was that engaging in a group discussion with a partner gave the students more confidence to speak and say what *they* wanted, rather than having to respond to an examiner. Only two students claimed that it was not a valid test of their oral ability: one because she believed that not enough time was allocated to the task, and the other because she believed she was 'too young' to be able to deal with topics like the national education system.

Student perceptions of tasks 1 and 2 are related to some degree to student ability levels. However, task 3 does appear, from these

results, to overcome any affective disadvantage which students may feel they have when being tested by other one-to-one task types.

VII Task enjoyment

Questions 8, 13 and 14 (which has a negative loading, see Appendix 2) appear to be related to whether or not students enjoy taking a test. It may seem rather foolish to consider this as a separate category from perceptions of validity or task difficulty, but in this study it transpired that the students who took part in this exercise were very sophisticated in their analysis of the testing experience, and could clearly distinguish between concepts of validity (in a nontechnical sense) and their own enjoyment of the experience.

It should be noted in the factor loadings that question 15, which is concerned with the students' perception of the fairness of the topic chosen for the tasks, loads on the same factor as questions 8, 13 and 14 in the questionnaires for tasks 2 and 3. It may be hypothesized that topic is related to student perceptions of enjoyment of the testing experience rather than test difficulty.

The most surprising response in the questionnaire data was that lower ability students reported enjoying taking task 2, whilst also reporting that they found the task difficult and the task taking experience stressful. This is the clearest possible indication that these concepts are separate in the minds of students, and that students may enjoy an experience which they consider to be too difficult or even invalid. On the other hand, the majority of students also reported enjoying task 1, because they found it easy. Those who reported not enjoying doing the task did so because they were not interested in the topic chosen: sports.

Almost half the students reported not enjoying task 2, and this was because of its perceived difficulty, although the number of lower-ability students responding in this way was small. Over a third of students also felt that the topic chosen, poverty in the third world, was unfair because of their lack of appropriate background knowledge and specialist vocabulary to complete the task.

Task 3 was seen as an enjoyable experience by well over half the students, reasons provided being that they could 'take the task with friends' and that 'it didn't feel like a test'. Two students specifically reported that their enjoyment was due to reduced anxiety in this test format, and a further two said that their enjoyment was increased because the examiner did not join in the discussion until the very end of the test. The main reason for not enjoying the task was the choice of topic, with only one student claiming that he preferred a one-to-one interview. In response to question 15, one further

student added that the topic may not have been fair in itself, but that the 10-minute preparation period meant that this possibly confounding factor had been removed as far as he was concerned.

Enjoyment seems to exist in a roughly inverse relationship to task difficulty, but this is somewhat complicated by the finding that weaker students still enjoy doing even the most difficult tasks. This may warrant further investigation.

VIII Students' preferences in task types

After taking all tasks, students were asked to state which of the three tasks they would prefer to take, if they were given the choice, in an oral test. The frequency of responses are classified in Appendix 3 according to the estimated ability level of the student. Task 3 was viewed as the most preferable, followed by task 1 and then task 2. This result merely confirms the adequacy of the questionnaire in eliciting consistent responses in relation to the students' perception of the tasks, and provides some confidence in the results reported.

IX Generalizability of test scores from one task to another

Wilkinson (1968: 125) claimed that 'It is not known whether we can speak of the candidate's speech ability as a general factor, or whether it can only be defined in relation to a specific situation'. In the field of testing, supporters of 'performance-based' testing '... seem to be arguing that it is only necessary to select certain representative communication tasks, as we do not use the same language for all possible communication purposes' (Weir, 1988: 15). That is, 'performance based' refers not only to the test format but also to the scoring methods. This immediately limits the generalizability of any scores to the types of task used in the test (Messick, 1994), as the scoring system adopted for the test would contain descriptors which were directly related to the test situation itself. This approach thus fuses inextricably together the testing method and what it is we wish to test. It unites trait and test method facet.

In the field of SLA, Tarone (1983: 147–52; 1985) has argued that underlying performance is 'capability' which is heterogeneous and varies by speech style, which is related to the nature of the task being undertaken. 'Capability' is therefore variable. Skehan (1987: 200) accepts the implications of this view for testing, stating that the main problem for language testers is one of sampling tasks to achieve a representative sample for the purpose for which the scores are going to be used (see Fulcher, forthcoming).

