Does thick description lead to smart tests? A data-based approach to rating scale construction

Glenn Fulcher *University of Surrey*

Within the fields of applied linguistics and language testing, there has been a recent interest in rating scales, and how rating scales are constructed (Upshur and Turner, 1995). This is not surprising, as there is increasing concern that scores from language tests should be meaningful in applied linguistics terms. However, applied linguistics research and second language acquisition research have done little to provide descriptions of language abilities or performances which can be operationalized by language testers. Many existing descriptors for bands in rating scales are therefore barely tenable as definitions of constructs.

This article looks at the definition of fluency in the literature, and proposes a qualitative and quantitative approach which may be used to produce a 'thick' description of language use, which can be used in rating scale construction. A fluency rating scale is described, and its reliability and validity assessed. The article suggests that validity considerations must be addressed in the construction phase of developing rating scales, through the careful consideration of the linguistic meaning of constructs, rather than merely as a *post hoc* enterprise.

I Introduction

Rating scales have tended to be a priori measuring instruments. By a priori it is meant that the descriptors of the rating scales are constructed by an expert, often using his or her own intuitive judgement concerning the nature of developing language proficiency, sometimes in consultation with a team of other experts, whether these are colleagues within an examination board, or a wider sample of specialists working within the field. A priori methods can be broken down into more specific development methodologies (North, 1994a), but they mostly have in common the lack of any empirical underpinning, except as *post hoc* validity studies (Jarvis, 1986: 21). Nevertheless, rating scales remain attractive scoring devices for tests, because their descriptors appear to be meaningful definitions of what students can do. This is especially welcome by score users. Sometimes the attraction of the rating scale has seduced researchers into attempting to

'index' norm-referenced numerical scores to rating scale descriptors (Boldt, 1991), whether this is a valid procedure or not.

Suggestions for an empirical basis to be introduced at the construction stage of rating scales have been made in the past (Fulcher, 1987; 1988; Shohamy, 1990), but rarely taken up. One refreshing exception is a study by Upshur and Turner (1995) which describes an empirically based approach to scale construction which they term 'empirically derived binary-choice, boundary-definition scales' (p. 6). The purpose of this study is to describe the concept of fluency in the applied linguistics and testing literature, and to present a methodology whereby observations from student performance can be utilized in the construction of a fluency rating scale. The adequacy of the methodology and the resulting rating scale will be assessed.

II Background: the concept of fluency

'Fluency', a term with as many definitions as there are commentators, has been chosen as a notion for this study because it is widely assumed in oral testing that the two constructs of fluency and accuracy are separate aspects of oral ability (Griffin, 1985), and in language teaching it is common for a strong difference to be drawn between fluency activities and accuracy activities (see, for example, Brown and Yule, 1983: 104; Brumfit, 1984; Klippel, 1984; Rixon, 1992: 81; Hedge, 1993).

In research literature this difference is variously referred to as that between 'norm-oriented' and 'communicative-oriented' learners (Clahsen, 1985), 'rule-formers' and 'data-gatherers' (Hatch, 1974) or 'planners' and 'correctors' (Seliger, 1980). In each case, the former of the two categories refers to students who concentrate on building up grammatical rules and aim for accuracy in production (at the expense of fluency), and the latter category refers to students who concentrate on communicating fluently, paying little attention to accuracy or, alternatively, using their knowledge of the grammar and lexicon of the language only to correct themselves after they have communicated. This 'basic polarity', as Brumfit (1984: 50–57) describes it, predicts that these two components of oral proficiency will tend to develop separately depending upon the learning orientation of the student. Such a polarity would also seem to be predicted by the distinction often drawn between 'gestalt' and 'analytic' learners (Dulay, Burt and Krashen, 1982: 237–38; Larsen-Freeman and Long, 1991: 196–97), and the two-dimensional modular theory of second language acquisition which posits knowledge-based and control-based language processing (Bialystok and Sharwood Smith, 1985).

Although little empirical research has been conducted into the distinction between students who are 'norm-oriented' and those who are 'communicative-oriented' to date, Ellis (1990) conducted a study in which it was hypothesized that students who focus on accuracy would acquire linguistic knowledge quickly, while students who focus on fluency would develop 'channel control mechanisms' (measured as speech rate) much more quickly. Using correlation and a principal components analysis, Ellis (1990: 89–90) concluded that

> ... those learners who showed the greatest gain in acquisition of the three word order rules manifested the smallest gain in general oral fluency and, conversely, those learners who developed the ability to process speech the most rapidly displayed the smallest gain in accuracy.

Douglas's (1994: 131–32) working definition of fluency was the ratio between types and tokens in the transcriptions of student speech. He reported a negative correlation of −.91 between the ratio scores and fluency rating scale scores for six students, leading him to conclude that there was 'very little relationship ... between the scores on the test and the language actually produced' (Douglas, 1994: 134). However, a negative correlation of −.91 indicates a very strong inverse relationship, which could mean either that the raters do not understand and appropriately apply the scale descriptors, or that the type/token ratio definition of fluency is an inadequate operationalization of the concept.

It must be stressed that empirical work is limited, and inconclusive. The definitions of fluency which exist seem to be inadequate for the purpose of operationalization in a test, even though the concept is widespread in the literature. A brief consideration of references to fluency in three tests confirms this.

In the history of testing, the development of a fluency rating scale has been plagued by a lack of operational specificity since the very earliest Foreign Service Institute (FSI) component oral rating scales. In the FSI component scale for fluency, the scale constructors relied on vague concepts such as 'slow and uneven speech' at band 2, and 'hesitant and jerky speech' at band 3. At band 4 'groping for words' and 'unevenness' are said to result in rephrasing, while by bands 5 and 6 speech is said to be 'smooth' (Sollenberger, 1978; Clark and Clifford, 1988). The criterion of 'hesitation' is also frequently found within rating scales of fluency. Some rating scales rely on the notion of hesitation alone, despite the fact that the phenomenon of hesitation is not well understood (Fulcher, 1987).

Two examples from existing rating scales will suffice to highlight the point that rating scales frequently work with a very basic and inadequate operationalization of the concept of fluency. These are

taken from the University of Cambridge Local Examinations Syndicate First Certificate (FCE) and Certificate of Proficiency (CPE) examinations (see Hamp-Lyons, 1987, and Davies, 1987, for reviews of these tests respectively):

The FCE Fluency rating scale

5 Comfortable and natural speed and rhythm in everyday contexts, though there may be some hesitation when speaking in more abstract topics.
4 In everyday context speaks with minimal hesitation. Hesitation when discussing abstract topics, but does not demand unreasonable patience of the listener.
3 Does not hesitate unreasonably in everyday contexts, though may experience some difficulty with more abstract topics.
2 Unacceptable hesitation, even in everyday contexts.
1 Speech very disconnected.
0 Not capable of connected speech.

The CPE Fluency rating scale

5 Virtually native-speaker speed and rhythm, and coherent presentation of thoughts, in all contexts.
4 Foreign, but with minimal hesitation in all contexts.
3 Minimal hesitation in everyday contexts, but some hesitation when discussing more abstract topics, though not such as to demand unreasonable patience of the listener.
2 Hesitation not unreasonable in everyday contexts, but impedes understanding on more abstract topics.
1 Speaks haltingly even in everyday contexts.
0 Not capable of connected speech.

It would appear that the only theory (if theory it can be called) underlying these two scales amounts to the probably untested assumptions that : 1) the better a student's fluency in English the less hesitation will be evident in performance; 2) hesitation is much more likely when talking about 'abstract' topics than 'everyday' topics; and, given the CPE band 5 descriptor, 3) native speakers do not hesitate a great deal.

