Chapter Four

Test Methods

Methods are used to construct tests but are not in themselves tests. Though it is possible to talk of a good or a bad test, or a valid or invalid test, this is obviously not possible for methods. The multiple choice procedure might produce a valid test in one realisation but not in another. This is the case for all methods and should be borne in mind when following the discussion on the potential advantages and disadvantages of the various test methods below.

The different approaches to language testing were outlined above in Chapter One and reference was made to the possible effect of test method on test scores. There is some evidence in the literature (see Murphy, 1978, 1980; Porter, 1983; Weir, 1983a; Boniakowska, 1986; and Alderson and Urquhart, 1985a) that test format might affect student performance. Given the limited state of knowledge concerning the effect of test formats, the only practical approach at present is to safeguard against possible format effect by spreading the base of a test more widely through employing a variety of valid, practical and reliable formats for testing each skill.

We have not yet discussed in any detail the different test methods currently in use. This chapter, therefore, gives a brief account of the main kinds of test formats and highlights some of their potential advantages and disadvantages. This is intended to provide a reference and set of guidelines for future test construction.

As a general rule it is best to assess by a variety of test formats, the scores on which are taken as a composite for reporting purposes. The main proviso for testing within a communicative framework is that the test tasks should as far as possible reflect realistic discourse processing and cover the range of contributory enabling skills that have been identified (see Appendix I for an example of this approach in the TEEP test). It is important that tests developed within this paradigm should have a strong washback effect on practice in the language classroom. What follows is an outline of some of the available options along the discrete point/integrative continuum.

4.1 Testing reading comprehension

4.1.1 Multiple-choice questions (MCQs)

The advice on the construction of multiple-choice items in this section is also applicable to the construction of tests of listening comprehension, structure and vocabulary. These are referred to later in this chapter.

A multiple-choice test item is usually set out in such a way that the candidate is required to select the answer from a number of given options, only one of which is correct. The marking process is totally objective because the marker is not permitted to exercise judgement when marking the candidate's answer; agreement has already been reached as to the correct answer for each item. Selecting and setting items are, however, subjective processes and the decision about which is the correct answer is a matter of subjective judgement on the part of the item writer.

Advantages

1. In multiple-choice tests there is almost complete marker reliability. Candidates' marks, unlike those in subjective formats, cannot be affected by the personal judgement or idiosyncrasies of the marker. The marking, as well as being reliable, is simple, more rapid and often more cost effective than other forms of written test.
2. Because items can be pre-tested fairly easily, it is usually possible to estimate in advance the difficulty level of each item and that of the test as a whole. Pre-testing also provides information about the extent to which each item contributes positively towards what the test as a whole is measuring. Ambiguities in wording of items may also be revealed by analysis of the pre-test data and can then be clarified or removed in the test proper.
3. The format of the multiple-choice test item is such that the intentions of the test compiler are clear and unequivocal; the candidates know what is required of them. In open-ended formats ambiguities in the wording of questions may sometimes lead to the candidates submitting answers to questions different from those which the examiner had intended to ask.
4. In more open-ended formats, e.g., short answer questions, the candidate has to deploy the skill of writing. The extent to which this affects accurate measurement of the trait being assessed has not been established. Multiple-choice tests avoid this particular difficulty.

Disadvantages

1. There are however a number of problems associated with the use of this format. If a candidate gets a multiple-choice item wrong because of some flaw in the question, the answer sheet on which he records his answer will not reveal this fact. In addition, we do not know whether a candidate's failure is due to lack of comprehension of the text or lack of comprehension of the question. A candidate might get an item right by

eliminating wrong answers, a different skill from being able to choose the right answer in the first place.

2. The scores gained in multiple-choice tests, as in true-false tests, may be suspect because the candidate has guessed all or some of the answers. This has the effect of narrowing the range of scores. The format of these tests encourages the candidate to guess and it is sometimes considered necessary to take steps to discourage candidates from doing so. It may also be possible to complete some items without reference to the texts they are set on, and, if this is so, whatever it is that is being tested, it cannot be comprehension of the text.

3. Multiple-choice tests take much longer and are more expensive and difficult to prepare than more open-ended examinations, e.g., compositions. A large number of items have to be written carefully by item writers who have been specially trained and these then have to be pre-tested before use in a formal examination. Each item has to be rigorously edited to ensure that:

   - There is no superfluous information in the stem.
   - The spelling, grammar and punctuation are correct.
   - The language is concise and at an appropriate level for candidates.
   - Enough information has been given to answer the question.
   - There is only one unequivocally correct answer.
   - The distractors are wrong but plausible and discriminate at the right level.
   - The responses are homogeneous, of equal length and mutually exclusive and the item is appropriate for the test.

4. It is extremely time-consuming and demanding to get the requisite number of satisfactory items for a passage, especially for testing skills such as skimming. A particular problem lies in devising suitable distractors for items testing the more extensive receptive skills. Heaton (1975) noted that, for these activities, it is more helpful to set simple open-ended questions rather than multiple-choice items; otherwise students will find it necessary to keep in mind four or five options for each item while they are trying to process the text.

5. A further objection to the use of multiple-choice format is the danger of the format having an undue effect on measurement of the trait. There is some evidence that multiple-choice format is particularly problematic in this respect. This has been evidenced by low correlations both with alternative reading measures and with other concurrent external validity data on candidates' reading abilities (see Weir, 1983a).

6. There is considerable doubt about their validity as measures of language ability. Answering multiple-choice items is an unreal task, as in real life one is rarely presented with four alternatives from which to make a choice to signal understanding. Normally, when required, an understanding of what has been read or heard can be communicated through speech or writing. In a multiple-choice test the distractors present choices that otherwise might not have been thought of. If a divergent view of the world is taken it might be argued that there is sometimes more than one right answer to some questions particularly at the inferential level. What the test constructor has inferred as the correct answer might not be what other readers infer, or necessarily be explicit in the text.

4.1.2 Short answer questions

These are questions which require the candidates to write down specific answers in spaces provided on the question paper. The technique is extremely useful for testing both reading and listening comprehension and the comments made below in reference to reading are, for the most part, also applicable to the testing of listening.

Advantages

1. Answers are not provided for the student as in multiple-choice: therefore if a student gets the answer right, one is more certain that this has not occurred for reasons other than comprehension of the text.

2. With careful formulation of the questions a candidate's response can be brief and thus a large number of questions may be set in this format, enabling a wide coverage.

3. If the number of acceptable answers to a question is limited it is possible to give fairly precise instructions to the examiners who mark them.

4. Activities such as inference, recognition of a sequence, comparison and establishing the main idea of a text, require the relating of sentences in a text with other items which may be some distance away in the text. This can be done effectively through short answer questions where the answer has to be sought rather than being one of those provided.

5. A strong case can be made in appropriate contexts, e.g., in EAP tests, for the use of long texts with short answer formats on the grounds that these are more representative of required reading in the target situation, at least in terms of length. They can also provide more reliable data about a candidate's reading ability (see Engineer, 1977 for evidence on the increased reliability resulting from the use of longer texts and Appendix I below for an example of this approach in the TEEP test).

Disadvantages

1. The main disadvantage to this technique is that it involves the candidate in writing and there is some concern, largely anecdotal, that this interferes with the measurement of the intended construct.

2. Care is needed in the setting of items to limit the range of possible acceptable responses and the extent of writing required. In those cases where there is more debate over the acceptability of an answer, e.g., in questions requiring inferencing skills, there is a possibility that the variability of answers might lead to marker unreliability. However, careful moderation and standardisation of examiners should help to reduce this.

4.1.3 Cloze

In the cloze procedure words are deleted from a text after allowing a few sentences of introduction. The deletion rate is mechanically set, usually between every fifth and eleventh word. Candidates have to fill each gap by supplying the word they think has been deleted. Alderson's research (1978a) established that more difficult texts were better measures of the lower order skills which cloze tests, than were easy texts. He found the semantically acceptable scoring procedure to be superior to any other.

In comparison of cloze and multiple-choice, Engineer (1977) concluded that the two techniques were measuring different aspects of the reading activity — namely that a timed cloze measured the *process* of reading, i.e., the reader's ability to understand the text while he is actually reading it; multiple-choice, on the other hand, measures the *product* of reading, namely the reader's ability to interpret the abstracted information for its meaning value.

There is a good deal of supportive evidence in the literature for using the cloze format. Klein-Braley (1981, p. 229) commented that: 'Up to now, in the main, the results of research with cloze tests have been extremely encouraging. They have shown high validity, high reliability, objectivity, discrimination and so on.' She quoted J.D. Brown (1979, p. 13): 'As demonstrated in this and other studies, it can be a valid and reliable test of overall second language proficiency.'

Alderson (1978a, p. 2) described how: 'The last decade, in particular, has seen a growing use of the cloze procedure with non-native speakers of English to measure not only their reading comprehension abilities but also their general linguistic proficiency in English as a Foreign Language.' He added (p. 39):

> The general consensus of studies into and with cloze procedure for the last twenty years has been that it is a realiable and valid measure of readability and reading comprehension, for native speakers of English . . . . As a measure of the comprehension of text, cloze has been shown to correlate well with other types of test on the same text and also with standardised testing of reading comprehension.

He pointed out that though this evidence is not available for non-native speakers (p. 63): 'it does seem cloze procedure is a potentially interesting measure of language proficiency for non-native speakers.'

The term 'cloze' was first introduced by W.L. Taylor (1953) who took it from the gestalt concept of 'closure' which refers to the tendency of individuals to complete a pattern once they have grasped its overall significance. Taylor (p. 416) described it as follows: 'A cloze unit may be defined as: any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product), by deciding from the context that remains, what the missing part should be.' The reader comprehends the mutilated sentence as a whole and completes the pattern. Alderson (1978a, p. 8) pointed out that: 'the cloze procedure becomes a measure of the similarity between the patterns that the decoder is anticipating and those that the encoder had used.'