If this view, irrespective of the different theoretical positions

Table 2 Results of a G-study

Source	Sum of squares	Df	Mean square	F-ratio	*p*
Rater	72.51	4	18.12	12.90	.001
Task	17.34	2	8.76	6.14	.002
Rater * task	15.90	8	1.99	.42	.186
Residual	948.22				

which generated it, were to be accurate, it would predict that in a multitask test which could be demonstrated to be reliable, a G-study would produce a low equivalent forms generalizability coefficient (EFGC) between the tasks. It would also predict that a Rasch partial credit study would generate high outfit statistics for some tasks. That is, the scores on one task could not be generalized to the other task and that they could not be assumed to be measurable on a unidimensional scale. Tables 2 and 3 present the results of a G-study for the three tasks used in this research. Although both rater and task are significant sources of variance, the EFGC is .98. It should be noted that in the ANOVA table the residual, that proportion of the variance which cannot be attributed to rater or task but to student ability, is much greater than the variance attributable to either rater or task. The conclusion can only be that while task does have a significant effect upon scores, this effect is so small that it does not seriously reduce the ability to generalize from one task to another.

In Table 1 we also note that the outfit statistics do not approach the critical figure of 2, which indicates that these tasks are operating on a unidimensional scale. Thus, both the G-study and the Rasch partial credit analysis provide evidence that, although there is a difference between the tasks, these are not sufficient to result in loss of generalizability of scores from one task to another. Without having to take a theoretical stance on the issues in SLA, these results would certainly support the position of Gregg (1990) that synchronic variability is a feature of performance, and that trait is a much more stable element which generates that performance.

The rating scale which was used in this study (Appendix 4) did not contain any reference to any of the test method facets of the three tasks. However, it is also hypothesized that, if the rating scale had confounded test method facets and traits, the EFGC would have

Table 3 Reliability coefficients

Reliability coefficient	.90
Inter-rater generalizability coefficient	.93
Equivalent forms generalizability coefficient	.98
Forms by raters generalizability coefficient	.99

been greatly reduced and the outfit statistics increased as a result. This hypothesis would bear further investigation in the future.

If ability is to be distinguished from behaviour (Bachman, 1990: 308) and trait from test method facet (Bachman and Savignon, 1986), as I believe they have to be for the study of construct validity, then whatever ability or abilities one hypothesizes underlie test performance, they must contribute to performance in a variety of situations and tasks. The problem of generalizability from one task to another is thus one which should be tackled through the development of the scoring system and not necessarily the design of the task. In relationship to indirect testing, Hughes (1989: 16) suggests that indirect tests '... offer the possibility of testing a representative sample of a finite number of abilities which underlie a potentially indefinitely large number of manifestations of them'.

The argument is that if the scoring system is related to underlying abilities and does not confound these with test method facets, one of which is the nature of the task, then what Hughes sees as a major advantage of an indirect test could also become a major advantage of a direct oral test, irrespective of the task type used.

The data and argument presented here suggest that, although there is a task effect in oral testing which has frequently been commented on in the literature, this may not be as large as has often been assumed when rating scales which do not contain descriptors which refer to test method facets are used. That is, large task effects may be an artifact of the rating scale used.

X Conclusions

Writing with knowledge of their own situation, dealing with students of economics, business studies and administration at the University of Tampere, Finland, Folland and Robertson (1976) introduced and developed a group oral test as an improvement to the situation as they found it. From their observations in this one context, they concluded that the group oral examination had many advantages over more traditional one-to-one interview oral tests. They argued that

> The position of the examiner is greatly altered. He is now an observer – of a 'real-life' situation. He must not interfere in the discussion and therefore cannot influence the testees, or be inconsistent in his way of holding the test. The discussion is controlled by the testees and the examiner is there to evaluate, according to fixed criteria, the linguistic content alone. There is also the advantage that because the examiner does not participate, the testees have more time, and perhaps even inclination, to speak the language.
>
> The situation created in a group test helps the testees feel more at ease, less under examination stress than when they must appear individually before a

> possibly unknown examiner. The discussion which develops gives more incentive to the testee to speak and exhibit his ability to use the language, especially since he can himself alter the course of the discussion (Folland and Robertson, 1976: 161–62).