Fluent speech is often 'disconnected', and 'speed' and 'rhythm' remain undefined: an analysis of native speaker talk reveals that unless the speech is deliberately preplanned, hesitations and reformulations will abound (Fulcher, 1987). For example, in the 16 brief conversational extracts in Crystal and Davy (1975), 146 examples of the hesitation marker/filler 'er(m)' were found, showing that assumption (3) above is likely not to be the case.

If the third assumption is false, it would follow that the first assumption is also likely to be false, but would benefit from empirical investigation. The real question is why the speech of both native speakers and foreign learners contains hesitation. The second assumption is merely baffling. Apart from the fact that 'everyday' and 'abstract' are not defined but left to the intuition of the

interviewer/assessor, there is no empirical evidence to suggest that talking about, say, politics in the third world would prompt more hesitation than describing one's kitchen. The opposite may very well be true, as in courses leading to the FCE and CPE examinations students often discuss such subjects – subjects on which they might reasonably be expected to have some opinion.

At present, the essential logic behind most rating scales currently in use is that there is a monotonic development of Fluency from 0 to 'perfect'. The rating scales can then be written by an individual (or team of experts) who may or may not be aware of the definitions of fluency which have been attempted by the researchers in the field. Alderson (1991) describes the process of the use of expert judges in drafting descriptors, and using marking bands to extract 'key features' from actual interviews. Alderson's chapter is a description of one of the most thorough attempts to construct rating scales by this method. However, in many cases even this may not be done. It is much more common for a priori notions to form the basis of scale development, such as the 'native speaker' yardstick in the top band of a rating scale (Frith, 1979), or the principle that each definition should exist 'in the context of the whole scale and in relation to adjacent definitions' (Ingram, 1982: 9). This is a circular, self-contained notion of scale development which appears to lack empirical support or theoretical credibility.

III Method

1 Subjects

The data which were used for this study came from 21 oral interviews conducted with Greek-speaking learners of English. Their average ELTS band score was 6, with a range of 4–9. The 21 interviews were all conducted in the same room using the same interviewer, on three separate occasions within the space of one and a half weeks. This was done in order to control for test method facets such as physical environment and the style of the interviewer as much as possible. All the interviews were transcribed for later analysis.

2 Materials

All subjects took the ELTS oral interview (see Weir, 1987) in a live test, which was recorded with permission of the British Council and the University of Cambridge Local Examinations Syndicate. The ELTS oral interview was used in this study because of the accessibility of students taking live interviews, and at the time the ELTS

system was due to be decommissioned and replaced by the International English Language Testing System (IELTS), making it possible to use transcripts from an operational test without threatening test security.

3 Procedure

The procedure used in this research was to analyse the transcripts of the interviews for fluency phenomena, in conjunction with the audiorecordings of the interviews. The initial working construct of fluency is discussed below. As the researcher was present at each of the interviews, and the transcriptions and interpretations were undertaken immediately after each session, the data for interpretation were rich. In that each of the interviews contributed to the development of a system of coding for fluency phenomena, which was checked against the interpretation of new data from each new interview, the methodology resembles that used in Grounded Theory methodology (Glaser and Strauss, 1967), and at the end of the iterative process of qualitative interpretation, the results were analysed statistically. Although essentially a qualitative methodology, Grounded Theory methodology admits the usefulness of quantitative approaches (Glaser and Strauss, 1967: 17–18). At the end of the iterative process, all transcripts were recorded for analysis, using the final coding system which had been developed.

It became clear after the project had been completed that it would have been appropriate, immediately after the final coding of transcripts, to obtain codings from additional applied linguists and investigate the reliability of the coding system. However, at the time it was decided that it would be adequate to use a method that would allow prediction of oral rating grades from the simple frequency counts taken from the coding system. This was clearly an error in the research design, but is partly compensated for by very impressive concurrent validity results (see later).

In this study, discriminant analysis was used to investigate to what extent frequencies of phenomena counts within codes could predict the band/level into which each learner had been placed by an ELTS test. This procedure only assumes that the ELTS test is capable of rank ordering the learners according to ability, not that its rating scale is valid. Discriminant analysis allows the researcher to look at the extent to which simple tallies of phenomena occurrence (as described below) are capable of categorizing students in the same way as they have been on a direct test. To the extent to which the codings allow such prediction, the discriminant analysis provides evidence of concurrent validity.

A fluency rating scale was then generated from the data, and used in a test on a new sample of students drawn from the same population. Rater reliability was calculated using a G-study, while validity was assessed using the group difference method, and a Rasch partial credit analysis. A description of these procedures is provided below.

a A note on qualitative methodologies: The use of qualitative approaches in test design is becoming more popular because of the need to achieve applied linguistics descriptions which make test scores meaningful. Grounded theory methodology is a useful qualitative approach in testing, as it is 'a general methodology for developing theory that is grounded in data sytematically gathered and analysed. Theory evolves during actual research, and it does this through continuous interplay between analysis and data collection' (Strauss and Corbin, 1994: 237).

The approach expects that theory will be generated directly from data, and that through an iterative process new data will be checked to see if they fit the theory being developed. Unlike other qualitative methodologies, such as analytic induction (Manning, 1991), quantitative approaches may be used in addition to the qualitative interpretation, adding to the process of theory development. That is, although Grounded Theory is a postmodernist qualitative approach which demands 'thick description' (Geertz, 1973), it is unique in that it admits of theory development from qualitative analysis with quantitative checks. As such, in the process of conducting a qualitative analysis, it is necessary to devise systematic coding procedures for data, so that '. . . theories are always traceable to the data that gave rise to them within the interactive context of data collecting and data analysing, in which the analyst is also a crucial significant interactant' (Strauss and Corbin, 1994: 278–79). However, theory generated by induction, like all theory, can never be said to be 'true', as it will always be underdetermined.

The use of qualitative methodologies in applied linguistics is not new. Long (1983: 23) notes that researchers recognize 'bias in one person reporting events', but that this is necessary to produce meaningful, or 'thick', descriptions. He develops the notions of 'low' and 'high' inference categories in coding data. It may also be argued that the constant interaction between data and theory development through systematic coding is at the heart of developments in revealing the generic structure of texts (Swales, 1990; Bhatia, 1993), and also in conversation analysis, especially as it applies to the description of 'test genre', using as few as six subjects/interviews (Perrett, 1990) or as many as 20 (Lazaraton, 1992).

b Generation and description of categories and coding systems: Initially it is important to define the observable speech phenomena that could be said to constitute an interruption in perceived fluency. For the purposes of this description, six phenomena were isolated as potentially interrupting fluency from an analysis of the transcripts of 21 oral interviews, probably affecting the score given by an interviewer during an oral test. The categories of phenomena generated in this study do not differ from the literature on fluency which discusses surface aspects of performance which interrupt fluency (Grosjean, 1980; Hieke, 1985). The data themselves did not suggest any further phenomena which might have been investigated. The phenomena are as follows:

1) Fillers such as 'er(m)'.
2) The repetition of the first syllable of a word or a full word.
3) The negotiation of reference indicated by the reselection of referring devices.
4) The reselection of lexical items.
5) Anacolouthon.
6) Longer pauses of three seconds or more.