Taylor first applied the procedure to gauging the readability of a text but later it came to be highly regarded as a measure of testing reading comprehension and even as a measure of overall language proficiency. For Bormuth (1962, p. 134): 'cloze tests are valid and uniform measures of reading comprehension ability.' Heaton (1975, p. 122) thought that: 'cloze tests measure the reader's ability to decode interrupted or mutilated messages by making the most acceptable substitutions from all the contextual clues available.'

Engineer (1977) found that a cloze test given under timed conditions provided valid and reliable indices of students' proficiency if two conditions were met: first, that the textual material used was of the appropriate level of difficulty for the population and second, that it contained a sufficient number of deleted items.

Advantages

1. Cloze tests are easy to construct and easily scored if the exact word scoring procedure is adopted. They are claimed to be valid indicators of overall language proficiency (see Bormuth, 1962; Brown, 1979; Engineer, 1977, and Oller, 1979).
2. With a fifth word deletion rate a large number of items can be set on a relatively short text and these can exhibit a high degree of internal consistency, in terms of Kuder-Richardson coefficients. This consistency may vary considerably, though, dependent on text selected, starting point for deletions and deletion rate employed.
3. In the literature cloze tests are often feted as valid and uniform measures of reading comprehension.

Disadvantages

1. Despite the arguments adduced in favour of cloze procedure, a number of doubts have been expressed, largely concerning its validity as a testing device. It has been shown to be irritating and unacceptable to students and doubt has been thrown on the underlying assumption that it randomly samples the elements in a text. Klein-Braley and Raatz (1984) in investigating its construct validity found that it fails to ensure random deletion of elements in a text.
2. Alderson (1978a, p. 392) had discovered that:

   > cloze procedure is not a unitary procedure, since there is a marked lack of comparability among the tests it may be used to produce. The fact emerges clearly that different cloze tests, produced by variations in certain of the variables, give unpredictably different measures, particularly of proficiency in English as a foreign language.

   If one changes the text, changes the deletion rate, starts at a different place or alters the scoring procedure, one gets a different test in terms of reliability and validity coefficients and overall test difficulty.
3. The evidence is contradictory about the differing scoring methods to be adopted in marking a cloze procedure. It has been suggested (Klein-Braley, 1985) that a cloze test is a much less effective measure for assessing 'general proficiency' in that it correlates less well with other established general proficiency measures when used on monolingual as against multilingual groups. In addition it seems that cloze is not suitable for restricted range groups (Klein-Braley, 1985); weak relationships have been found between cloze and teachers' judgements (Klein-Braley, 1981; 1985); cloze does not seem to correlate well with productive tests of speaking and writing and scores on cloze

cannot easily be related to native speaker performance since native speaker performance varies considerably from one cloze test to another (Alderson, 1978a).

4. The cloze procedure seems to produce more successful tests of syntax and lexis at sentence level than of reading comprehension in general or of inferential or deductive abilities, what might be termed higher order abilities (see Darnell, 1968). This would seem to accord with Alderson's (1978a, p. 99) findings that:

> cloze is essentially sentence bound . . . . Clearly the fact that cloze procedure deletes words rather than phrases or clauses must limit its ability to test comprehension of more than the immediate environment, since individual words do not usually carry textual cohesion and discourse coherence (with the obvious exception of cohesive devices like anaphora, lexical repetition and logical connectors).

5. Perhaps the most crucial reservation is the question of what performance on a cloze test really tells us about a candidate's language ability. It is difficult to translate scores on a cloze test to a description of what a candidate can or can't do in real life.

4.1.4 Selective deletion gap filling

In the light of recent negative findings on mechanical deletion cloze, increasing support has developed for the view that the test constructor should use a 'rational cloze', selecting items for deletion based upon what is known about language, about difficulty in text and about the way language works in a particular text. Linguistic reasoning is used to decide on deletions and so it is easier to state what each test is intended to measure (see Alderson, 1978a, p. 397; Klein-Braley, 1981, p. 244; and Weir, 1983a). This technique is better referred to as selective deletion gap filling as it is not 'cloze' in the proper sense.

Advantages

1. Selective deletion enables the test constructor to determine where deletions are to be made and to focus on those items which have been selected a priori as being important to a particular target audience.
2. It is also easy for the test writer to make any alterations shown to be necessary after item analysis and to maintain the required number of items. This might involve eliminating items that have not performed satisfactorily in terms of discrimination and facility value.

Disadvantages

1. It is important to stress that this technique restricts one to sampling a much more limited range of enabling skills (i.e., those abilities which collectively represent the overall skill of reading) than do the short answer and multiple-choice formats (see Weir, 1983a). Whereas short answer and multiple-choice questions allow the sampling of the range of reading enabling skills, gap filling is much more restrictive where only single words are deleted. Gap filling only normally allows the testing of sentence bound reading skills.

2. If the purpose of a test is to sample the range of enabling skills including the more extensive skills such as skimming, then an additional format to gap filling is essential.

4.1.5 C-Tests

Recently an alternative to cloze and selective deletion gap filling has emerged for testing comprehension of the more specifically linguistic elements in a text. An adaptation of the cloze technique called the C-test has been developed in Germany by Klein-Braley (1981, 1985; Klein-Braley and Raatz, 1984) based on the same theoretical rationale as cloze, viz., testing ability to cope with reduced redundancy and predict from context.

In the C-test every second word in a text is partially deleted. In an attempt to ensure solutions students are given the first half of the deleted word. The examinee completes the word on the test paper and an exact word scoring procedure is adopted.

Advantages

1. With C-tests a variety of texts are recommended, and given the large number of items that can be generated on small texts this further enhances the representative nature of the language being sampled. Normally a minimum of 100 deletions are made and these are more representative of the passage as a whole than is possible under the cloze technique.
2. The task can be objectively scored because it is rare for there to be more than one possible answer for any one gap.
3. Whereas in cloze the performance of native speakers on the test is highly variable, according to Klein-Braley (1985) it is much more common for native speakers to be able to score 100 per cent on C-tests. This may be of some help in setting cutting scores, e.g., what percentage constitutes a pass.
4. The C-test is economical and the results obtained to date are encouraging in terms of reliability and internal and external validity. It would seem to represent a viable alternative to cloze procedure and selective deletion gap filling.

Disadvantages

1. Given the relatively recent appearance of the technique in this form there is little empirical evidence of its value. Most concern has been expressed concerning its public acceptability as a measure of language proficiency. It is interesting to note that Davies (1965) has a version of this technique in his battery where the first letter of a word is given.
2. This technique suffers from the fact that it is irritating for students to have to process heavily mutilated texts and the face validity of the procedure is low.

4.1.6 Cloze elide

A technique which is generating interest recently is where words which do not belong

are inserted into a reading passage and candidates have to indicate where these insertions have been made. There is in fact nothing new about this technique and Davies was using it much earlier (Davies, 1965). In its earlier form it was known as the intrusive word technique.

Advantages

1. In comparison with multiple-choice format or short answer questions the candidate does not have the problem of understanding the question. It has approximately the same item yield as a cloze test.

Disadvantages

1. Scoring is highly problematic as candidates may, for example, delete items which are correct, but redundant.

4.1.7 Information transfer

In testing both reading and listening comprehension we have referred to the problem of the measurement being 'muddied' by having to employ writing to record answers. In an attempt to avoid this contamination of scores several Examination Boards in Britain have included tasks where the information transmitted verbally is transferred to a non-verbal form, e.g., by labelling a diagram, completing a chart or numbering a sequence of events (see Appendix V for interesting examples of this in the JMB Test).

Advantages

1. Information transfer techniques are particularly suitable for testing an understanding of process, classification or narrative sequence and are useful for testing a variety of other text types. It avoids possible contamination from students having to write answers out in full.
2. It is a realistic task for various situations and its interest and authenticity gives it a high face validity in these contexts.

Disadvantages

1. A good deal of care needs to be taken that the non-verbal task the students have to complete does not itself complicate the process. In some tasks students may be able to understand the text but not what is expected of them in the transfer phase.
2. There is a danger of cultural and educational bias. Students in certain subject areas may also be disadvantaged, e.g., some students in the social sciences may not be as adept in working in a non-verbal medium as their counterparts in science disciplines.

4.1.8 Conclusion

For testing reading abilities we would recommend the use of short answer questions together with selective deletion gap filling. The C-test is an interesting alternative to the latter and its acceptability to students and validity are worthy of further investigation. If we are to develop the communicative nature of our tests it is perhaps important to focus on performance tasks in reading tests, and the use of information transfer techniques and other restricted response formats is advocated.

4.2 Testing listening comprehension

4.2.1 Testing extensive listening skills

The rationale behind the construction of many of the earlier listening comprehension tests was described by Valette (1967, p. 49): 'The main object of a listening test is to evaluate the student's comprehension. His degree of comprehension will depend on his ability to discriminate phonemes, to recognise stress and intonation patterns, and to retain what he has heard.'

It was thought that, if a learner was tested in phoneme discrimination, stress and intonation, the sum of the 'discrete' sub-tests would be equivalent to his proficiency in listening comprehension. An example of the test of this type is the ELBA test battery constructed by Ingram (1964) which placed the emphasis on 'discrete' listening items such as sound recognition, intonation and stress, using short items rather than continuous passages of discourse or dialogue. As Ryan (1979) pointed out, even the section described as listening comprehension seemed more a test of appropriate-response mechanisms than a test of comprehension of continuous speech in an authentic context.

A noticeable trend in recent years has been the attempt to differentiate between tests of auditory discrimination and contextualised tests of listening comprehension. Templeton (1973) outlined how research began to focus on these integrative tests of listening comprehension in preference to discrete point tests of phoneme discrimination, intonation and word and sentence stress.

Since 1969 the JMB no longer tests individual aural skills in isolation, but instead tests listening comprehension in an integrated context of lecturettes or dialogues (see McEldowney, 1976, and Appendix IV). This paradigm shift can also be observed in the 1977 version of EPTB (see Davies, 1978) which substituted for the earlier analytical, phoneme discrimination, stress and intonation tasks, an overall listening comprehension sub-test containing, for example, an integrated test of listening comprehension based on a lecture with simulated note-taking.