Whether or not the discourse of the group oral more closely approximates some situation external to that of the test and less like 'test talk' is still a claim which remains to be empirically investigated. However, research into affective responses of students to different task types bears out their initial intuitions, if not completely, for the most part.

The call for more 'human' tests (Underhill, 1987) and the frustrations often felt by those who teach and test oral skills, exemplified by Robertson (1993), can be tackled by serious research into the fields discussed in this article. Alderson (1985: 101) argued that introspective and, more practically in the field of oral testing, retrospective data from students would provide valuable information about the ways in which students deal with test items. What they think about the tests, and what effect the tests have upon them as individuals, are equally important areas for investigation. The more we know about such issues, the more likely we are to limit the effect of variables other than the abilities we wish to measure on the test scores. It is hoped that this research will provide ideas which may be investigated by others working in the field of oral testing, in the interests of providing students with reliable and valid oral tests, which they may also perceive as being reliable, valid, as stress free as possible and enjoyable.

Important as task design is, and however much we may wish to strive to create tasks which are more likely to result in desirable discourse outcomes, it must also be remembered that this endeavour may not necessarily be the solution to the problem of generalizability of oral test scores. Evidence has also been presented here to suggest that however intuitive it may now appear that the nature of the task will effect not only student performance but also scores, this effect is limited if a rating scale which does not refer to test method facets in its descriptors is used. This finding should be welcomed by all those whose aim in testing remains to predict from the testing situation, to the universe of tasks which exist in the world around us. That is, generalization from one task to another is possible.

Further research has been suggested in the following areas:

- Comparative analysis of the discourse produced by different task types, keeping participant variables steady.
- The assessment of the comparative difficulty of a variety of oral task types.

- Studies of the relative factors contributing to task design which account for the 'success' or 'failure' of the task in a test.
- Student perceptions of task validity and difficulty in relation to student ability, and enjoyment of taking tasks.
- An investigation of the hypothesis that the more trait and method are confounded in rating scale descriptors, the lower the equivalent forms generalizability coefficient and the higher the outfit statistics will be in a G-study and a Rasch partial credit analysis, respectively.

Acknowledgements

This study is an updated version of one section of a doctoral dissertation submitted to the University of Lancaster in 1993. I should like to thank Professor Charles Alderson for all the help and advice which I received while working in this field with him. Responsibility for any errors, and the views expressed, remain mine.

XI References

ACTFL 1986: *ACTFL proficiency guidelines.* Hastings-on-Hudson, NY: American Council on the Teaching of Foreign Languages.

Alderson, J.C. 1985: Innovations in language testing? In Portal, M., editor, *Innovations in language testing.* Slough: NFER-Nelson, 93–105.

Alderson, J.C., Krahnke, K.J. and **Stansfield, C.**, editors, 1987: *Reviews of English language proficiency tests.* Washington, DC: TESOL.

Alderson, J.C. and **North, B.**, editors, 1991: *Language testing in the 1990s.* London: Modern English Publications and the British Council.

Bachman, L.F. 1990: *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L.F. and **Savignon, S.J.** 1986: The evaluation of communicative language proficiency: a critique of the ACTFL oral interview. *Modern Language Journal* 70, 380–90.

Bartz, W.M. 1979: *Testing oral communication in the foreign language classroom. Language in Education: Theory and Practice* 17. Arlington, VA: ERIC Clearinghouse on Languages and Linguistics.

Berkoff, N.A. 1985: Testing oral proficiency: a new approach. In Lee, Y.P., editor, *New directions in language testing.* Oxford: Pergamon Institute of English, 93–100.

Brown, A. 1993: The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10, 277–303.

Ellis, R. 1992: *Second language acquisition and language pedagogy.* Philadelphia, PA: Multilingual Matters.

Folland, D. and **Robertson, D.** 1976: Towards objectivity in group oral testing. *English Language Teaching Journal* 30, 156–67.

Fulcher, G. forthcoming, 1995: Variable competence in second language acquisition: a problem for research methodology? *System* 23, 1.