Category 6 needs special comment as this is an area related to the observation of fluency-related phenomena which has recently seen much controversy. Studies involving the analysis of pauses in the second language literature have mainly concerned themselves with teacher speech and the role of pauses in adapting to communication with non-native speakers/learners (Ellis, 1985: 145–46; Chaudron, 1988: 69–70), the suggestion being that speech rate is slower and the length of pauses greater in native speaker–non-native speaker communication, similar to that observed between adults and children (Gaies, 1977, as discussed in Allwright, 1988: 215), and that these phenomena help the comprehension of non-native speakers. Most of these studies do not use spectograph analysis, and the methodology used has come under severe criticism for lack of accuracy in measurement and drawing unsubstantiated conclusions from poor-quality data (Griffiths, 1991).

Here, we are concerned with the effect that the impression of hesitation phenomena exhibited by non-native speakers under test conditions, including pauses, have upon raters. It is acknowledged that Griffiths' criticism of second language research in the field of what he calls 'pausology' is equally applicable to the way in which pauses in category 6 of this data were observed and recorded, that is, in terms of the impression created for raters. However, in defence of the

analysis of this category, it must be said that the purpose of the analysis is somewhat different from that carried out in the currently available literature: it is not claimed that rate of speech and number of pauses are directly linked to comprehension, which is the major issue at stake. What is claimed is that the length and, perhaps, the nature of pauses in an oral test may be related to the impression given to the rater regarding the proficiency of the student and hence be related to the score which is awarded. For this purpose, precise measurement of pause length is advocated by Griffiths does not seem – at least for the present – to be a methodological consideration of primary importance.

However, these surface phenomena are of little use in themselves. Surface phenomena may be coded easily enough, but this method would not provide an *explanation* of the phenomena in terms of language use. When encountering one of the phenomena listed above, the researcher must attempt to develop a coding system which is interpretive. That is, to say why fluency appears to be disrupted by the occurrence of a particular phenomenon, or why it is not. For example, some longer pauses in speech may be interpreted by raters as indicating that communication has broken down, and the student would therefore be penalized on the rating scale. However, it is quite possible that the rater will consider certain pauses as being 'thinking time' in which the student is seriously considering the content of the next utterance, and thus be prepared to award a higher grade for 'natural' language behaviour (Meredith, 1978). There is, therefore, a need for a set of 'explanatory' categories for the coding of data from interviews. The explanatory categories are created iteratively in the interplay between data and interpretation.

The data suggested that observed interruptions in fluency would be accounted for by eight categories. These are as follows:

1) End-of-turn pauses: pauses indicating the end of a turn.
2) Content planning hesitation: pauses which appear to allow the student to plan the content of the next utterance.
3) Grammatical planning hesitation: pauses which appear to allow the student to plan the form of the next utterance.
4) Addition of examples, counterexamples or reasons to support a point of view: these pauses are used as an oral parenthesis before adding extra information to an argument or point of view, or break up a list of examples.
5) Expressing lexical uncertainty: pauses which mark searching for a word or expression.
6) Grammatical and/or lexical repair: hesitation phenomena which appear to be associated with self-correction.

7) Expressing propositional uncertainty: hesitation phenomena which appear to mark uncertainty in the views which are being expressed.
8) Misunderstanding or breakdown in communication.

Although the explanatory categories are high inference (in the terminology of Long, 1983), they do vary in the amount of interpretation which is necessary on the part of the researcher. Category one, for example, is fairly low inference in that it merely involves observing those pauses which occur at the end of turns where the propositional content of the utterance is coherent. Category two, on the other hand, is high inference in that one must attempt to decide when hesitation occurs because the learner is deciding what to say next, as opposed to how to say it, which is the definition of category three. This means that when an example of one of the six phenomena described as constituting aspects of fluency is observed, it must be decided into which of the explanatory categories to place it. The issue is one of data coding.

Any nontrivial descriptive system will suffer from such problems in the developmental stage. However, quantitative checks on the coding may lessen such concerns.

IV Results and discussion

1 Explanatory categories and data coding

The following is a description of the explanatory categories used in the qualitative study. Examples of each category are not provided, but a sample coded transcript is provided in Appendix 1. Although the reader does not have access to the tapes of the interview, or the first-hand experience of observation, it provides some indication of the methodology used.

a End-of-turn pauses: The most frequent use of extended pauses in the data was at the end of a student's turn. In the speech of less fluent speakers the students pause because they are not able (or willing) to continue speaking. The interviewer frequently does not begin his or her turn immediately in such cases, as he or she is waiting for the student to continue. The interviewer in the oral test appears to be highly sensitive to the possibility that the student needs time to plan what is going to be said next, and therefore the amount of overlapping speech may be much less than in less formal interaction, and the amount of silence between turns increases, something which would be highly embarrassing in informal talk (Sacks, Schegloff and Jefferson, 1974). This may be one aspect of oral testing situations which gives

rise to the frequent comments that the language produces is not as 'natural' as nontest language (see Fulcher, 1996).

When turning to the high-scoring students, it was discovered that the amount of this type of hesitation is somewhat different. For example, the number of times interlocutor contributions overlap increases, and pausing allows speakers to continue the conversation if they wish to; pausing is used as an effective turn-taking device, to initiate a change in speaker.

b Content planning hesitation: Hesitation seems to have a significant psycholinguistic role in that it allows the speaker time to formulate a plan for the next chunk of language. One aspect of this planning is the content of the utterance.

Students at intermediate and advanced levels often overtly indicate that content planning is going on. This was particularly evident in data generated by tasks which required the description of graphic information, flow charts or processes. Overt indication of planning takes the form of the student repeating a question or the previous utterance of the interviewer, almost as if the student were 'thinking out loud' or introspecting on his or her own thought processes. The following example is a particularly clear example of this:

> *Interviewer*: What reason can you think of for it happening in this way?
> *Student*: what reason (pause) I I (repetition) should say it must be the er. our er. con er (lexical search) contribution to . . .

c Grammatical forward planning: Grammatical forward planning occurs when the student may know what he or she wishes to say, but does not know how to say it. Thus, the planning stage needs time, and the execution may also contain other hesitations and structural reformulations, as in the following example, where A is the interviewer, and B the student:

> A> . . .kind of a job do you want:: B> when I chose this subject: I:: I chose this sub I chose this subject because I think I can . . .

In this example, it appears that the student wishes to express the view that the subject of study was chosen because of the ease with which (she goes on to say) it would be possible to get any kind of job. The purpose clause, beginning with 'because . . .' is therefore the most important part of the message. However, having begun with 'when' the student seems to find difficulty in actually formulating the message intended as it requires a purpose clause which cannot be used given the colligational restrictions which have been activated, unless the student can use the cohesive form 'I did so (because)'. Hence, the medium length pause after 'I' and the two reformulations. In the

process the use of 'when' is avoided, and the purpose clause is successfully introduced.

d Addition of examples, counterexamples or reasons to support a point of view: It was observed that a pause or a filled pause often precedes the addition of an example or reason when the student is presenting an argument, adding to or supporting what has already been said. In such cases, a pause acts as a verbal parenthesis.

Students of higher ability were noticed to be particularly adept at using pauses or fillers to introduce an additional element to their argument. In the speech of students of lower ability very few occurrences of giving examples, counterexamples or reasons were observed. When they did occur the most common use was that of adding to an utterance in order to give content to a general word which otherwise would have remained empty. Such general words (like 'problem' and 'solution') are often referred to as 'delexical' in that they do not of themselves contain any specific meaning, and have to be filled out with reference to context or other parts of the discourse (Winter, 1978).

e Expressing lexical uncertainty: When choice of a lexical item was a problem for a more able student there was a tendency for it to be overtly marked, while not causing any strain on the interaction. Often, this was associated with overt markers such as 'what do you call it' or 'I think that is how it is called'.