Davies (1978, pp. 146–8) illustrated how similar changes had occurred between the listening tasks described in the first and second editions of Valette's book on testing (cf. Valette, 1967; 1977): 'we can characterise the difference between Valette (1967) and Valette

(1977) as a move from linguistics to sociolinguistics, from structuralism to functionalism, from taxonomy and breaking down into skills, into discrete parts, to integration and building up into wholes.' In the second edition of Valette (1977) Davies noted (p. 147): 'a move from a concentration on sound, the production of speech, the phonology, to meaning and communication.'

A strong argument against auditory discrimination as a test of proficiency in listening comprehension was that the ability to distinguish between phonemes, however important, did not necessarily imply an ability to understand verbal messages. Furthermore, as Ryan (1979) pointed out, occasional confusion over selected pairs of phonemes does not matter too greatly, because in real life situations the listener has contextual clues to facilitate understanding. For Valette (1977, p. 102): 'The key concern of the evaluator is to determine whether the students have received the message that was intended and not whether they made certain sound discriminations or identified specific structural signals.'

J.W. Morrison (1974) after assessing the listening comprehension needs of science students at the University of Newcastle-upon-Tyne, concluded that at the ESP/EST level, performance needs to be considered at a level beyond phonology and grammatical structure, thus taking into account the communicative context of spoken discourse. Chaplen (1970a, p. 19) had earlier concluded that: 'Whatever the contribution of the elements of oral/aural communication — intonation, stress and phonemic discrimination — to a test of oral/aural communication, their importance appears to be minimal at any level of proficiency beyond a very elementary stage.'

Holes (1972) developed test instruments which focused on the ability to handle academic lectures, a communicative task regarded by departments as both crucial and difficult for their overseas students. He approached test design from a 'job-sampling' viewpoint and attempted to assess the more global, less 'pure' ability of students to interpret 'message content' as well as eliciting data on their linguistic competence. He used the Davies Test as part of his concurrent validation procedures and an interim academic success/failure rating for predictive validity purposes. Though the predictive validity correlations of tests versus subject examination results were non-conclusive, Holes concluded (p. 134) that: 'The value of the tests lay rather in what they revealed about the kinds of difficulty which overseas students experience in lectures.'

In line with the paradigm shift described above it is usual to provide ongoing and sequential texts as stimuli in a test battery, though, in terms of the tasks, items and scoring, it might be desirable in certain components of the test to focus on discrete items. As with tests of reading comprehension, a balance of integrative and 'discrete point' is felt to be the most satisfactory approach for maximising reliability and validity in a test.

Multiple-choice questions

In our consideration of the use of this technique in the assessment of reading in Section 4.1.1. above it is clear that the disadvantages of employing the technique far outweigh any advantages it might have. These disadvantages are equally applicable for using this format in the testing of listening.

Because of the problems associated with the serial nature of the listening process there are additional difficulties in employing this technique as a measure of listening ability, for example, the extra burden that is placed on processing by having to keep four options in mind (Heaton, 1975). The format is artificial and is increasingly perceived as an invalid method for assessing comprehension by teachers, materials designers and language testers. The new General Certificate of Secondary Education (GCSE) examinations in the United Kingdom will not be employing multiple-choice format largely because of hostile comment from teachers' organisations on its validity as a teaching and testing technique.

The RSA CUEFL examination, despite conscientious efforts to maximise authenticity in the stimulus texts selected, is heavily dependent on this format and its variants (for example true-false items) and has been criticised for the retreat from realistic discourse processing that this involves (see Appendix III). The use by the RSA of these more objective formats highlights the need for trying to establish realism in both text stimulus and the tasks that are expected of the student, and the sometimes contradictory pulls of reliability and validity.

Short answer questions (SAQs)

Advantages

1. Short Answer Questions can be a realistic activity for testing listening comprehension, for example, if one wishes to simulate real life activities where notes are taken as somebody communicates a spoken message. With sufficient care the responses can be limited and so the danger of the writing process interfering with the measurement of listening is largely avoided (see Appendix I).
2. In contrast to the multiple-choice or the true-false formats employed in some examinations, one can be more certain that correct answers have not been arrived at by chance.

Disadvantages

1. If the candidate has to write an answer at the same time as listening to continuous discourse there are obvious problems. An unnecessary load might be placed on the memory and vital information in the ongoing discourse might be missed while the answer to a previous question is being recorded.

Information transfer techniques

This technique was discussed above in connection with reading and it is worthy of consideration for similar reasons in a listening test (see Appendix V).

Advantages

1. A particular advantage for using this technique in testing listening is that the student

does not have to process written questions while trying to make sense of the spoken input. It is particularly efficient for testing an understanding of sequence, process, relationships in a text and classification.

Disadvantages

1. It is often very difficult to find spoken texts which lend themselves to a non-verbal format. Whereas in reading a certain amount of editing of texts is feasible and in general a greater variety of texts are more readily available, this is not the case for listening texts taken from authentic sources.

Limitations on the testing of extensive listening

It is important to note that, if one wishes to make test tasks more like those in real life, the serial nature of extended spoken discourse and the greater processing problems associated with understanding spoken English preclude items which focus on the more specifically linguistic skills such as working out the meaning of words from context or recognising the meaning value of specific features of stress or intonation. It is, for example, extremely difficult for students to backtrack and focus on very specific features of discourse while listening to and attempting to understand a non-interactive, uninterrupted monologue. To preserve the integrative nature of the test, therefore, we have to focus questions on the more global processing skills such as inferencing, listening for specifics or identifying the main ideas.

A serious problem in testing extensive listening by use of the tape recorder is that the visual element, the wealth of normal exophoric reference and paralinguistic information, is not available to the candidate and perhaps, therefore, the listening task is made that much more difficult for the candidate. The listener does not normally have to process disembodied sounds from a tape recorder in real life (apart from the obvious exceptions such as listening to the radio).

Until there is greater accessibility to video equipment the artificiality of a straight audio listening task will remain a problem. Even video is likely to have its own practical difficulties though, e.g., the number of screens required so that all viewers are treated equally or the incompatibility of various systems. Whatever the instrumentality, in a test situation the student is in any case denied the natural context provided by the experience of using the language in contiguous situations.

There is a great danger in listening tests that the candidates might be expected to cope with additional difficulties arising from the restricted context available and steps need to be taken to compensate for this or we might seriously underestimate the ability to process spoken language.

4.2.2 *The testing of intensive listening*

Reference was made above to the difficulty of focusing on specific listening points while candidates are exposed to ongoing discourse. Given the need to enhance the reliability of our test batteries it is often advisable to include a more discrete format with the possibility

this gives of including a greater number of specific items. A dictation or listening recall test can provide this discreteness as well as being valid in content terms for certain groups of candidates, particularly those involved in academic study through the medium of English.

Dictation

It is important that the candidates should be assessed in situations as close as possible to those in which they will be required to use the language. For dictation, this involves them listening to dictated material which incorporates oral messages typical of those they might encounter in the target situation.

Advantages

1. Given our concern with reliability as well as validity, it is perhaps advisable to improve the overall reliability of a listening battery by including a format which has a proven track record in this respect. A dictation can provide this reliability through the large number of items that can be generated as well as being valid for specific situations where dictation might feature as a target group activity.
2. There is a lot of evidence which shows dictation correlating highly with a great variety of other tests, particularly with other integrative tests such as cloze and it is often employed as a useful measure of general proficiency. There is some evidence that the use of a semantic scoring scheme (see Weir, 1983a) as against an exact word system serves to enhance the correlations with other construct valid tests of listening.
3. Criticisms of dictation in the past stemmed from a viewpoint heavily influenced by structural linguistics that favoured testing the more discrete elements of language skills and wished to avoid the possibility of muddied measurement. Heaton (1975) commented: 'as a testing device it measures too many different language features to be effective in providing a means of assessing any one particular skill'. The proponents of dictation, however, consider its very 'integrative' nature to be an advantage since it reflects more faithfully how people process language in real life contexts.
4. The new interest in dictation reflected the paradigm shift in testing values and objectives referred to above. Whereas in 1967 Valette had observed that foreign language specialists were not in agreement on the effectiveness of dictation as an examination for more advanced students, significantly ten years later she was able to state that dictation was a precise measure of overall proficiency and an excellent method of grouping incoming students according to ability levels.
5. An important factor in the return of dictation to popularity as a testing device was the research carried out by Oller, which formed part of a wider interest in integrative testing. Oller (1979) rejected current criticisms of dictation and argued that it was an adequate test of listening comprehension because it tested a broad range of integrative skills.