Galassi, J.P., Frierson, H.T. and **Siegel, R.G.** 1984: Cognitions, test anxiety, and test performance: a closer look. *Journal of Consulting and Clinical Psychology* 52, 319–20.

Gregg, K. 1990: The variable competence model of second language acquisition, and why it isn't? *Applied Linguistics* 11, 364–81.

Halliday, M.A.K. and **Hasan, R.** 1985: *Language, context and text: aspects of language in a social-semiotic perspective.* Australia: Deakin University Press.

Harder, P. 1980. Discourse as self-expression – on the reduced personality of the second-language learner. *Applied Linguistics* 1, 262–70.

Hatch, E. 1992: *Discourse and language education.* Cambridge: Cambridge University Press.

Hilsdon, J. 1991: The group oral exam. Advantages and limitations. In Alderson, J.C. and North, B., editors, *Language testing in the 1990s.*

Hughes, A. 1989: *Testing for language teachers.* Cambridge: Cambridge University Press. Modern English Publications and the British Council 187–9.

Ingram, D. 1985: Assessing proficiency: an overview on some aspects of testing. In Hyltenstam, K. and Pienemann, M. editors, *Modeling and assessing second language acquisition.* San Diego, CA: College Hill Press, 215–276.

Kenyon, D.M. and **Stansfield, C.** 1991: A method for improving tasks on performance assessments through field testing. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Lazaraton, A. 1992: The structural organization of a language interview: a conversation analytic perspective. *System* 20, 373–86.

Linacre, J.M. 1991: *FACETS computer program for many faceted Rasch measurement.* Chicago, IL: Mesa Press.

Linacre, J.M. and **Wright, B.D.** 1990: *Facets: many faceted Rasch analysis.* Chicago, IL: Mesa Press.

Long, M. 1983: Native speaker/non-native speaker conversation and the negotiation of meaning. *Applied Linguistics* 4, 126–41.

Lowe, P. 1983: The ILR oral interview: origins, applications, pitfalls, and implications. *Die Unterrichtspraxis* 16, 230–44.

—— 1987: Interagency language roundtable proficiency interview. In Alderson, J.C., Krahnke, K.J. and Stansfield, C., editors, *Reviews of English language testing.* Washington DC: TESOL, 43–47.

Lowe, P. and **Liskin-Gasparro, J.** 1986: *Testing speaking proficiency: the oral interview.* Washington, DC: ERIC Digest.

MacPhail, J. 1985: Oral assessment interviews: suggestions for participants. Unpublished MA dissertation, University of Lancaster, Department of Linguistics and Modern Languages.

Madsen, H.S. and **Murray, N.** 1984: Retrospective evaluation of testing in ESL content and skills courses. Unpublished manuscript, Brigham Young University.

Messick, S. 1994: The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher* 23, 13–23.

Morrison, D.M. and **Lee, N.** 1985: Simulating an academic tutorial: a test validation study. In Lee, Y.P., editor, *New directions in language testing.* Oxford: Pergamon Institute of English, 85–92.

Morrow, K. 1982: Testing spoken language. In Heaton, J.B., editor, *Language testing.* London: Modern English Publications, 56–58.

Nobuyoshi, J. and **Ellis, R.** 1993. Focused communication tasks and second language acquisition. *English Language Teaching Journal* 47, 203–10.

Nunan, D. 1989: *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Perrett, G. 1990: The language testing interview: a reappraisal. In de Jong, J.H.A.L. and Stevenson, D.K., editors, *Individualizing the assessment of language abilities.* Philadelphia, PA: Multilingual Matters, 225–38.

Reves, T. 1980: The group-oral test: an experiment. *English Teachers' Journal* 24, 19–21.

—— 1991: From testing research to educational policy: a comprehensive test of oral proficiency. In Alderson, J.C. and North, B., editors, *Language testing in the 1990s.* London: Modern English Publications and the British Council, 178–88.

Robertson, E.F. 1993: Can oral tests be made more communicative? *Testing Newsletter (IATEFL Testing Special Interest Group)* December, 15–16.