When a proficient student cannot recall a lexical item which is required in the discourse, it is frequently the case that the student can explain what is intended in another way. The use of circumlocution marks the student as a competent communicator. At times, the failure to retrieve a lexical item required and the inability of the student to circumlocute can lead to difficulty in retrieving the meaning of an utterance. Circumlocution should therefore be seen as a positive skill in language learning, which many speakers use to great effect every day.

f Grammatical or lexical repair: When analysing the data the separation of grammatical and lexical repair was found to be virtually impossible. The two appear to overlap to such an extent that to separate individual examples would have involved what was considered to be an undue amount of subjective decision, even for a qualitative interpretation. Grammatical and lexical repair were thus treated as a single category.

Although it was the case that a number of students could not repair utterances, which led to a rapid change of subject, many intermediate-ability students were easily able to monitor and repair their speech

in such a way that there was no strain on the interaction. Strategies typically included repeating an entire phrase or sentence with the correct form, often twice, before continuing, and anacoluthon. Anacoluthon is, in fact, a phenomenon which is observable in native speaker speech almost every time they engage in conversation whieh is not highly planned. More proficient students may have been correcting errors of which they were aware, but the effect was very 'natural'. It is only in oral tests that such phenomena tend to be penalized in the marking systems.

g Expressing propositional uncertainty: It was observed that students tended to hesitate when expressing uncertainty about a point of view or an argument which was being presented. This may be connected to the fact that the interviewer is the socially dominant partner in the interaction, as observed by Channell (1985). The student may attempt to avoid strong claims where points of view are presented as fact and seen as potentially challenging a dominant partner; this would not occur when the interviewer is offered the possibility of disagreeing.

The expression of propositional uncertainty often coincides with the use of lexical items such as: perhaps, maybe, think, believe, don't know or sometimes. This co-occurrence of the hesitation phenomena with the overt lexical items isolated gives to the utterance a sense of reservation on the part of the speaker which would admit the possibility of a challenge from the interlocutor.

h Misunderstanding or breakdown in communication: Examples of misunderstandings which needed to be clarified by the interviewer, or complete breakdown in communication, were only observed in the speech of lower-ability students.

i Variable interpretation of data: From the brief discussion of the categories above, it should have become clear that some of the categories do not reflect a linear/monotonic relationship between phenomena, interpretation and ability. This can be demonstrated in Figure 1, which shows the pattern of the uses of pausing as turn-taking device for students at each of five levels of oral proficiency. Students at level 1 are those who are least proficient, and simply 'dry up'. Here the pause is a plea for the interlocutor to take over the conversation again, before the propositional content of the student utterance is complete. As proficiency increases, the use of the pause for negotiating turn taking decreases, and again increases with the most able students, whose utterances are propositionally complete, and are indicating to their interlocutor that it is their turn to speak. Similarly, Figure 2

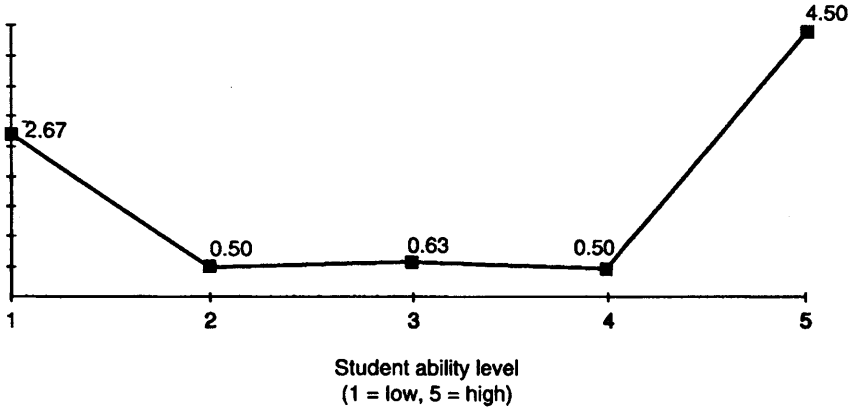

Figure 1 Pattern of the uses of pausing as a turn-taking device

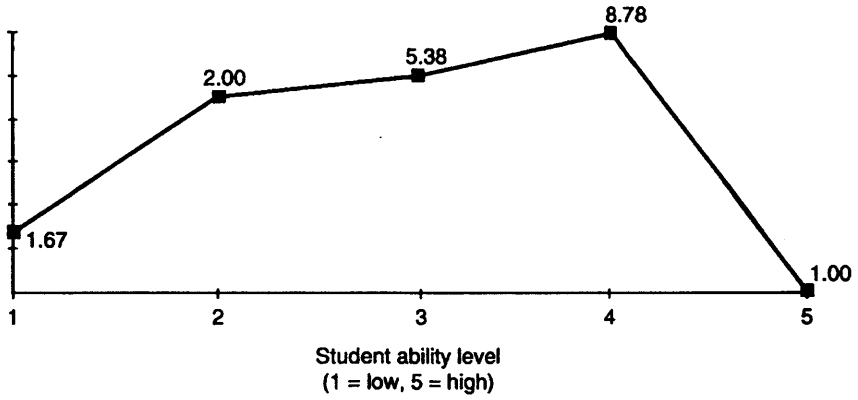

Figure 2 Pattern of the expression of proportional uncertainty

demonstrates that as proficiency increases, students express propositional uncertainty more frequently. It may be hypothesized that this increases with the development of sociolinguistic competence, but then dramatically decreases at the higher-ability level, perhaps because students of great proficiency are much more confident, even in oral interviews. Obviously, any such hypothesizing must be tentative explanations of these curves, until adequate research can be designed which can provide a causal explanation. That is, the explanatory categories are operating at a level of delicacy which is not deep enough to be able to account for developing language processing capacity in students. However, this research quickly highlighted the fact that oral rating scales cannot be linear/monotonic in the way a priori rating scales assume they are. More errors may occur in the

speech of more able students because they are in the process of acquiring more language, and experimenting with it. Similar results were discovered by Meisel (1980) who noticed increasing 'wrong use' of forms as earlier systems were extended in the direction of the second language system, and called this transitory interlanguage 'elaborative simplification' in contrast to 'restrictive simplification'. (The number printed on Figures 1 and 2, respectively, refers to the mean number of pauses or expression of propositional uncertainty respectively, used by students at the ability level marked on the *X* axis, in approximately 15 minutes of interview. The number of pauses or expression of propositional uncertainty was corrected for the amount of speech produced, to maintain comparability of data across interviews.)

2 *Discriminant analysis*

As a check on the way in which the researcher coded the data into categories, and to create a fluency rating scale in which the band descriptors are generated by the data, it was necessary to make tallies of observations from the speech of students coded into each of the eight explanatory categories described above. Discriminant analysis allows the researcher to investigate 1) the extent to which all categories together discriminate between the students; and 2) the extent to which students would have been reliably placed in the bands/levels which they actually received on an Oral Proficiency Interview, had they been rated *only* on the categories developed to describe fluency phenomena (Crocker and Algina, 1986; 256–63). Clearly, an assumption is being made in the use of such a technique in the development of a scale: that the students were at least rank ordered appropriately in terms of ability by the Oral Proficiency Interview. For the moment, it will be assumed that this was indeed the case.

It was also assumed that if it proved possible to discriminate accurately between students using only the coded data, it should also prove possible to develop a rating scale on the basis of the data which could then be crossvalidated in use on other groups of students. That is, if the categories are found to discriminate between students placed in certain bands, they may then be used to create verbal descriptors for new bands by returning to the description which generated the categories. The new descriptors thus contain definitions which are directly linked to actual L2 production in test conditions. It is this process which links test development to applied linguistics concerns, and avoids the problems associated with flawed *post hoc* validation studies.

