Tests of oral performance: the need for data-based criteria

Glenn Fulcher

It has become almost axiomatic for 'communicative' testing theory that the tests should contain exercises which are based on 'real-life' communicative situations (Morrow 1979). In essence, this makes the communicative testing enterprise an exercise in content validity. The notion of 'sampling real life' has, however, been extensively criticized (Oller 1979:184; Alderson 1981:57), mainly because testing has been seen as the reliable prediction of success in some behavioural performance in a non-test situation. Communicative testing theory tries to tap the performance directly, and here lies the problem.

This article investigates one aspect of this problem with regard to the assessment of the English Language Testing Service,[1] and attempts to uncover what appears to be a problem in the theory on which such tests are constructed—or, to put it another way, a problem in construct validity.

The basis of the oral interview

The majority of communicative tests in current use, like the English Language Testing Service (ELTS), were derived in part from J. L. Munby (1978), and the specifications follow his taxonomy of skills closely (Carroll 1981:78–9). This has led the tester to come to a compromise between the 'constructive interplay with unpredictable stimuli' and 'scientific measurement' by trying on the one hand to mirror real life in the oral interview, while on the other providing the interviewer with assessment scales based upon the taxonomy of skills to be rated (Carroll 1980:53–5).

It is in the oral interview that communicative testers therefore feel they have had their greatest success. 'Nowhere is the contrast between the old and the new approaches to language testing clearer than in the testing of speaking—or, more precisely, of oral interaction' (Carroll 1985:49). The 'interview' can either be of the traditional one-to-one variety, or involve the preferred group situation. Activities may involve role play, the discussion of some graphic material given to the testees prior to the test, a discussion of some burning issue chaired by a member of the group, or some problem-solving activity (Carroll 1985:50–56).

Problems inherent in this approach, such as personality factors, will be overlooked here. We are primarily interested in the examiner's approach to the role of fluency in the rating procedure.

The notion of assessment of fluency

In ELTS, the Interview Assessment Scale is a nine-band marking system, 9 representing an expert speaker and 1 a non-speaker (Carroll 1980:135). It is rare for the extreme ends of the scale to be used, because even though it is criterion-referenced, where the criteria have been established in relation to norms for certain types of speaker in certain educational courses, scores

tend to bunch up at band 5. A testee with such a mark may be said to be an 'average' speaker, and this is to be defined by the categories directly above and below him or her.

What do the bands describe as increased fluency? At Band 6 the guidelines are: 'Stumbles and hesitates at times but is reasonably fluent otherwise.' So we may assume that at band 5 these things happen more often than 'at times', and that fluency involves not stumbling or hesitating. At Band 7 the reference point is: 'Some hesitation and repetition due to a measure of language restriction but interacts effectively.' The question here is how the examiner could know that lack of fluency on any occasion was due to language restriction—which presumably means not having an appropriate word or expression on hand when needed. It is interesting to note that in band 6 the speaker is 'able to maintain theme of dialogue' while at band 7 he or she 'presents case clearly and logically'. We can conclude that Carroll believes increased fluency to go hand in hand with an increase in ability to deliver propositional content clearly and precisely.

Below Band 5, in Band 4, we learn that the testee 'lacks fluency and probably accuracy in speaking'. Lack of fluency, stumbling, and hesitating seem to be allied to grammatical inaccuracy, and as the testee also 'gives the impression that he (*sic*) is in touch with the dialogue' but isn't, there is also a drop in ability to handle propositional content development. At Band 3, the 'extremely limited speaker', we are told that 'dialogue is a drawn-out affair punctuated with hesitations and misunderstandings'.

So the factors involved in fluency and their connection with the assessment scale could be summarized as shown in Table 1.

Factors	*Higher Band*	*Lower Band*
Repetition	Low	High
Hesitation	Low	High
Stumbling	Low	High
Propositional development	High	Low
Grammatical accuracy	High	Low

Table 1

So far, so good. But a communicative oral test claims to be 'communicative'—i.e. to reflect 'real-life' communication. How far is this picture in the assessment scale 'real'?