6 Oller (1979) claimed that a dynamic process of analysis by synthesis was involved Dictation draws on the learner's ability to use all the systems of the language in conjunction with knowledge of the world, context, etc , to predict what will be said (synthesis of message) and after the message has been uttered to scrutinise this via the short term memory in order to see if it fits with what had been predicted (analysis)

7 Dictation for Oller tests not only a student's ability to discriminate phonological units but also his ability to make decisions about word boundaries, in this way an examinee discovers sequences of words and phrases that make sense and from these he reconstructs a message The identification of words from context as well as from perceived sounds is seen by Oller as a positive advantage of dictation in that this ability is crucial in the functioning of language The success with which the candidate reconstructs the message is said to depend on the degree to which his internalised 'expectancy grammar' replicates that of the native speaker Fluent native speakers nearly always score 100 per cent on a well-administered dictation while non native learners make errors of omission, insertion, word order, inversion, etc , indicating that their internalised grammars are, to some extent, inaccurate and incomplete, they do not fully understand what they hear and what they reencode is correspondingly different from the original

8 According to Oller, research showed that dictation test results were powerful predictors of language ability as measured by other kinds of language tests (see Oller, 1971, Valette, 1977)

Disadvantages

1 Alderson (1978a) concludes that the evidence concerning dictation is inconclusive and that it is useful only as part of a battery of listening tests rather than a single solution He points out (1978a, p 365) that

> The reason it correlates more with some sub-tests than with others does not appear to be due to the claimed fact that it is an integrative test, but because it is essentially a test of low level linguistic skills Hence the dictation correlates best with those cloze tests texts and scoring methods which themselves best allow the measurement of these skills

2 Dictation will be trivial unless the short term memory of the students is challenged and the length of the utterances dictated will depend on the listeners' ability up to the limit that native speaker counterparts could handle

3 Marking may well be problematic if one wishes to take into account seriousness of error or if one wishes to adopt a more communicatively oriented marking scheme where a mark is given if the candidate has understood the substance of the message and redundant features are ignored

4 If the dictation is not recorded on tape, the test will be less reliable, as there will be differences in, for example, the speed of delivery of the text to different audiences

5 The exercise can be unrealistic if the texts used have been previously created to be read rather than heard

Listening recall

In contrast to dictation, very little evidence is available about listening recall tests (see Furneaux, 1982, Henning, 1982, and Beretta, 1983 for a full description of this procedure) The student is given a printed copy of a passage from which certain content words have been omitted (these deletions are checked in advance to ensure that they cannot be repaired by reading) The words deleted are normally content words felt to be important to an understanding of the discourse and the blanks occur at increasingly frequent intervals

Students are given a short period of time to read over the text, allowing for activation of their expectancy grammars They have to fill in the blanks, having heard a tape recording of the complete passage twice They are advised to listen the first time and then attempt to fill in the banks during a short period allowed for writing in the answers They hear the passage a second time and then are allowed a short period of time to write in any remaining missing words These limited write-in times draw on their short term memories The format involves many of the linguistic factors outlined above for dictation and this is reflected in the other names which have been given to the test spot dictation and combined cloze and dictation

Advantages

1 Like dictation it can be administered rapidly and scored objectively and it allows the tester to focus on items which are deemed to be important (as in selective deletion gap filling)

2 High correlations have been reported with other more direct tests of listening (Beretta, 1983) and with test totals for listening batteries

3 It has advantages in large scale testing operations in that it is easy to construct, administer and mark

Disadvantages

1 The difficulty in this technique lies in stating what it is that is being tested As only one word is deleted it may not be testing anything more than an ability to match sounds with symbols aided by an ability to read the printed passage containing the gaps

2 It is an inauthentic task and involves reading ability as well as listening Careful construction is needed to ensure that the students cannot fill in the blanks simply by reading the passage without having to listen at all

3 Given the high correlations that have been discovered between listening recall and dictation (see Furneaux, 1982, Beretta, 1983), and roughly equivalent practicality and reliability, the greater potential validity of dictation for certain groups, e g , for students

studying through the medium of English, might lead to a preference for dictation over listening recall.

4. Problems may occur in the marking if consideration is given to anything other than exact spelling.

4.2.3 Conclusion

Where possible listening tests should include an authentic performance task. An attempt should be made to incorporate information transfer techniques (see Appendix V) where appropriate. We might usefully include short answer questions and consideration could be given to dictation (see Appendix I).

4.3 Testing writing

Two different approaches for assessing writing ability can be adopted. Firstly, writing can be divided into discrete levels, e.g., grammar, vocabulary, spelling and punctuation, and these elements can be tested separately by the use of objective tests. Secondly, more direct extended writing tasks of various types could be constructed. These would have greater construct, content, face and washback validity but would require a more subjective assessment.

4.3.1 Indirect methods for assessing linguistic competence

In Section 4.1 above we examined the formats of cloze, selective deletion gap filling and C-tests, and commented on the value of these techniques for testing the more specifically linguistic, sentence bound reading skills, viz., items focusing on an understanding of vocabulary, structure or cohesion devices.

Both productive and receptive skills can be broken down into levels of grammar and lexis according to a discrete point framework. McEldowney (1974, p. 8) commenting on the syllabus of the JMB Test in English (Overseas) stated:

> To be able to operate these four skills (listening, reading, speaking and writing) in the various function areas it is necessary to be able to manipulate items from three levels of language. That is, to communicate, it is necessary to have an adequate vocabulary, to know basic items of English grammar and to be able to handle English sounds, stress and intonation.

The JMB Test in English (Overseas), as well as including tasks testing written production, also has tasks which test knowledge of 'basic productive vocabulary' and 'minimum grammatical items' (see Appendix V). The problems which face the constructors of vocabulary tests are manifold. Chaplen (1970a) who constructed the sub-tests for the vocabulary sections of the early JMB tests noted two main problem areas:

1. The selection of lexical items for testing.
2. Methods used to test the lexical items.

If the examinees are studying a variety of different subjects, as is the case in an EAP context, then there is a serious problem of selection. The more generalised the subject matter, the more difficult it is to select for testing purposes. In specialised courses, where there is an identifiable, agreed register, selection is easier but still exacting.

An additional problem occurs in the relative weighting that should be given to items selected from future reading materials as against the items that are likely to be employed in extended writing tasks. Do we test active or passive vocabulary? Furthermore, how do we establish the frequency and importance levels of the lexical items intended for use in the test?

Given the constraints on testing these items discretely it is by no means clear whether the results of such tests should form part of a profile of reading or of writing ability. They do not fit comfortably into either.

Similar problems occur in the selection of grammatical items for inclusion in an indirect test of linguistic competence. A quantitative survey of the occurrence of structural items in the receptive and productive written materials a test population will have to cope with in future target situations is obviously beyond the scope of most test constructors. A more pragmatic, subjective method of taking decisions on which items to be included is needed. It would seem sensible to examine the content of existing tests and coursebooks at an equivalent level to determine what experts in the field have regarded as suitable items for inclusion for similar populations.

There also seems to be a problem in reporting on what is being tested in these discrete point grammar tests. Should the performance on an indirect test of grammatical knowledge be reported on under the profile for reading or writing? Additionally indirect techniques are restricted in terms of their perceived validity for test takers and the users of test results. An interesting attempt to retain the objectivity and coverage of the discrete point approach whilst enhancing validity can be found in the editing task in Paper Two of the TEEP test (see Weir, 1983a and Appendix I).

Editing task

In the editing task the student is given a text containing a number of errors of grammar, spelling and punctuation of the type noted as common by remedial teachers of students in the target group and is asked to rewrite the passage making all the necessary corrections.

Advantages

1. As well as being a more objective measurement of competence, this task may have a good washback effect in that students may be taught and encouraged to edit their written work more carefully.
2. It is certainly more face valid than other indirect techniques as it equals part of the writing process.

Disadvantages

1. If the student rewrites the passage in his own words instead of just correcting the errors, the problems of marking are considerable. There is also some doubt as to whether the ability to correct somebody else's errors equates with an ability to correct one's own.

2. Marking can be problematic also if a candidate alters something which is already correct; a sin of commission rather than omission.

4.3.2 The direct testing of writing

With a more integrative and direct approach to the testing of writing, we can incorporate items which test a candidate's ability to perform certain of the functional tasks required in the performance of duties in the target situation. For doctors in a hospital this might involve writing a letter to a local GP about a patient on the basis of a set of printed case notes. For a student in an EAP context it might involve search reading of an academic text to extract specified information for use in a written summary (see Appendix I).

Essay tests

This is a traditional method for getting students to produce a sample of connected writing. The stimulus is normally written and can vary in length from a limited number of words to several sentences. The topics are often very general and rely heavily on the candidate providing the content out of his or her head. The candidates are not usually guided in any way as to how they are expected to answer the question.

Advantages

1. The essay has traditionally been accorded high prestige as a testing technique which may explain a widespread reluctance to discard it despite the problems in marking that have been encountered (see Coffman, 1971; Gipps and Ewen, 1974).
2. The topics are extremely easy to set and it is a familiar testing technique to both the candidates and the users of test results. It thus has a superficial face validity in particular for the lay person.
3. It is a suitable vehicle for testing skills, such as the ability to develop an extended argument in a logical manner, which cannot be tested in other ways.
4. The big advantage it shares with other tests of extended writing is that a sample of writing is produced which can provide a tangible point of reference for comparison in the future.

Disadvantages

1. Free, open-ended writing is problematic. An ability to write on general open-ended topics may depend on the candidate's background or cultural knowledge, imagination or creativity. These may not be factors we wish to assess.
2. The candidate may not have any interest in the topic he is given and if a selection of topics is provided it is very difficult to compare performances especially if the production of different text types is involved.

3. Candidates tend to approach an open ended question in different ways and examiners have to assess the relative merits of these different approaches. This increases the difficulty of marking the essays in a precise and reliable manner.
4. Time pressure is often an unrealistic constraint for extended writing and writing timed essays is not normally done outside of academic life. For most people the writing process is lengthier and may involve several drafts before a finished version is produced.
5. The inclusion of an extended writing component in an examination is time consuming in terms of the total amount of test time that is available for testing all the skills.

Controlled writing tasks

There is obviously a very strong case for including a test of writing on the grounds of the perceived content validity of 'job sample' tasks. It tests important skills which no other form of assessment can sample adequately. To omit a writing task in situations where writing tasks are an important feature of the student's real life needs might severely lower the validity of a testing programme.

Wall (1982) carried out an illuminating investigation of the kinds of writing task engineering students were required to perform as part of their coursework and compared these with the types of essay they were set in the Michigan Battery used for assessing students' language proficiency on entry to the university. She (p. 166) summarised the differences as follows:

> The main difference seems to be that in the engineering tasks there is much prior input and the task itself is explicitly outlined, whereas in the composition the writer has only a suggestion to respond to and must not only create the content of the writing but a context, audience and purpose as well. The criteria for marking would also seem different.

The conclusion to the investigation was disturbing. The research produced: 'a correlation study between the Michigan Battery total and part scores and the student's first term GPA, in which no significant relationship between tests and the criterion for academic success could be found.' In other words no relationship at all could be found between writing performance in the test and subsequent indicators of performance in the course of study.