The multivariate results are presented in Table 1. This shows the

Table 1 Discriminant analysis results for prediction from eight explanatory categories of fluency to Oral Proficiency Interview scores

Categorial variable	Band scores
Wilk's lambda	**0.02**
F	**2.06**
df	32, 34
p	0.02

First discriminant function: $\chi^2 = 52.903$, $p = .00$.
Second discriminant function: $\chi^2 = 27.704$, $p = .14$.

degree to which all categories taken together are capable of discriminating between the students. Wilk's lambda is the statistic produced by the test of the multivariate hypothesis that subjects can be divided into groups from observations (Wilkinson, 1988: 538). The result of the multivariate analysis is significant at $p = .02$, indicating that when taken together the categories discriminate well between students.

Finally, by using discriminant analysis it is possible to analyse the relationship between the band score actually awarded to each of the students and the band score which would be predicted on the basis of the categories. The results of such prediction are presented in Table 2.

In Table 2, we have the bands which were actually awarded on an Oral Proficiency Interview, and these are compared with the bands which would have been awarded if they had been awarded on the basis of the explanatory categories in the operational description of fluency used in the discriminant analysis (Wilkinson, 1988: 589). Thus, for example, eight students were actually awarded a band 3, whereas if these eight students had been awarded their score on the basis of the significantly discriminating explanatory categories only seven would have been awarded a band 3 and one would have been awarded a band 4.

Table 2 The relationship between actual band scores and predicted band scores

Band awarded	Predicted band					Total
	1	**2**	**3**	**4**	**5**	
1	***3***	**0**	**0**	**0**	**0**	3
2	**0**	***4***	**0**	**0**	**0**	4
3	**0**	**0**	***7***	**1**	**0**	8
4	**0**	**0**	**0**	***4***	**0**	4
5	**0**	**0**	**0**	**0**	***2***	2
Total	3	4	7	5	2	21

It may be seen from Table 2 that only one candidate would have been given a different band score. It is not possible to investigate which components of the definition of fluency contribute to the overall discrimination reported, as substantially more data would be required before reliable univariate statistics could be produced. However, as prediction appears from this data to be very accurate, this may be taken as an indication that the data coding was accurate enough to allow reasonable prediction from data to student ability, as predicted by an ELTS band/level. If this is the case, then such a data-driven approach to rating scale development may be worth pursuing.

3 Additional components of fluency

From an analysis of the transcripts, it also appeared that in the speech of higher-ability students there was significant evidence of back-channelling (students using utterances such as 'hmm' or 'yeah' when listening to the interviewer). It was hypothesized that the number of back-channels was also related to the perception of fluency, and would also discriminate well between lower- and higher-ability students. For the purposes of investigating back-channelling the original bands were collapsed to form just three, on the basis of mean counts of back-channels in student speech, corrected for amount of language produced. These counts were then also submitted to discriminant analysis. The results (not presented here) were significant, and so information on back-channelling was included in the final fluency band descriptors in Appendix 2.

V Producing fluency descriptors

It was initially hypothesized that the results of the discriminant analysis could be further investigated using a *post hoc* Tukey HSD test to investigate where the eight explanatory categories distinguished between students achieving different band scores. However, this was not possible because the Tukey HSD test divides the sample of 21 students into such small groupings that extremely large differences in means would have to be recorded for them to register as significant. However, from the discriminant analysis it is known that differences in means are significant, otherwise the results of the multivariate tests would not have been significant.

The process of producing scale descriptors was attempted in the following way. First, the means of each category score for the students in a particular band may be calculated and plotted, as in Figures 1 and 2. Secondly, we return to the characteristics of the discourse which were observed to create the eight explanatory categories. It is

these explanatory categories and their definitions which were produced by the analysis which can now be used to produce definitions for bands or groups of bands which will most effectively discriminate between students.

The fluency rating scale contains five bands which were generated by the data, labelled bands 1–5. Two further bands were attached to the scale: band 0 was merely described a 'less fluent' than band 1, while band 6 was described as 'more fluent' than band 5. The purpose of this was twofold. First, it is not being claimed that the rating scale taps the entire range of 'potential' fluency. That is, bands 1 and 5 are not seen as absolute extremes. It was merely not possible to describe anything less or more fluent from the data, and this avoids the further problem of suggesting that there is such a thing as zero and perfect 'proficiency' in fluency. Secondly, it was hypothesized that raters would tend to avoid the highest and lowest bands on the scale, and thus the inclusion of two additional bands would mean that there would be a greater chance of raters using the full range of bands which can be described.

The fluency rating scale is given in full in Appendix 2, and an explanation of the relationship of the descriptors to the eight explanatory categories is contained in [square brackets]. It will also be noted that some parts of the descriptors are contained in (round brackets). A discussion of these elements of the rating scale is not contained in this article, but a full discussion may be found in Fulcher (1993).

VI An assessment of the fluency rating scale

The fluency rating scale which has been described here may well avoid two problems. The first of these is the problem of falling into the trap of failing to define abilities in enough detail to make it impossible to investigate the validity of the scale. Secondly, the descriptors are specific enough to be relatable to actual language performance if we assume, for it must at the moment be an assumption, that the Grounded Theory approach to the analysis of the original data used in this study to develop eight explanatory categories is a useful model for test development with an applied linguistics basis.

The data-driven approach to scale development is therefore different from traditional approaches in the FSI and Interagency Language Roundtable (ILR) mould, and also different from the 'absolute' scales which researchers are currently investigating (Bachman and Clark, 1988; Bachman, 1990). The degree to which such a rating scale is successful can only be evaluated through trialling and reliability and validation studies.

1 Reliability

Reliability of the use of the rating scale was estimated following procedures outlined in Bolus, Hinofotis and Bailey (1981), Cronbach (1984: 161–63), Crocker and Algina (1986: 157–85), Feldt and Brennan (1989: 128–36) and Bachman (1990). Using five raters and three tasks, a reliability coefficient of .9 was recorded, with an inter-rater generalizability coefficient of .93, and an equivalent forms generalizability coefficient of .98.

2 Validity by group differences

Although validity by group differences is a 'weak' form of validity (Messick, 1989: 55), the principle behind its use is simple: if the test-taking sample can be divided into ability groups in advance of taking the test (by teachers or other tests) then the test results may be described as valid if they do discriminate between the various groups. On the first use of the fluency rating scale, students were classified as 'high', 'average', and 'low' ability students by their teachers. Table 3 presents the ANOVA results for this analysis, and it can easily be seen that the between-groups variance is larger than the within-groups variance. It would appear that the fluency rating scale is capable of distinguishing between groups.

Table 3 Validity by group differences

ANOVA

Source	SS	df	MS	F	*p*
Between groups	136.24	2	68.12	34.26	.00
Within groups	83.50	42	1.99		

	Absolute mean differences			Probabilities	
	Good language learners	Average language learners		Good language learners	Average language learners
Average language learners	2.67		Average language learners	.00	
Poor language learners	4.34	1.67	Poor language learners	.00	.00

3 Rasch partial credit analysis

Whether a rating scale is operating as a measurement instrument can be determined by the extent to which rating scale calibrations represent 'a regular progression in scaled proficiency from lower to higher levels' (Henning, 1992: 368). Table 4 provides the delta statistic for each of the bands on the rating scale, on each of three tasks. This is the difficulty of scoring the particular band on the rating scale in logits.