Some evidence

The only way to answer the question is to collect large data banks of recorded material, transcribed and analysed. The following extract against which to measure the above criteria is taken from a (regrettably small) personal collection of recorded data. The topic under discussion was the relative difficulties of learning certain languages, and the recording was made without the knowledge of the participants, who were informed of the presence of the recorder after the discussion had ended.

In the discussion, speaker **A** is a teacher of business studies; speaker **B** is a lecturer in Biology; speaker **C** is the only non-native speaker in the group, but was educated in England (Cambridge) and is currently a TEFL lecturer overseas; speaker **D** is a medical practitioner.

A: . . . and suddenly the mother said something in Dutch to this girl, and this girl said to me 'My mother wants to know if you are a teacher?' and I said yeah.
B: Oh well.
A: Huu and she said and she'd gathered from the way I was talking, you know, and I wa I tend to wave my hands about and lay the law down a bit if I'm explaining something and and this girl said 'My mother wants to know if you are a teacher?' so . . .
B: Dutch is a funny language, but . . .
A: So it hadn't taken her very long to sum me up, had it?
C: Well . . .
B: German's easier to understand.
A: I believe Dutch is very difficult.
B: Yes.
A: Hmm.
B: But there's a, it it's a erm a lot like English (inaudible) . . . but it's the low German that's like English though.
A: Yeah, that's it, there are two Germans aren't there? There are low German and high German.
B: Not considered very good in Germany, I mean I think if you speak low German you are considered not to be very educated.
A: Hmm. The kids at school tell me that they find it easier to do German than French.
B: It's nearer, I think.
A: Hmm.
C: I think . . .
A: There are quite a few of them who drop French and keep on with German and if I say to them 'Why?' they say 'Oh it's easier, Miss, to do German.'
C: I was told that German is a very difficult language.
D: Hmm.
A: Well I wouldn't know because I've never done it, but . . .
C: I mean I I've seen thi well I've seen I haven't tried to learn any German, but I've seen a couple of things an I mean it looks difficult.
A: Hmm.
B: Written I think, but I think when they actually speak it . . .
C: It sounds difficult.
B: It it's much more like English than French.
A: It all sounds to me if people speak in German as if they are swearing.
C: (Laughter)
B: Yes, it's a bit like that isn't it.
D: Quite hard as well.
C: I prefer French personally, rather than German.
B: Yeah.
A: But nothing looks more difficult to me than Greek, I mean just looking at Greek is enough to put me off.
C: That's because the letters are completely different.
A: Hmm.
B: Yeah.

To conduct a full analysis of this extract would take a great deal of time, but we will draw out a number of points which are pertinent to the assessment scales provided by Carroll.

1 *Repetition* is high: the main function of lexical repetition appears to be to reintroduce a personal discourse topic which is ignored by the following speaker. Indeed, for part of the conversation the speakers refer back to their last utterance and not to the interlocutor's previous utterance. This is especially so at the beginning, when there is a fight for topic control. Later the repetition is much more concerned with propositional development.

2 *Hesitation* fulfils at least two roles: firstly, in connection with fillers 'so', 'well' or 'hmm' at the end of an utterance. Incomplete and in low key (see Brazil 1985), hesitation is used to indicate that the speaker wishes to give up the turn. Secondly, hesitation is connected very closely to the interesting phenomenon of having begun a sentence with a proposition in mind, but not having planned the sentence grammatically in advance. This leads to a rapid search for a structure which will allow the speaker to complete the utterance while preserving the original intention of that utterance.

3 *Stumbling* can now be seen as connected closely to a problem with grammatical forward planning, although in L2 learners, this may indeed be connected with a lack of knowledge of the grammatical system of the target language.

4 *Propositional development* is perhaps one of the most interesting properties of this particular piece of data. Once the discourse topic has been stabilized, the conversation focuses on the issue of whether or not German is easy or difficult to learn compared with French and, at the end, Greek. Once it has been established by speakers **A** and **B** that German is the easiest, speaker **C** contradicts this opinion, speaker **A** declares that she does not know, and speaker **D** agrees with speaker **C**. Speaker **B** provides resistance, but does little other than restate the original case. Hence the close connection with repetition, both lexical and grammatical. As already mentioned, speakers pursue propositions which are of little relevance to the propositions of the other speakers in the early part of the conversation. For these reasons, it can hardly be said that propositional development is smooth.