Rodriguez, M.C. 1984: The current status of the OPI in intensive language programs. Paper presented at the annual meeting of the Southwest Conference on the Teaching of Foreign Languages, Colorado Springs, CO.

Savignon, S.J. 1972: *Communicative competence: an experiment in foreign language teaching.* Philadelphia, PA: The Center for Curriculum Development.

Scarcella, R.C. 1983: Discourse accent in second language performance. In Gass, S.M. and Selinker, L., editors, *Language transfer in language learning.* Rowley, MA: Newbury House, 306–26.

Scott, M.L. 1986: Student affective reactions to oral language tests. *Language Testing* 3, 99–118.

Shohamy, E. 1983: The stability of oral proficiency assessment in the oral interview procedure. *Language Learning* 33, 527–40.

—— 1988: A proposed framework for testing the oral language of second/foreign language learners. *Studies in Second Language Acquisition* 10, 165–79.

—— 1990a: Language testing priorities: a different perspective. *Foreign Language Annals* 23, 385–94.

—— 1990b: Discourse analysis in language testing. *Annual Review of Applied Linguistics* 11, 115–28.

Shohamy E., Reves, T. and **Bejarano, Y.** 1986: Introducing a new compre-

hensive test of oral proficiency. *English Language Teaching Journal* 40, 212–20.

Shohamy, E. and **Stansfield, C.** 1990: The Hebrew speaking test: an example of international cooperation in test development and validation. In de Jong, H.A.L., editor, *Standardization in language testing.* Philadelphia, PA: 79–90.

Silverman, D. 1976: Interview talk: bringing off a research instrument. In Silverman, D. and Jones, J., editors, *Organizational work: the language of grading, the grading of language.* London: Collier Macmillan, 133–50.

Skehan, P. 1987: Variability and language testing. In Ellis, R., editor, *Second language acquisition and language pedagogy.* Philadelphia, PA: Multilingual Matters, 195–206.

Stansfield, C.W. and **Kenyon, D.M.** 1992: Comparing the scaling of speaking tasks by language teachers and by the ACTFL guidelines. Paper presented at the 14th annual Language Testing Research Colloquium, Vancouver, BC.

Stevenson, D.K. 1985a: Authenticity, validity and a tea party. *Language Testing* 2, 41–47.

—— 1985b: Pop validity and performance testing. In Lee, Y.P., editor, *New directions in language testing.* Oxford: Pergamon Institute of English, 111–18.

Tarone, E. 1983: On the variability of interlanguage systems. *Applied Linguistics* 4, 146–63.

—— 1985: Variability in interlanguage use: a study of style-shifting in morphology and syntax. *Language Learning* 35, 373–403.

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

Upshur, J.A. 1971: Objective evaluation of oral proficiency in the ESOL classroom. *TESOL Quarterly* 5, 47–59.

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

van Weeren, J. 1981: Testing oral proficiency in everyday situations. In Klein-Braley, C. and Stevenson, D.K., editors, *Practice and problems in language testing. Volume 1.* Frankfurt: Bem, 54–59.

Weir, C. 1988: *Communicative language testing.* Exeter: University of Exeter.

Wilds, C. 1979: The measurement of speaking and reading proficiency in a foreign language. In Adams, M.L. and Frith, J.R., editors, *Testing kit: French and Spanish.* Department of State: Foreign Service Institute.

Wilkinson, A. 1968: The testing of oracy. In Davies, A., editor, *Language testing symposium.* Oxford University Press, 117–32.

Young, D. 1986: The relationship between anxiety and foreign language proficiency ratings. *Foreign Language Annals* 19, 439–45.

Zuengler, J. 1993: Encouraging learners' conversational participation: the effect of content knowledge. *Language Learning* 43, 403–32.

Appendix 1

In the following questionnaire, it should be noted that questions 1–15 were asked for each task taken by the students. Questions 16–19 were separate questions added to the end of the three other questionnaires. This example is from task 1, with questions 16–19 added to the end.