From Table 4, we may see that not only does the fluency rating scale meet the requirements set out by Henning for a measurement instrument but the difficulty estimates of achieving the level on the scale is also very stable across task types. This is a second requirement for a measurement instrument which Henning does not mention, namely, that tests *in which scores need to be generalizable* should yield similar results under different data-collection circumstances. In terms of rating scale development, we may term this the principle of 'coherence' in measurement: that the operationalization of the construct in a rating scale, in this case fluency, operates in a stable fashion both synchronically and diachronically. The rating scale described in this study would appear to be coherent, according to this definition.

VII Conclusion

Further studies are urgently required into the study of constructs such as fluency (Fulcher, 1994). Evidence has frequently been presented to suggest that what we traditionally call 'skills' (e.g., reading vs. speaking) are distinctive traits (Bachman and Palmer, 1982). However, there is little evidence available upon which we are able to judge the more specific constructs which are claimed to make up the skill of speaking, or the skill of reading. Establishing discriminant validity using multitrait–multimethod studies (Campbell and Fiske, 1959) and/or maximum likelihood studies (Joreskog, 1969), requires extremely stringent criteria which are difficult to meet, but such work must be undertaken.

Table 4 Delta for the bands of the fluency rating scale on three tasks

Task	Band on the fluency rating scale				
	2	3	4	5	6
1	–3.96	–1.72	0.24	1.10	4.34
2	–3.99	–1.46	0.11	1.73	3.60
3	–3.88	–1.73	0.07	1.19	3.64

Unfortunately, it appears to be the case that many testing instruments do not contain a rigorous applied linguistics base, whether the underpinning be theoretical or empirical. The results of validation studies are, therefore, often trivial. For example, in one study of the relative contribution of constructs to the scores of students at various levels of the FSI rating scale (Adams, 1980), there is no principled way of deciding why the constructs are jumbled up by level, apparently randomly, in the way they appear to be. Until test researchers and developers take seriously the validity of tests at the development phase rather than as a *post hoc* notion, the problem of the indeterminacy of validation studies and the uninterpretability of test scores will remain serious. Research in progress, such as that of North (1994a; 1994b) into use of various rating scales, including the one described in this article, should throw more light on the utility of scales produced by different construction methods. Research is also currently underway to examine the use of this rating scale on a sample of students drawn from a different population from for which it was constructed and on which it was validated.

However, it is tentatively suggested that a data-based approach to rating scale development appears to be promising, and that further research should be carried out into the description and operationalization of constructs for language testing, reinforcing the necessary link between applied linguistics, second language acquisition research and language testing theory and practice.

VIII References

Adams, M.L. 1980: Five cooccuring factors in speaking proficiency. In Frith, J.R., editor, *Measuring spoken language proficiency*. Washington, DC: Georgetown University Press, 1–6.

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

Allwright, D. 1988: *Observation in the language classroom*. London: Longman.

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

Bachman, L.F. and **Clark, J.L.D.** 1988: The measurement of foreign/second language proficiency. *Annals of the American Academy of Political and Social Science* 490, 20–33.

Bachman, L.F. and **Palmer, A.S.** 1982: The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–65.

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.

Bhatia, V.K. 1993: *Analysing genre: language use in professional settings.* London: Longman.

Bialystok, E. and **Sharwood Smith, M.** 1985: Interlanguage is not a state of mind: an evaluation of the construct for second language acquisition. *Applied Linguistics* 6, 101–17.

Boldt, R.F. 1991: Second language constructs as indexed by ACTFL ratings and TOEFL scores. Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Bolus, R.E., Hinofotis, F. and **Bailey, K.M.** 1981: An introduction to generalizability theory in second language research. *Language Learning* 32, 245–58.

Brown, G. and **Yule, G.** 1983: *Discourse analysis.* Cambridge: Cambridge University Press.

Brumfit, C. 1984: *Communicative methodology in language teaching: the roles of fluency and accuracy.* Cambridge: Cambridge University Press.

Campbell, D.T. and **Fiske, D.W.** 1959: Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin* 56, 81–105.

Channell, J. 1985: Vagueness as a conversational strategy. *Nottingham Linguistic Circular* 14, 3–24.

Chaudron, C. 1988: *Second language classrooms: research on teaching and learning.* Cambridge: Cambridge University Press.

Clahsen, H. 1985: Profiling second language development: a procedure for assessing L2 proficiency. In Hyltenstam, K. and Pienemann, M., editors, *Modeling and assessing second language acquisition,* San Diego, CA: College Hill Press.

Clark, J.L.D. and **Clifford, R.T.** 1988: The FSI/ILR/ACTFL proficiency scales and testing techniques: development, current status and needed research. *Studies in Second Language Acquisition* 10, 129–47.

Crocker, L. and **Algina, J.** 1986: *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Cronbach, L.J. 1984: *Essentials of psychological testing.* New York: Harper & Row.

Crystal, D. and **Davy, D.** 1985: *Advanced conversational English.* London: Longman.

Davies, A. 1987: Certificate of Proficiency in English. In Alderson, J.C., Krahnke, K. and Stansfield, C.W., editors, *Reviews of English language proficiency tests,* Washington, DC: TESOL.

Douglas, D. 1994: Quantity and quality in speaking test performance. *Language Testing* 11, 125–44.

Dulay, H., Burt, M. and **Krashen, S.** 1982: *Language two.* Oxford: Oxford University Press.

Ellis, R. 1985: *Understanding second language acquisition.* Oxford: Oxford University Press.

—— 1990: Individual learning styles in classroom second language development. In de Jong, J.H.A.L. and Stevenson, K., editors, *Individualizing*

the assessment of language abilities, Philadelphia, PA: Multilingual Matters.

Feldt, L.S. and **Brennan, R.L.** 1989: Reliability. In Linn, R.L., editor, *Educational measurement*, New York: American Council on Education, Macmillan.

Frith, J.R. 1979: Preface. In Adams, M.L. and Frith, J.R., editors, *Testing kit: French and Spanish*. Washington, DC: Department of State Foreign Services Institute.

Fulcher, G. 1987: Tests of oral performance: the need for data-based criteria. *English Language Teaching Journal* 41, 287–91.

—— 1988: *Lexis and reality in oral testing*.Washington, DC: ERIC Clearinghouse on Languages and Linguistics (document reference: ED 298 759).

—— 1993: The construction and validation of rating scales for oral tests in English as a foreign language. Unpublished PhD, University of Lancaster.

—— 1994: Some priority areas for research in oral language testing. *Language Testing Update* 15, 39–47.

—— 1996: Testing tasks: issues in task design and the group oral. *Language Testing* 13, 23–51.

Gaies, S.J. 1977: The nature of linguistic input in formal second language learning: linguistic and communicative strategies in ESL teachers' classroom language. In Brown, H.D., Yorio, C.A. and Crymes, R.H., editors, *Teaching and learning English as a second language: trends in research and practice. Selected papers from the 1977 TESOL convention*, Washington, DC: TESOL.

Geertz, C. 1973: Thick description: toward an interpretative theory of culture. In Geertz, C. *The interpretation of cultures: selected essays*, New York: Basic Books.

Glaser, B. and **Strauss, A.L.** 1967: *The discovery of Grounded Theory: strategies for qualitative research*. Chicago; IL: Aldine.

Griffin, P.E. 1985: The use of latent trait models in the calibration of tests of spoken language in large-scale selection-placement programs. In Lee, Y.P., editor, *New directions in language testing*, Oxford: Pergamon Institute of English.