5 *Grammatical accuracy* appears to be something which is not possessed by native speakers engaging in informal conversation. It is clearly related to all the other aspects of fluency already mentioned, and is a function of their relationships. Only L2 learners are trained to be self-conscious monitors even when discussing the weather!

Conclusions

It would appear that our four participants in the above conversation would not meet the appropriate criteria embedded in the ELTS assessment scales if they are measured on the recorded material sampled from 'real life'. It is suggested that the present assessment scale, based on the functional–notional categories, is attempting to describe not what actually happens in communicative situations, but what communicative theorists think happens in communicative situations. This has two consequences.

First, in the rush to improve content validity, construct validity has been ignored. Within the communicative approach we see evidence that notional–functional considerations have so far ruled the roost, at the expense of discourse considerations, which are only just being able to make their influence felt in testing circles. And secondly, the concern with content validity has resulted in serious mistakes in test development. Take, for example, the report of a paper on communicative progress tests delivered at the IATEFL conference of 1986. The author declares:

> The risk of reduced reliability in such tests has to be acknowledged. However, their validity is high and the fact that the ELTS-type tests account for one half of the entire assessment means that reliability on that part of the test system is ensured. (McNeill 1986:45).

Reliability is a prerequisite for validity, and the validity of a test can never exceed its reliability. It is quite simple, really: if a test is not reliable, it is not actually measuring anything, and so cannot possess validity! Low reliability and high validity are a contradiction in terms, and the problem is compounded on tests like the ELTS where profiles are drawn up on the basis of sub-test scores. One has to be certain that the reliability of each sub-test is high, and calculate the reliability of the difference of scores in order to be fairly sure that any difference reflects a true difference of actual ability in the testee. Such an uncritical approach to reliability and construct validity cannot be allowed to go unchallenged when the future of the testee depends upon it.

This article has argued that communicative oral tests which are currently popular have claimed high content validity and slipped the theory in under the door, as it were. This has led to a confusion between types of validity and reliability, and to a concern that the assessment scales are actually based on theory with little empirical justification. It is further suggested that a new approach to construct validity in which the construct can be empirically tested can be found in discourse analysis. A discourse analysis orientation could then lead to the development of new communicative 'discourse' tests in all skills. The first stage in this process is to build up banks of recorded data, and to conduct research into how they can be utilized in the testing situation. □

Received November 1986

Note

1 The ELTS (English Language Testing Service), introduced in 1980, is conducted jointly by the University of Cambridge Local Examinations Syndicate and The British Council. For further information about the service, please contact The British Council, 10 Spring Gardens, London SW1A 2BN.

References

Alderson, J. C. (ed.) 1981. *Issues in Language Testing* (ELT Documents 111). London: The British Council.

Brazil, D. 1985. *The Communicative Value of Intonation in English*. University of Birmingham: English Language Research.

Carroll, B. J. 1980. *Testing Communicative Performance*. Oxford: Pergamon.

Carroll, B. J. 1981. 'Specifications for an English Language Testing Service' in Alderson (ed.) 1981.

Carroll, B. J. 1985. *Make Your Own Language Tests*. Oxford: Pergamon.

McNeill, A. 1986. 'Progress tests: can the format of tests reflect the activities of the course?' *IATEFL Newsletter*, August 1986.

Morrow, K. 1979. 'Communicative language testing: revolution or evolution?' in C. J. Brumfit and K. Johnson (eds.): *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press.

Munby, J. L. 1978. *Communicative Syllabus Design*. Cambridge: Cambridge University Press.

Oller, J. 1979. *Language Tests at School*. London: Longman.

The author

Glenn Fulcher is a lecturer in English Language at the Forum Language Institute, Nicosia, Cyprus, and is also responsible for test development. He is currently a post-graduate student of Applied English Linguistics at the University of Birmingham.