Free, uncontrolled writing would seem to be an invalid test of the writing ability required by most students. It is easier to extrapolate from writing tests when care is taken in specifying for each task: the media, the audience, the purpose and the situation in line with target level performance activities (see Wall, 1982). When the task is determined more precisely in this manner it is also easier to compare performances of different students and to obtain a greater degree of reliability in scoring. If the writing task is uncontrolled examinees may also be able to cover up weaknesses by avoiding problems.

There are various types of stimuli that can be used in controlled writing tasks. Stimuli can be written, spoken or most effectively non-verbal, e.g., a graph, plan or drawing which the student is asked to interpret in writing (see Appendix V; Dunlop, 1969; McEldowney, 1974, 1976, 1982; Weir, 1983a, and Appendix I for examples of these).

Advantages

1. The advantage of non-verbal stimuli is that if they present information in a clear and precise way the candidate does not have to spend a long period of time decoding a written text. The task is most effective when the candidate is asked to comment on particular trends shown in a graph, or to compare and contrast one set of figures with another. Different stimuli can be used to elicit written performance of a number of different language functions such as argumentation, description of a process, comparison and contrast or writing a set of instructions.

Disadvantages

1. Problems have arisen when, through a desire to favour no particular groups of candidates, a test has resorted to extremely specialised areas such as bookbinding or mediaeval helmets for its visual stimuli. Often candidates are unable to cope with the mental challenge of taking this sort of test and give up rather than jump through the intellectual hoops necessary to get into the writing task.

   Problems are always likely to occur when the complexity of the stimulus obstructs the desired result, i.e., one needs to understand a very complex set of instructions and/or visual stimuli to produce a relatively straightforward description of a process or a classification of data.

2. The difficulties generated by these information transfer type tasks may arise through educational or cultural differences in ability to interpret graphs or tables or line drawings.

Summary

Advantages

1. Summary can be a valid test in certain domains, for example, it is very suitable for testing a student's writing ability in terms of the tasks he has to cope with in an academic situation. The writing of reports and essays requires the ability to select relevant facts from a mass of data and to re-combine these in an acceptable form. Summary of the main points of a text in this fashion involves the ability to write a controlled composition containing the essential ideas of a piece of writing and omitting non-essentials.

Disadvantages

1. The problem of the specificity of the text candidates are expected to produce arises in summary tasks as in other controlled writing tasks. There is often a difficulty in selecting appropriate stimulus texts because their subject specificity would create too many problems for non-specialists in the subject and the test might therefore be invalid. One alternative is to choose deliberately obscure texts which in theory favour nobody, to get at underlying abilities. This might bring into play features we do not want to test such as imagination.

If, however, students in the fields of science and engineering have to read 'general' or 'neutral' texts and then summarise using a wide range of non-scientific vocabulary and demonstrating qualities of literary style and imagination there might be serious validity problems for these students. Though a science student may not be able to summarise a piece on why a cat might make a suitable pet for an old lady he might be able to summarise the salient features of a process.

2. The main difficulty with an integrated writing component of this type is making the marking reliable and consistent. To assess students' responses reliably one needs to formulate the main points contained in the extract, construct an adequate mark scheme and standardise markers. Some subjectivity inevitably remains and it is easy to underestimate the difficulty of marking reliably.

   The relative merits of impressionistic and analytic approaches to marking for improving the reliability and validity of a writing sub-test are examined below. Little attention is normally accorded to improving the reliability of marking extended writing in the literature and so an attempt has been made to survey the field and bring together what is known about the main approaches to this problem. In terms of the overall structure of this book it occupies a relatively large amount of the discussion on test method but given the crucial importance of this particular skill to students studying through the medium of English the extended treatment is considered worthwhile. The comments made on the marking of writing apply, *mutatis mutandis*, to the assessment of spoken production. In both we have an identifiable product which can be evaluated in terms of previously specified criteria.

4.3.3 Analytical and general impression marking

Comparison of the two approaches

We have discussed how, by controlling the writing tasks in our battery, we might improve their validity and reliability. We concluded that there was a need for 'controlled' writing sub-tests in which the register, context and scope of the writing task were determined for the candidate. This would facilitate marking and allow more reliable comparison across candidates. In this section we examine how the standardised application of impressionistic and analytic approaches to marking might also aid us in our attempt to improve the reliability and validity of our writing sub-tests.

Analytical marking refers to a method whereby each separate criterion in the mark scheme is awarded a separate mark and the final mark is a composite of these individual estimates.

The impression method of marking usually entails two or more markers giving a single mark based on their total impression of the composition as a whole (see Wiseman, 1949; E. Ingram, 1970). Each paper is scored using an agreed scale and an examinee's score is the average of the combined marks. The notion of impression marking specifically excludes any attempt to separate the discrete features of a composition for scoring purposes.

According to Francis (1977), in its purest form, impression marking usually requires each marker to read a sample of scripts, perhaps 10–25 per cent, to establish a standard in his mind and thereafter to read all scripts quickly and allocate each script to a grade or mark range.

Hartog *et al.* (1936) conducted one of the earliest studies into the relative effectiveness of analytical and general impression marking for assessing English composition. They were intent on finding out which method produced the superior results in terms of ability to reduce marker error. Their research found (p. 123) that variation between markers was, to some extent, reduced by the analytic method: 'there are greater discrepancies between marks awarded by impression than between marks awarded by details ... it appears that these discrepancies are entirely due to greater differences in the standards of marking of different examiners when they mark by impression.'

The investigation also demonstrated that a large number of examiners were consistently biased in terms of either leniency or severity in their marking. The evidence they produced of discrepancies in rank order placements was in many ways more serious, since disagreements of this kind are not susceptible to correction in the same way as differences deriving from mark range bias. Both could, however, have been corrected by the provision of a detailed mark scheme and by the efficient standardisation of examiners prior to the marking exercise.

Like Hartog *et al.* (1936), Cast (1939) found the analytical method slightly superior in a single marker system. His criticisms of the impression method were that, though it discriminated more widely among individual candidates, it judged them on more superficial characteristics than the analytic method. However, although the analytical method was considered the more suitable, Cast felt that the results did not provide definitive evidence of the superior reliability of analytical marking and, therefore, refused to advocate the exclusive use of either method.

Cast pointed to important characteristics inherent in the two systems. An important feature of the analytical method to which he drew attention (pp. 263–4) was: 'on averaging their marks for all the questions, the range inevitably shrinks .... This "regression" is the inevitable consequence of all forms of summation of incompletely correlated figures.' In comparison, he noted (p. 263) that impression marking discriminated more widely among individual candidates and that the range of marks awarded by different examiners to the same script tended to be unusually wide.

Cast (p. 264) also noted the tendency of impression marking:

> to seize on a few salient or superficial points — errors of spelling, grammar or fact, perhaps — and weight those out of all proportion to the rest: on the other hand, the analytic methods, by dealing with numerous isolated and possible inessential points, may overlook certain general qualities that characterize the essays as a whole.

Francis (1977) also pointed out that a great danger of impression marking a piece of writing is that impression of the quality as a whole will be influenced by just one or two aspects of the work. He argues that the prejudices and biases of the marker may play a greater part in determining the mark than in the analytical scheme.

Multiple marking

Wiseman (1949) investigated the possibilities of improving assessment by summing the multiple marks of four independent, unstandardised markers, using a rapid impression method. He found that multiple marking by impression method improved reliability and was much quicker than comparable analytic procedures. He (p. 205) estimated that if the average inter-correlation of a group of four impression markers was as low as 0.6 with each other: 'the estimate of the probable correlation of averaged marks with "true" marks is 0.92. This is very much higher than we could expect from *one* analytic marker.'

Wiseman (p. 208) took pains to stress that: 'The efficiency of markers should be judged primarily by their self-consistency.' He pointed out (p. 204) that the consistency coefficient obtained by a pure mark, re-mark correlation (intra-marker reliability), using the same marking method on both occasions: 'is the one single measure which is quite clearly a true consistency, and one which is closest allied to the normal concept of test reliability.' By using a system of multiple marking based on this principle of self consistency he was able to achieve very high levels of reliability.

The work of Coffman and Kurfman (1968) and Wood and Wilson (1974) similarly drew attention to the problem of the instability of examiner marking behaviour. They produced evidence that marking behaviour does not remain stable during the whole marking period, when a large number of scripts are involved (see Edgeworth, 1888). They argued for subjecting each script to more than one judgement, which might help to neutralise the effects of inconsistent marking behaviour over a protracted period of assessment.

Though some doubt has been expressed in the past (see Edgeworth, 1888) about the expediency of having more than one marker, more recently Britton (1963), Britton *et al.* (1966), Head (1966), Lucas (1971) and Wood and Quinn (1976) all found that multiple marking improved the reliability of marking English essays.

Britton *et al.* (1966), in an experiment designed to devise a more reliable marking apparatus for use by examining boards, compared experimental multiple marking with the single marking carried out by a GCE examining board. They found (p. 21): 'The figures clearly indicate that in this case marking by individual examiners with very careful briefing and elaborate arrangements for moderation was in fact significantly less reliable than a multiple mark.' When the official marking and multiple marking were correlated with external criteria of coursework produced by candidates throughout the year, multiple marking was found to correspond more closely.

Head (1966) conducted an experiment to discover whether the added impression marks of two examiners would be more reliable than individual examiners. He found (p. 71): 'The raising of the coefficient from 0.64 for single mark correlations to 0.84 for paired mark correlations shows clearly that the added marks were more reliable.'

Lucas (1971) found that despite using somewhat inconsistent markers (mean mark/remark correlation only 0.65) multiple marking by impression increased the reliability of the mark awarded significantly. The greatest increase in reliability occurred in the change from one to two markers.

Wood and Quinn (1976) using 'O' level English Language essay and summary questions

found that impression marking by pairs of markers was more reliable than a single marking. They suggested, however, that there is no more to be gained in reliability from a single analytic marking than from a single impression marking. The real improvement is in double marking.