Griffiths, R. 1991: Pausological research in an L2 context: a rationale, and review of selected studies. *Applied Linguistics* 12, 345–64.

Grosjean, F. 1980: Temporal variables within and between languages. In Dechert, H. and Raupach, M., editors, *Towards a cross-linguistic assessment of speech production*, Frankfurt: Peter Lang.

Hamp-Lyons, L. 1987: Cambridge First Certificate in English. In Alderson, J.C., Krahnke, K. and Stansfield, C.W., editors, *Reviews of English language proficiency tests*, Washington, DC: TESOL.

Hatch, E. 1974: Second language universals. *Working Papers on Bilingualism* 3, 1–17.

Hedge, T. 1993: Key concepts in ELT: fluency. *English Language Teaching Journal* 47, 275–77.

Henning, G. 1992: The ACTFL Oral Proficiency Interview: validity evidence. *System* 20, 365–72.

Hieke, A.E. 1985: A componential approach to oral fluency evaluation. *Modern Language Journal* 69, 135–42.

Ingram, D. 1982: Introduction to the ASLPR. Mimeo, Brisbane College of Advanced Education.

Jarvis, G.A. 1986: Proficiency testing: a matter of false hopes? *ADFL Bulletin* 18, 20–21.

Joreskog, K.G. 1969: A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34, 183–202.

Klippel, F. 1984: *Keep talking: communicative fluency activities for language teaching*. Cambridge: Cambridge University Press.

Larsen-Freeman, D. and **Long, M.** 1991: *An introduction to second language acquisition research*. London: Longman.

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

Long, M. 1983: Inside the 'black box': methodological issues in classroom research on language learning. In Seliger, H.W. and Long, M., editors, *Classroom oriented research in second language acquisition*, Rowley, MA: Newbury House.

Manning, P.K. 1991: Analytic induction. In Plumber, K., editor, *Symbolic Interactionism. Vol. 2. Contemporary issues*. Brookfield, VT: Edward Elgar, 401–30.

Meisel, J.M. 1980: Linguistic simplification. In Felix, S., editor, *Second language development: trends and issues*, Tubingen: Gunter Narr.

Meredith, R.A. 1978: Improved oral test scores through delayed response. *Modern Language Journal* 62, 321–27.

Messick, S. 1989: Validity. In Linn, L., editor, *Educational measurement*. New York: American Council on Education/Macmillan.

North, B. 1994a: *Scales of language proficiency: a survey of some existing systems*. Washington, DC: Georgetown University Press.

—— 1994b: Item and descriptor-banking: objective underpinning of subjective assessment in schools. Paper presented at the Language Testing Forum, University of Cambridge Local Examinations Syndicate, 11 December.

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

Rixon, S. 1992: English and other languages for younger children: practice and theory in a rapidly changing world. *Language Teaching* 25, 73–93.

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

Seliger, H.W. 1980: Utterance planning and correction behaviour: its function in the grammar construction process for second language learners. In Dechert, H. and Raupach, M., editors. *Towards a cross-linguistic assessment of speech production*, Frankfurt: Peter Lang.

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

Sollenberger, H.E. 1978: Development and current use of the FSI Oral Interview Test. In Clark, J.L.D., editor, *Direct testing of speaking proficiency: theory and application*, Princeton, NJ: Educational Testing Service.

Strauss, A. and **Corbin, J.** 1994: Grounded theory methodology: an overview. In Denzin, N.K. and Lincoln, Y.S., editors, *Handbook of qualitative research*, London: Sage.

Swales, J. 1990: *Genre analysis*. Cambridge: Cambridge University Press.

Upshur, J.A. and **Turner, C.E.** 1995: Constructing rating scales for second language tests. *English Language Teaching Journal* 49, 3–12.

Weir, C. 1987: English Language Testing Service. In Alderson, J.C., Krahnke, K. and Stansfield, C.W., editors, *Reviews of English language proficiency tests*, Washington, DC: TESOL.

Wilkinson, A. 1988: *Systat: the system for statistics*. Evanston, IL: Systat Inc.

Winter, E.O. 1978: A look at the role of certain words in information structure. In Jones, K.P. and Horsnell, V., editors, *Informatics 3*, London: ASLIB.

Appendix 1 Sample coded transcript

In the following transcript, the categories of fluency are given between slashes. The notation '/3/' would therefore indicate grammatical planning hesitation. 'A>' indicates the speech of the interviewer and 'B>' the speech of the student. Utterances enclosed by square brackets indicate speech overlap, while round brackets indicate noises, the removal of identifying information or sections of inaudible speech. The use of a colons represents pauses. One colon represents a short (unobtrusive) pauses, while three colons represents a lengthy (obtrusive) pause. When reading the following transcript it must be remembered that the researcher was present at the oral interview, and when coding the transcript was working with audiorecordings of the interview data. The student concerned was awarded a band 7 on the ELTS Oral Interview rating scale.

A> Hello::I'm (NAME) B> I'm (NAME) not (NAME) A> (NAME) [yes B> yeah] A> that's what that's what I've got that's [what B>ah] A> that's what I said [the B> but] they called me (NAME) A> ah yeah they it was the the person who I asked him to call you but er never mind: anyway I understand you're a pharmacist B> yeah A> is that right tell me something about your work B> ah: erm /2/ I work at I don't know if you know (COUNTRY) at all A> yes I do B> I work at (NAME) hospital A> hm hm B> and I

work as a pharmacist as a hospital pharmacist we are three pharmacists there we actually erm /2/ dispense medicines and er: /4/ of course we have other things like preparing: some kinds of medicines the (inaudible) A> yeah B> sceptic solution: A> hm: so it's generally dealing with drugs in the hospital that B> yeah and of course we supply the medicines to most other wards and casualty and we're actually responsible: A> yeah so you have to check to make sure that none of them go missing B> yeah but their expiry dates erm: /4/ if they're kept well if they're kept properly A> yeah it sounds very interesting is (NAME) a big hospital B> no actually it's a small one A> hm: right but you you live out there in (NAME) B> I have to (laughter) A> yeah B> because I am er: I am /2/ on call every other day A> yeah B> er but I /4/ come from (NAME) A> hm so you would you like to transfer to (NAME) General if [you B> yeah] A> were given half the [chance B> for sure] A> or (NAME) probably even nicer B> yes [(laughter)] A> yeah because (NAME) is a modern hospital whereas (NAME) is beginning to fall to bits B> yes A> a little isn't it B> one of the most er: /3/ it's the most A> yeah okay B> but I'm interesting in drug quality control A> hm: okay well I'll ask you about that in a few minutes B> ah A> but you chose to do the (NAME) the (NAME) version this morning B> (laughter) yes A> so I'm going to ask you to have another look at this could you turn to page one it's the very first page in there B> first page yeah A> yeah you remember reading this this morning very [quickly B> yes] A> I guess B> very very quickly [(laughter)] A> yeah B> that was the trick A> okay well this was about plutonium in the body: and I'm going to have to ask you about this diagram here er again: in the bottom left hand side left hand corner down here there's a box with an X in it: my first question is can you identify what X [is B> bones] A> the bones B> yeah A> okay can you tell me how the bones how how it how it's related to the rest of the diagram I mean what happens to plutonium and how it gets into the bones and so on B> ah: er /2/ plutonium enters the body through inhalation: okay A> hm hm B> I start from the beginning [(laughter) A > Okay] B> erm /2/ then it goes to the digestive system A> hm hm B> and to the (inaudible) lungs some in the lungs :: /2/ from the lungs it goes er /2/ through the reticuloendothelial system A> hm which is: [what B> erm] /2/ A> does that what does it mean that word B> well ah it's a system that er /2/ has to do with erm: /5/ fibrocytosis: A> no you're going to have to explain that for me B> erm: consisted for from er: /2/ cells A> hm hm B> that actually are responsible for er: /6/ to take: er foreign material A> hm hm B> out of the body erm /4/ and this is done: er /2/ when they put into the circulation A> okay right B> (laughter) A>