A. Please complete these details.
Name: _____
Age: _____ years, _____ months.
Class at school: _____

B. Please complete the following by placing a circle around the most appropriate answer.
For example:

Question: It is useful to study the day before an oral test.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

1. I believe that the picture task would provide an examiner with an accurate idea of my ability to speak English.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

2. I felt nervous before the picture task.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

3. I felt nervous while I was doing the picture task.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

4. I believe I did well on the picture task.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

5. If I had done the picture task on another day, I would have done better.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

6. I believe that the picture task provided me with an adequate opportunity to demonstrate my ability to speak English.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

Please explain why: _____

_____

_____

7. The time allowed for the picture task was too short.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

8. I liked doing the picture task.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

Please explain why: _____

_____

_____

9. I understood what I was supposed to do in the picture task.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

10. I thought that the picture task was related to what I learn in class.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

11. If a different teacher had conducted the picture task, I would have done better.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

12. I thought that the picture task was too difficult.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

Please explain why: _____

_____

_____

13. I thought that the picture task was interesting.

Strongly agree | Agree | No opinion | Disagree | Strongly disagree

14. I thought that doing the picture task was an unpleasant experience.

Strongly agree	Agree	No opinion	Disagree	Strongly disagree

15. Did you think that the subject chosen for the picture task was particularly fair or unfair? Please give your reasons.

Very fair	Fair	No opinion	Unfair	Very unfair

Reasons: _____

_____

_____

16. If you were going to take an oral test in an examination, which one of the three tasks would you prefer to do? Put a '1' next to the task you would prefer most, a '2' next to your second choice, and a '3' next to the task you would least like to do.

Task 1: Picture task _____
Task 2: Discussion of passage _____
Task 3: Group discussion _____

17. If you felt nervous during any of the tasks, what would have made you feel less nervous?

_____

_____

_____

_____

18. How would you rate your own proficiency in spoken English?

Generally:	Very good	Good	Average	Poor	Very poor
Grammatical accuracy:	Very good	Good	Average	Poor	Very poor
Fluency:	Very good	Good	Average	Poor	Very poor

19. For how many years have you been studying English?

_____

Appendix 2: Factor analysis of the questionnaire by task

Varimax rotated factor analysis of the questionnaire for task 1

	Factor 1	Factor 2	Factor 3	Factor 4
Question 1	**.23**	−.25	−.06	.19
Question 2	.07	**.70**	.21	.04
Question 3	.07	**.71**	.08	.22
Question 4	.02	.06	.05	.18
Question 5	.05	**.60**	.22	−.34
Question 6	**1.00**	−.02	.13	−.10
Question 7	.07	−.39	.29	−.18
Question 8	−.07	.02	**.82**	.50
Question 9	.04	−.32	−.01	.03
Question 10	.08	.41	−.32	−.03
Question 11	−.07	.26	−.21	−.02
Question 12	.03	.17	.03	.47
Question 13	.05	.13	**.65**	.02
Question 14	−.07	−.04	**−.12**	−.63
Question 15	**.24**	.16	.23	.26
Variance explained by rotated factors	1.99	1.93	1.51	1.21
% Variance explained by rotated factors	13.28	12.84	10.07	8.09
Total % of variance explained				44.28

Varimax rotated factor analysis of the questionnaire for task 2

	Factor 1	Factor 2	Factor 3	Factor 4
Question 1	.00	**.76**	.17	−.03
Question 2	−.24	.03	**.83**	.38
Question 3	**.06**	.17	**.61**	−.02
Question 4	.35	−.06	.04	−.24
Question 5	−.14	−.07	**.16**	.47
Question 6	**.55**	**.41**	−.01	−.08
Question 7	−.01	**.20**	−.15	.32
Question 8	**.80**	.22	−.20	.07
Question 9	**.69**	−.22	−.13	.08
Question 10	.08	−.06	.11	.55
Question 11	−.27	.25	.22	.05
Question 12	−.59	**.20**	.05	.34
Question 13	**.44**	.29	.13	.24
Question 14	**−.75**	−.05	.40	.03
Question 15	**.71**	.04	.17	−.08
Variance explained by rotated factors	3.30	1.12	1.44	1.03
% Variance explained by rotated factors	21.97	7.45	9.62	6.85
Total % of variance explained				45.89