As regards the advantages of impression as against analytic marking though, there is evidence which indicates that multiple impression marking is not necessarily superior to multiple analytic marking. Penfold (1956) compared impression marking with analytic marking and found the latter much more effective in reducing inter-marker variance than the impression scheme. R.B. Morrison (1968) found that using impression marking did not produce more reliable marks than the standard analytical marking procedures employed by the Examining Board at the time. R.B. Morrison's findings (1968) were confirmed by a similar study he conducted a year later (R.B. Morrison, 1969).

In many of the studies we have examined there often seems to be an undisputed belief that work marked independently by two different markers, with their marks being averaged, is a more reliable estimate than if it were marked by a single marker. This general viewpoint needs qualifying though, for as was noted by Coffman and Kurfman (1968), Wiseman (1949) and Wood and Wilson (1974) in the discussion above, it is dependent on the markers being equally consistent in their own individual assessments for the duration of the marking period. If this is not the case the reliability of the more consistent marker on his own might in fact be superior to the combined reliability estimate for two markers who exhibit unequal consistencies.

These provisos must be borne in mind in considering the potential value of a double marking system. With an adequate marking scheme and sufficient standardisation of examiners, however, a high standard of inter-marker and intra-marker reliability should be feasible and the advantages of a double as against a single marker system would obtain.

Logistical considerations (time, money, computing, personnel) affecting multiple marking have, however, led to a widespread reluctance especially among examining boards to adopt it in large scale marking operations (see Penfold, 1956). A serious problem with multiple marking is that examiners sometimes find it difficult to avoid annotating a script to help them form their impression. If this script is to be remarked then either the second examiner approaches it in a dissimilar state to the first, the marks have to be tediously removed, or multiple copies of the script need to be made. In addition, practical difficulties in getting results out in a reasonable period after the conduct of an examination and the cost effectiveness of the procedure have led examining boards to employ single markers for all their examinations and this situation is not likely to change easily.

Holistic scoring

Jacobs *et al.* (1981) offer a different perspective on the various approaches to composition evaluation. They made a primary distinction between holistic scoring and frequency-count marking as against the rather overlapping division into impression and analytic marking used by the body of researchers referred to above. It was based on a classification by Cooper (1977). Jacobs *et al.* (p. 29) described the division as follows: '"Holistic" in Cooper's terms means "any procedure which stops short of *enumerating* linguistic, rhetorical, or informational features of a piece of writing".'

In holistic evaluations, markers base their judgements on their impression of the whole composition: in frequency-count marking (see Steel and Talman, 1936), markers total or enumerate certain elements in the composition such as: cohesive devices, misspelled words, misplaced commas, or sentence errors. Jacobs *et al.* argue that the latter method is highly objective and, therefore, also highly reliable. Not so certain is its validity because a composition evaluated by a frequency-count method has been judged not for its communicative effect, but for its number or kinds of elements.

Holistic evaluation would appear to be more subjective as it depends on the impressions formed by the markers. Jacobs *et al.* (p. 29) point out though:

> In spite of (or perhaps because of) this subjectivity, holistic evaluation has been shown capable of producing highly reliable assessments. Most of the studies cited ... were, in fact, based on holistic evaluation of one type or another and all of those studies obtained reader reliabilities in the mid-to-high eighties or nineties. Intuitively it would seem that composition scores based on holistic responses from readers who attend to the writer's message must be more valid than those based on frequency-count methods, which at best pay only lip service to the writer's meaning and ideas. As Cooper (1977) puts it, 'holistic evaluation by a human respondent gets us closer to what is essential in communication than frequency-counts do'.

Holistic evaluation is obviously to be preferred where the primary concern is with evaluating the communicative effectiveness of candidates' writing. This was the case in the TEEP project (see Weir, 1983a and Appendix I) where the preference was for an analytic, holistic marking scheme over an impressionistic one, favouring an explicit rather than implicit list of features or qualities to guide judgements.

It was felt strongly that too little attention had been paid in the past to the actual criteria to be applied, implicitly or explicitly, to samples of written production. Even in the analytic schemes referred to in the studies above, there is too much room for idiosyncratic interpretation of what constitutes the criterion that is being applied to a script. The application of clear, appropriate criteria was felt to be important.

Chaplen (1970a) had suggested that more reliable results might be obtained from the impression method of marking if the scale employed was one in which each grade was equated with a distinct level of achievement which was closely described. This was the approach initially adopted by the British Council in the ELTS testing system. It may be described as an impression based banding system. An example of such a banded mark scheme can be found in B.J. Carroll (1980b, p. 136).

Carroll's approach is fine in conception as it allows a more detailed description to be presented to institutions. The problem is that, as with Chaplen's (1970a) band system, it fails in practice because it does not cater for learners whose performance levels vary in terms of different criteria. A candidate may be a band 7 in terms of 'fluency', but a band 5 in terms of 'accuracy'. This leaves aside other trenchant criticisms we might have, such as the vagueness of such descriptions as 'authoritative writing', 'good style', 'fluency', etc.

This problem of collapsing criteria is avoided by a more 'analytic' mark scheme, whereby a level is recorded in respect of each criterion and to a certain extent one of the most integrative of measures is brought back somewhat to a discrete point position. This method had the added advantage in that it would lend itself more readily to full profile reporting

and could perform a certain diagnostic role in delineating students' strengths and weaknesses in written production.

Additionally an analytic mark scheme is seen as a far more useful tool for the training and standardisation of new examiners. Francis (1977) pointed out that, by employing an analytic scheme, examining bodies can better train and standardise new markers to the criteria of assessment. A measure of agreement about what each criterion means can be established and subsequently markers can be standardised to what constitutes a different level within each of these criteria. Analytic schemes have been found to be particularly useful with markers who are relatively inexperienced. The data reported by Adams (1981) and Murphy (1982) are consistent with this view.

Analytic mark schemes are devised in an attempt to make the assessment more objective, insofar as they encourage examiners to be more explicit about their impressions. Although one of these criteria may take account of the relevance and adequacy of the actual content of the essays, they are normally concerned with describing the qualities which an essay is expected to exhibit. Brooks (1980) pointed out that the qualities assessed by analytical mark schemes in the past were often extremely elusive. She cited as examples the qualities 'gusto' and 'shapeliness of rhythm' outlined in the *Schools Council Working Paper — Monitoring Grade Standards in English*, as being particularly nebulous and inaccessible to assessment. Thus, although analytic schemes may facilitate agreement amongst examiners as to the precise range of qualities that are to be evaluated in any essay, the actual amount of subjectivity involved in the assessment in many schemes may be reduced very little because of lack of explicitness with regard to the applicable criteria, or through the use of vague criteria.

Establishing appropriate criteria for assessing written production: the test in English for educational purposes (TEEP) experience

The failings of analytic mark schemes in the past have been in the choice and delineation of appropriate criteria for a given situation. In the design work for the TEEP test (see Weir, 1983a and Appendix I) it was felt that the assessment of samples of written performance should be based on appropriate, behaviourally described, analytic criteria, graded according to different levels of performance. The criteria needed to be comprehensive and based on empirical job sample evidence.

The data informing the selection of criteria of assessment came from a survey carried out on language teachers in ARELS schools, and more particularly from the returns to that part of a national questionnaire to academic staff in the United Kingdom which had requested an estimation of the relative importance of the different criteria they employed in assessing the written work of their students. Empirical evidence was gathered from 560 lecturers to help decide upon those criteria which could be used for assessing the types of written information transfer exercises that occur in an academic context.

As a result of the investigation the criteria of relevance and adequacy, compositional organisation, cohesion, referential adequacy, grammatical accuracy, spelling and punctuation were seen as the most suitable for assessing writing tasks. From the returns to the staff questionnaire it appeared there was a need for evaluation procedures that would

assess students, particularly in relation to their communicative effectiveness and in such a way that a profile containing a coarse diagnosis of candidates' strengths and weaknesses could be made available.

To apply these 'valid' criteria reliably an attempt was made to construct an analytic marking scheme in which each of the criteria is sub-divided into four behavioural levels on a scale of 0–3 (see Table below). A level 3 corresponds to a base line of minimal competence. At this level it was felt that a student was likely to have very few problems in coping with the writing tasks demanded of him by his course in respect of this criterion. At a level 2 a limited number of problems arise in relation to the criterion and remedial help would be advisable. A level 1 would indicate that a lot of help is necessary with respect to this particular criterion. A level 0 indicates almost total incompetence in respect of the criterion in question.

TEEP Attribute Writing Scales

A. *Relevance and adequacy of content*

0. The answer bears almost no relation to the task set. Totally inadequate answer.
1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
2. For the most part answers the tasks set, though there may be some gaps or redundant information.
3. Relevant and adequate answer to the task set.

B. *Compositional Organisation*

0. No apparent organisation of content.
1. Very little organisation of content. Underlying structure not sufficiently apparent.
2. Some organisational skills in evidence, but not adequately controlled.
3. Overall shape and internal pattern clear. Organisational skills adequately controlled.

C. *Cohesion*

0. Cohesion almost totally absent. Writing so fragmentary that comprehension of the intended communication is virtually impossible.
1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.
2. For the most part satisfactory cohesion though occasional deficiencies may mean that certain parts of the communication are not always effective.
3. Satisfactory use of cohesion resulting in effective communication.

D. *Adequacy of vocabulary for purpose*

0. Vocabulary inadequate even for the most basic parts of the intended communication.

1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.

2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.

3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.

E. *Grammar*

0. Almost all grammatical patterns inaccurate.

1. Frequent grammatical inaccuracies.

2. Some grammatical inaccuracies.

3. Almost no grammatical inaccuracies.

F. *Mechanical accuracy I (punctuation)*

0. Ignorance of conventions of punctuation.

1. Low standard of accuracy in punctuation.

2. Some inaccuracies in punctuation.

3. Almost no inaccuracies in punctuation.

G. *Mechanical accuracy II (spelling)*

0. Almost all spelling inaccurate.

1. Low standard of accuracy in spelling.

2. Some inaccuracies in spelling.

3. Almost no inaccuracies in spelling.

The set of criteria developed for TEEP and the behavioural descriptions of the levels within each of them are not seen as irrevocable, but they represent the outcome of a long process of practical trialling and revision. In all, the behavioural descriptions of the levels within the criteria went through five major revisions.