I think I've got that B> yeah: and er /2/ from that system goes to the blood A> hm hm B> and from the circulation: er /2/ to bones kidneys and the liver A> yes okay how's erm: you said that the plutonium was taken into the body through inhalation B> yeah A> yeah: does that mean to say that plutonium is in the air:: [I mean we breath it in B> there is some] there: yes plutonium exists there in very small quantities very small quantities A> but not enough to damage us: B> oh no I don't think so A> hm B> except if you live near or next to nuclear reactors where plutonium (laughter) A> yes but we don't have any nuclear reactors in (NAME) B> no A> so this is not a problem that we would come up in: would you you we would find this in (NAME) at all B> no no A> if erm let's imagine that you did come across a case like this B> yes A> what kind of drug would you treat it with would you treat it with drugs at all: B> drugs A> yes B> I mean do you mean plutonium poisoning A> yeah:: B> the symptoms yes: A> hm B> but not actually not the er: /4/ the real disease A> hm what what would the symptoms be B> oh: it might be symptoms of er: radiation exposure like diarrhoea vomiting A> hm B> headaches and er: /4/ ataxia erm:: /4/ convulsions sweating A> it sounds pretty horrible B> yeah it is (laughter) A> okay well I'll leave that now I I won't I won't press you any further on plutonium poisoning the course you're going to going to you want to take in England I I presume B> yeah A> yeah what you what is it B> it's a course on I'm not sure because er /7/ it says United Kingdom or or Denmark Denmark /7/ I think A> sorry I I didn't B> er it was for the United Kingdom or Denmark A> or Denmark B> yeah A> so you don't know which one you're going to B> no A> do you speak Danish B> no (laughter) B> probably the United Kingdom well unless of course they teach [it B> yeah] A> in En in En in English in Denmark but but you don't know where you're going to do it B> no not [yet A> no] hm how about the the course itself quality control: drugs B> yeah A> well what does it involve B> erm: ah /2/ first we've to control drugs A> hm B> ah you want me to say what we mean by quality control [something A> yeah] B> like that A> please B> we mean ah: we mean /2/ (inaudible) and testers tests which are used to determine the density purity potency quantity and of course quality of pharmaceutical preparations A> okay I I thought that most pharmaceutical preparations they were produced by companies B> yeah A> and then they were sold I mean does this not mean to say that: say you were going to buy a drug from an English company or a company from the States: but when it first came to (NAME) you you would have to check it first: before you actually used it in (NAME) [or are B> or yeah] A> you checking the quality of [drugs produced B>

for the first time] A> in (NAME) B> yes A> [I mean erm: B> (inaudible)] A> quality of what I mean [to say (inaudible) B> I mean all] pharmaceutical companies have their own: er /2/ quality control departments A> yeah B> they check their products for their quality but er all products registered and sold in the (NAME) market: /2/ they are first checked from their quality and then registered here A> right B> yeah A> hm B> and of course: /4/ I mean the government (inaudible): er er /2/ has two pharma two: erm /5/ labs two laboratories the pharmaceutical laboratories and the chemical state laboratory A> A hm B> in the pharmaceutical laboratory are checked are checked er /6/ products for (inaudible) A> hm B> you see and: erm also on erm: /4/ medicines bought by the government for the use of to be used in /6/ the hospitals A> hm it sounds like very interesting work B> it is A> quite [fascinating B> yeah] A> erm B> has to do with chemical reactions A> yeah hm: yes it it's fascinating I think it's much more it's much more interesting than being involved in this side of hospital work than perhaps er you know the medical practitioners going around and seeing patients and things I mean I would find this more [fascinating B> yes it's] more fascinating A> yeah: but are you B> it's fascinating with chemical reactions er: /1/ A> yeah B> erm: /4/ has to do with photometric methods [and A> hm] B> (inaudible) A> yeah: okay well I don't think I need to ask you anything else B> yeah A> unless you have anything to ask me B> (laughter) no

---

Appendix 2 The fluency rating scale

Band 0

Candidates in band 0 do not reach the required standard to be placed in band 1.

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 [categories 1 and 8]. (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 [category 2]. However, hesitation is frequently evident when the candidate has to plan the utterance grammatically [category 3]. 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 [categories 1 and 8]. (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, counterexamples or reasons, to support a view expressed [category 4].

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) [category 6].

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 [category 5]. (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 [category 8]. 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 reap the question or clarify what he or she wishes the candidate to do [category 8]. 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 [categories 3 and 6]. 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 circumlocute even if this is not very successful [categories 5 and 6]. 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 [category 4].

Band 3

A candidate in band 2 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 [category 8].

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 [category 2]. Time is sometimes needed to plan a sentence grammatically in advance, especially after making an error which the candidate then rephrases [category 3].

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 he or she considers to be better in the context. The candidate may even question the interviewer overtly regarding the appropriacy of the word which has been chosen [category 5].

Often candidates in this band will give examples, counterexamples or reasons to support their point of view [category 4].

At band 3 and above there is an increasing tendency for candidates to use 'back-channelling' – the use of 'hm' or 'yeah' – when the interviewer is talking, giving the interview a greater sense of normal conversation, although many better candidates still do not use this device.)

Band 4

A band 4 candidate will only very rarely misunderstand a question of the interviewer, fail to respond or dry up in the middle of an utterance [categories 1 and 8].

A candidate in this band will exhibit a much greater tendency than candidates in any other band to express doubt about what he or she is saying. They will often use words such as 'maybe' and 'perhaps' when presenting their own point of view or opinion [category 7]. More often than not, they will back up their opinion with examples or provide reasons for holding a certain belief [category 4]. They will pause frequently to consider exactly how to express the content of what they wish to say and how they will present their views [category 2]. (They will only rarely respond with a single word unless asked a polar question by the interviewer.)

There will be far fewer pauses to consider the grammatical structure of an utterance [category 3] and pausing to consider the appropriacy of a lexical item chosen is rare [category 5]. A candidate in this band will reformulate a sentence from time to time if it is considered to be inaccurate or the grammar does not allow the candidate to complete the proposition which he or she wishes to express [category 6].

Band 5

A candidate at band 5 almost never misunderstands the interviewer, fails to respond or dries up when speaking [categories 1 and 8]. The majority of pauses or hesitations which occur will be when the candidate is considering how to express a point of view or opinion

[category 2], or how to support a point of view or opinion by providing appropriate examples or reasons [category 4]. However, a candidate at band 5 will not express uncertainty regarding these views or opinions as frequently as a candidate at band 4, and so there are fewer hesitations when introducing new propositions [category 7].

Very rarely does a band 5 candidate have to pause to consider the grammatical structure of an utterance [category 3] and almost never hesitates regarding choice of lexis [category 5]. Band 5 candidates demonstrate a confidence in their ability to get things right the first time. While they do sometimes pause to reformulate sentences this is always because they cannot put across the propositional content of their utterance without changing grammatical form [category 6].

It may be noticed by the interviewer that the candidate responds to questions and prompts so quickly and efficiently that the next question or prompt has not been prepared, resulting in a pause in the interview while the interviewer plans his or her next utterance [category 1].

Band 6

Candidates in band 6 reach a standard higher than that described in band 5.