Varimax rotated factor analysis of the questionnaire for task 3

	Factor 1	Factor 2	Factor 3	Factor 4
Question 1	.33	−.02	−.13	**.95**
Question 2	.02	**.85**	.04	.12
Question 3	−.08	**.74**	.04	.14
Question 4	.54	−.45	.22	.09
Question 5	−.02	**.69**	−.04	−.13
Question 6	.52	−.10	−.03	**.49**
Question 7	.17	−.19	.56	−.12
Question 8	**.91**	−.12	.02	.04
Question 9	.15	−.23	.04	−.32
Question 10	.35	.09	.22	−.06
Question 11	−.13	.16	.73	.00
Question 12	−.55	.35	.10	.17
Question 13	**.78**	.09	−.07	.19
Question 14	**−.57**	.34	.19	.03
Question 15	**.70**	.05	.16	.26
Variance explained by rotated factors	3.41	2.35	1.04	1.46
% Variance explained by rotated factors	22.72	15.66	6.96	9.72
Total % of variance explained				55.06

Appendix 3: Task preference by estimated ability level of students

Ability level	High	Medium	Low
No. of cases	10	17	18
Preference for task 1	4	5	7
%	40	29	39
Preference for task 2	2	2	2
%	20	12	11
Preference for task 3	4	10	9
%	40	59	50

Appendix 4

Band 1

The candidate frequently pauses in speech before completing the propositional intention of the utterance, causing the interviewer to ask additional questions and/or make comments in order to continue the conversation. Utterances tend to be short, and there is little evidence of candidates taking time to plan the content of the utterance in advance of speaking. However, hesitation is frequently evident when the candidate has to plan the utterance grammatically. This often involves the repetition of items, long pauses and the reformulation of sentences.

Misunderstanding of the interviewer's questions or comments is fairly frequent, and the candidate sometimes cannot respond at all, or dries up part way through the answer. Single word responses followed by pauses are common, forcing the interviewer to encourage further contribution. It is rare for a band 1 candidate to be able to give examples, counter examples or reasons, to support a view expressed.

Pausing for grammatical and lexical repair is evident (selection of a new word or structure when it is realized that an utterance is not accurate or cannot be completed accurately.)

Candidates at band 1 may pause because of difficulty in retrieving a word, but when this happens will usually abandon the message rather than attempt to circumlocute. It is rare for a band 1 candidate to express uncertainty regarding choice of lexis or the propositional content of the message (the message itself is often simple).

Band 2

A band 2 candidate will almost always be able to complete the propositional intention of an utterance once started, causing no strain on the interviewer by expecting him or her to maintain the interaction. However, just like a band 1 candidate, a band 2 candidate will frequently misunderstand the interviewer's question or be completely unable to respond to the interviewer's question, requiring the interviewer to read the question or clarify what he or she wishes the candidate to do. Similarly, single word responses are common, forcing the interviewer to encourage further contribution.

Although the candidate will spend less time pausing to plan the grammar of an utterance, it will be observed that there are many occasions on which the candidate will reformulate an utterance having begun using one grammatical pattern and conclude with a different form. Similarly, with lexis, there will be evidence that the candidate pauses to search for an appropriate lexical item and, if it is

not available, will make some attempt to circumlocate even if this is not very successful. From time to time a band 2 candidate may pause to consider giving an example, counterexample or reason for a point of view. However, this will be infrequent and when it does occur the example or reason may be expressed in very simplistic terms and may lack relevance to the topic.

Band 3

A candidate in band 3 will hardly ever misunderstand a question or be unable to respond to a question from the interviewer. On the odd occasion when it does happen a band 3 candidate will almost always ask for clarification from the interviewer.

Most pauses in the speech of a band 3 candidate will occur when they require 'thinking time' in order to provide a propositionally appropriate utterance. Time is sometimes needed to plan a sentence grammatically in advance, especially after making an error which the candidate then rephrases.

A band 3 candidate is very conscious of his or her use of lexis, and often pauses to think about the word which has been used, or to select another which they consider to be better in the context. The candidate may even question the interviewer overtly regarding the appropriacy of the word which has been chosen.

Often candidates in this band will give examples, counterexamples or reasons to support their point of view.

(At band 3 and above there is an increasing tendency for candidates to use 'backchannelling' – the use of *hm* or *yeah*.)