The nature of the problems encountered in the evolution of the criteria provide useful background for the development of similar schemes. The first problem in earlier versions of these assessment criteria was that in some of the criteria an attempt was made to assess two things, namely communicative effectiveness and degrees of accuracy. As a result great difficulty was encountered in attempting to apply the criteria reliably. It was necessary

to refine the criteria so that the first four related to communicative effectiveness and the latter three to accuracy. It may well be that the latter three criteria contribute to communicative effectiveness or lack of it, but attempts to incorporate some indication of this into these criteria proved unworkable.

Secondly distinctions between each of the four levels were only gradually achieved and it was also necessary to try and establish roughly equivalent level distinctions across the criteria. Great problems were experienced in the trial assessments in gaining agreement as to what was meant by certain of the descriptions of levels within the criteria. Most sources of confusion were gradually eliminated and this seemed inevitably to result in a much simplified scale for these descriptions of level particularly in the accuracy criteria E–G.

4.3.4 Further considerations in designing writing tasks for inclusion in a test battery

Number of writing tasks

In the discussion of writing so far the concern has been with how marker reliability might be achieved. There are, however, other factors contributing to the reliability of a test which merit attention. Firstly, the number of samples of a student's work that are taken can help control the variation in performance that might occur from task to task.

Both reliability and validity have been found to be increased by sampling more than one composition from each candidate. Finlayson (1951, p. 132) found that: 'the performance of a child in one essay is not representative of his ability to write essays in general.' The research of Vernon and Milligan (1954, p. 69) also threw: 'very grave doubt on the common practice ... of trying to assess English ability in general from a single essay marked by a single examiner.'

Ebel (1972) showed that the more samples there were of a student's writing in a test, the more reliable the result. Ebel outlined how a test score comprised two elements: the true score and the error measurement. He showed (pp. 250–1) that: 'the contribution (i.e. variability) of the true component in the total score is proportional to the number of elements (items) comprising it ... increasing test length increases the true score variance more rapidly than it increases the error variance.' In other words, reliability of a test score tends to increase as the number of items in the test is increased (see Willmott and Nuttal, 1975).

Murphy (1978) also found that an important factor in determining the varying reliability of the eight GCE examinations under review was:

> the number of marks for individual parts contributing to the final examination marks. This effect of increasing reliability, by having more parts of an examination is well demonstrated by the case of English 'A' level. This observation is consistent with the established principle that combinations of unreliable measurements are more reliable than the individual measurements themselves.

Jacobs *et al.* (1981, p. 15) recommended that:

> In general, it is advisable to obtain at least two, if not more, compositions from each student.

> This helps ensure that the test provides a representative sampling of a writer's ability, by reducing to some extent the effects of variation in an individual's performance from topic to topic or from one test period to another Our experience and that of others suggests that two carefully formulated writing tasks are probably sufficient for most testing situations

Obviously the more samples of students' writing that are taken the better this will be for reliability and validity purposes, provided each sample gives a reasonable estimate of the ability

Question choice

As regards selection of topic(s) it is necessary to ensure that students are able to write something on the topic(s) they are presented with Whether this means allowing a choice of topics is an important decision that has to be made, for it too could affect the reliability of the test

Jacobs *et al* (1981, p 1) advised

> For large-scale evaluations, it is generally advisable for all students to write on the same topics because allowing a choice of topics introduces too much uncontrolled variance into the test — i e , are *observed* differences in scores due to *real* differences in writing proficiency or to the different topics? There is no completely reliable basis for comparison of scores on a test unless all of the students have performed the same writing task(s), moreover, reader consistency or reliability in evaluating the test may be reduced if all of the papers read at a single scoring session are not on the same topic

Heaton (1975) suggested that offering a choice means, in addition, that some students may waste time trying to select a topic from several given alternatives Where tests are to be conducted under timed conditions, forcing all students to write on the same topic might also be an advantage for indecisive candidates Jacobs *et al* (1981, p 17) concluded

> In view of the problems associated with offering a choice of topics, the best alternative, unless skill in choosing a topic is among the test objectives, would seem to be to require all students to write on the same topic, but to provide them more than one opportunity to write

By basing writing tasks on written and/or spoken text supplied to the candidates or on non-verbal stimuli, it is possible to ensure that in terms of subject knowledge all start equally, at least in terms of the information available to them All are required to write on the same topic, but they would write on a variety of topics

Amount of time allowed for each writing task the ramifications of time limits

Jacobs *et al* (1981, p 17) pointed to a need to give due consideration to the purpose of the writing test

> Is the test a direct outgrowth of certain learning activities, including perhaps, advance preparation for the test composition (reading certain books or conducting research on an assigned topic, practising with a similar topic or the same mode in class and so forth), or is it an impromptu test, which focuses almost entirely on the composing *product*, rather than the composing *process*?

About the only real life parallel of the closely-timed test is that students may encounter examination essays in their academic courses If we were to replicate reality more closely the test tasks would not be timed at all and students would be allowed maximum opportunity and access to resources for demonstrating their abilities with regard to this construct Considerations such as time constraints, reliability and test security requirements make longer, process-oriented tests impractical for most testing of this kind

Jacobs *et al* (1981, p 17) pointed to some of the ramifications of this distinction

> a closely-timed impromptu test can hardly begin to tap the writer's resources in the whole composing process, other than to require that all of the process skills be compressed into a speeded time frame, with the result resembling only vaguely what writers usually do in processing written discourse It is important to remember this serious limitation of a timed, impromptu test

As regards an appropriate time for completion of product-oriented writing tasks in an actual examination setting, Jacobs *et al* (1981, p 18) argued

> A composition test given in conjunction with a battery of other measures must of necessity be limited in time if the total test time is to be practical and not introduce too much variance due to fatigue in the examinees We have used this thirty minute time limit for composition tests given as part of the Michigan Test and believe this time allowance probably provides most students enough time to produce an adequate sample of their writing ability

They found in their research (p 19) that 'with a thirty-minute composition test students at all but the most basic level of proficiency can generally write about a page or more '

4.3.5 Conclusion

The writing component of any test should concentrate on controlled writing tasks where features of audience, medium, setting and purpose can be more clearly specified Attention needs to be paid to the development of adequate and appropriate scoring criteria and examiners trained and standardised in the use of these

4.4 Testing speaking

Testing speaking ability offers plenty of scope for meeting the criteria for communicative testing, namely that tasks developed within this paradigm should be purposive, interesting and motivating, with a positive washback effect on teaching that precedes the test, interaction should be a key feature, there should be a degree of intersubjectivity among participants, the output should be to a certain extent unpredictable, a realistic context should be provided and processing should be done in real time Perhaps more than in any other skill there is the possibility of building into a test a number of the dynamic characteristics of actual communication (see Section 3 1 above)

Set against this are the enormous practical constraints on the large-scale testing of spoken language proficiency, e g , the administrative costs and difficulties, and the resources

necessary for standardising and paying a large number of examiners. Most GCE Examining Boards in England were said to lose money on every candidate who sits an 'O' level language examination in which there is an oral component.

The problems of assessing speech reliably are even greater than those for assessing writing because the interaction, unless captured on tape or video, is fleeting and cannot be checked later. The essential task for the test designer is to establish clearly what activities the candidate is expected to perform, how far the dynamic communicative characteristics associated with these activities can be incorporated into the test, and what the task dimensions will be in terms of the complexity, size, referential and functional range of the discourse to be processed or produced.

Having established what it is that needs to be tested there is available a range of formats of varying degrees of directness. It embraces more direct types such as the face to face interview and the more indirect multiple choice, pencil and paper tests of speaking ability which can be scored by computer. What follows is a brief review of some of the more useful and potentially valid formats for testing speaking ability.

4.4.1 Verbal essay

The candidate is asked to speak (sometimes directly into a tape recorder) for three minutes on either one or more specified general topics.

Advantages

1. The candidate has to speak at length which enables a wide range of criteria including fluency to be applied to the output. More discrete short questions or posited situations to which the student has to respond often severely limit the range of criteria that are applicable.

Disadvantages

1. The problems associated with the free uncontrolled writing task above apply equally to this type of oral task. The topic specified may not be of interest to the candidate and it is not something we are normally asked to do extempore in real life. At the very least a period is normally given for preparation. Where there is a choice of topic it is difficult to compare performances.
2. The more open-ended the topic, the more successful performance in it might be dependent on background or cultural knowledge and draw upon factors such as imagination or creativity. The greater the deviation of responses from what is expected in terms of content, the more difficult it might be to maintain reliability in assessment.
3. The use of tape recorders for the conduct of this task might be stressful to some candidates. It is not possible to set the candidates at their ease as easily as it is in, say, an interview.

4.4.2 Oral presentation

The candidate is expected to give a short talk on a topic which he has either been asked to prepare beforehand or has been informed of shortly before the test. This is different from the 'Spoken Essay' described above in so far as the candidate is allowed to prepare for the task.

Advantages

1. It is often very effective to get the candidate to talk about himself. In the TEEP Test (an oral test carried out on tape in a language laboratory) this was intended as a warm up exercise, but it was found that the one minute given to the candidate to talk about *specified* features of his personal life provided a good overall indicator of his spoken language proficiency in terms of the criteria used in assessing all the other tasks. What is important in assessing spoken production is eliciting a sufficient sample of a candidate's speech for sensible assessments to be made. This is one technique which permits of this.
2. By integrating the activity with previously heard or previously read texts the oral task can be made to equate realistically with real life tasks that the candidate might have to perform in the target situation.

Disadvantages

1. If the candidate knows the topic well in advance there is a danger that he can learn it by heart. If little time is given for preparation then one faces the problem that what is being tested may be knowledge rather than linguistic ability. If the task is integrated say with a prior reading passage to ensure that all candidates have a common set of information available to them then one is faced with the problem of reading possibly interfering with the measurement.
2. The multiplicity of interpretations of broad topics may create problems in assessment.

4.4.3 The free interview

In this type of interview the conversation unfolds in an unstructured fashion and no set of procedures is laid down in advance.

Advantages

1. Because of its face and content validity in particular, the interview is a popular means of testing the oral skills of candidates.
2. Free interviews are like extended conversations and the direction is allowed to unfold as the interview takes place. The discourse might seem to approximate more closely to the normal pattern of informal social interaction in real life where no carefully formulated agenda is apparent.

Disadvantages

1. As there are no set procedures for eliciting language the performances are likely to vary from event to event not least because different topics may be broached and differences may occur in the way the interview is conducted.
2. The procedure is time consuming and difficult to administer if there are large numbers of candidates.

4.4.4 The controlled interview

In this procedure there are normally a set of procedures determined in advance for eliciting performance. The FSI interview is close to this model (see Adams and Frith, 1979 and Wilds, 1975).

Advantages

1. There is a greater possiblity in this approach of candidates being asked the same questions and thus it is easier to make comparisons across performances.
2. The procedure has a higher degree of content and face validity than most other techniques apart from the role play and information gap exercises in the UCLES/RSA Certificates in Communicative Skills in English (see Appendix III).
3. It has been shown elsewhere that with sufficient training and standardisation of examiners to the procedures and scales employed, reasonable reliability figures can be reached with this technique. Clark and Swinton (1979) report average intra-rater reliabilities of 0.867 and inter-rater reliability at 0.75 for FSI type interviews.
4. A particularly effective oral interview can occur when the candidate is interviewed and assessed by both a language and a subject specialist who have been standardised to agreed criteria. The procedures followed in the General Medical Council's PLAB oral interview which assesses both medical knowledge and spoken English merit consideration in this respect.

Disadvantages

1. One of the drawbacks of the interview is that it cannot cover the range of situations candidates might find themselves in even where the target level performance is circumscribed as in the case of the FSI. In interviews it is difficult to replicate all the features of real life communication such as reciprocity, motivation, purpose and role appropriacy.
2. Even when the procedures for eliciting performance are specified in advance there is still no guarantee that candidates will be asked the same questions in the same manner, even by the same examiner.

4.4.5 Information transfer: description of a picture sequence

The candidate sees a panel of pictures depicting a chronologically ordered sequence of events and has to tell the story in the past tense. Time is allowed at the beginning for the candidate to study the pictures.

Advantages

1. The task required of the candidates is clear. It does not require them to read or listen and thereby avoids the criticism of contamination of measurement provided the pictures are not culturally or educationally biased.
2. It can be an efficient procedure and one of the few available to get the candidate to provide an extended sample of connected speech which allows the application of a wide range of criteria in assessment. It is also useful for eliciting the candidate's ability to use particular grammatical forms such as the past tense for reporting.
3. Because all candidates are constrained by common information provided by pictures or drawings it allows a comparison of candidates which is relatively untainted by background or cultural knowledge given that the drawings themselves are culture free.
4. The value of the technique is dependent on the pictures being clear and unambiguous and free from cultural or educational bias. The technique is straightforward and much favoured by school examination boards in Britain. In the study of suitable formats for a spoken component for TOEFL (Clark and Swinton, 1979) this proved to be one of the most effective formats in the experimental tests.

Disadvantages

1. The authenticity of this task is limited though it could be said to represent the situation of having to describe something which has happened i.e. an informational routine. This may well be an important function in some occupations. It tells us very little, however, about the candidate's ability to interact orally.
2. If the quality of the pictures is in any way deficient then the candidate may not have the opportunity of demonstrating his best performance. Differences in interpretation might also introduce unreliability into the marking.

4.4.6 Information transfer: questions on a single picture

The examiner asks the candidate a number of questions about the content of a picture which he has had time to study. The questions may be extended to embrace the thoughts and attitudes of people in the picture and to discuss future developments arising out of what is depicted.

Advantages

1. There may be considerable benefit in investigating this technique, which already performs a valuable role in the oral component of the PLAB English test for overseas doctors. In PLAB candidates are shown slides, X-rays, pictures of medical conditions, ECG printouts, etc., and are asked to comment on them as well as answering specific questions relating to them. The task could be useful for testing other groups where relevant pictorial material could be developed.

Disadvantages

1. The candidate is cast in the sole role of respondent and is denied the opportunity to ask questions. The criterion of reciprocity, a normal feature of most spoken interaction, is not usually met.
2. The pictures need to be clear and unequivocal for the reasons stated above in discussion of a sequence of pictures. If a large number of candidates are to be examined over several days then the question of test security arises if the same pictures are to be employed. If different pictures are to be used the issue of comparability must be faced.

4.4.7 Interaction tasks

Information gap student–student

In these tasks students normally work in pairs and each is given only part of the information necesary for completion of the task. They have to complete the task by getting missing information from each other. Candidates have to communicate to fill in an information gap in a meaningful situation.

The UCLES/RSA Certificates in Communicative Skills in English have particularly realistic examples of this (see Appendix III). As a development from this interaction an interlocutor appears after the discussion and the candidates might, for example, have to report on decisions taken and explain and justify these decisions.

Advantages

1. There can be few test tasks which represent the act of communication better than this as it fulfils most of the criteria laid down by Morrow (1979) for what makes a test communicative, e.g., it should be reciprocal, purposeful, contextualised and interactive. The candidates should be free to choose their partners so that they are interacting under normal time constraint with somebody they know and feel happy communicating with.
2. As a normal feature of the interaction they can use question forms, elicit information, make requests, ask for clarification and paraphrase in order to succeed in the task, i.e., deploy improvisational as well as interactional skills.
3. The task is highly interactive and as such comes much closer than most other tasks in this section to representing real communication. It recognises the unpredictability of communicative situations and demands an ability to generate original sentences and not simply an ability to repeat rehearsed phrases.

Disadvantages

1. There is a problem if one of the participants dominates the interaction as the other candidate may have a more limited opportunity to demonstrate communicative potential.
2. Similarly if there is a large difference in proficiency between the two this may influence performance and the judgements made on it.
3. There is also a problem if one of the candidates is more interested in the topic or the task as the interaction may become one sided as a result.
4. Candidates are being assessed on their performance in a single situation and extrapolations need to be made about their ability to perform in other situations from this.
5. There are also practical constraints such as the time available, the difficulties of administration and the maintenance of test security.

Information gap student–examiner

To avoid the possibility of an imbalance in candidates' contributions to the interaction some boards have the examiner as one of the participants or employ a common interlocutor, e.g., a familiar teacher with whom candidates would feel comfortable.

To examine candidates separately they can be given a diagram, a set of notes, etc., from which information is missing and their task is to request the missing information from the examiner.

Advantages

1. The main advantage is that there is a stronger chance that the interlocutor will react in a similar manner with all candidates allowing a more equitable comparison of their performance.

Disadvantages

1. Interacting with a teacher, let alone an examiner, is often a more daunting task for the candidate than interacting with his peers.
2. There is some evidence that where the examiner is a participant in the interaction, he is sometimes inadvertently assessing his own performance in addition to that of the candidate (Fisher 1979).

4.4.8 Role play

A number of examining boards, for example the AEB and UCLES/RSA, include role play situations where the candidate is expected to play one of the roles in an interaction which might be reasonably expected of him in the real world. The interaction can take place between two students or, as in the GCE mould, the examiner normally plays one of the parts. The disadvantage of the latter is that it is difficult to make an assessment at the

same time as taking part in the interaction. (The same is of course true for all face to face tests.) As in the information gap exercise involving teacher as interlocutor and examiner there is a danger that the mark awarded will reflect the latter's view of his own performance as well as of the student's.

Advantages

1. The technique can be valid in both face and content terms for a wide variety of situations and the experience of the examination boards suggests that it is a practical and potentially a highly valid and reliable means of assessing a candidate's ability to participate effectively in oral interaction.

Disadvantages

1. There is a danger that the histrionic abilities of some candidates may weigh in their favour at the expense of the more introverted. There is also the danger in all oral interactions that a candidate cannot think what to say. The question of role familiarity arises in this technique in the sense that some candidates may not know what it is normal to do in certain situations. Another problem is that candidates often use the language of reporting and say what they would say rather than directly assuming the role.

2. Practical constraints operate here as well, especially in large scale testing operations. If it is necessary to use different role plays then great care needs to be taken to ensure that they are placing equal demands on candidates.

4.4.9 The training and standardisation of oral examiners

The relationship between a task and the criteria that can be applied to the product it results in is an essential factor in taking decisions on what to include in a test of spoken or written production. Tasks can not be considered separately from the criteria that might be applied to the performances they result in. Having established suitable tasks and appropriate assessment criteria to accompany them, consideration needs to be given as to how best to apply the criteria to the elicited samples of performance.

In Section 4.3.3 the advantages of analytical as against impressionistic approaches to assessment of written production were examined. The comments made there in relation to the need to establish clear criteria for assessment and to standardise examiners to these criteria apply *mutatis mutandis* to the assessment of oral ability.

The assessment of spoken language is potentially more problematic, though, given that no recording of the performance is usually made. Whereas in writing the script can be reconsidered as often as is necessary, assessments have to be made in oral tests either while the performance is being elicited or shortly afterwards. If the examiner is also an interlocutor then the problems are further compounded.

In oral testing, as in the assessment of written production, there is a need for explicit, comprehensive marking schemes, close moderation of test tasks and mark schemes, and rigorous training and standardisation of markers in order to boost test reliability. These

aspects of examining will be briefly discussed below. Though made in the context of oral testing, they apply equally to the testing of the other skills.

Marking schemes

Murphy (1979, p. 19) outlined the nature of the marking scheme demanded by the Associated Examining Board: 'A marking scheme is a comprehensive document indicating the explicit criteria against which candidates' answers will be judged: it enables the examiner to relate particular marks to answers of specified quality.'

In assessing the productive skills it is necessary to provide comprehensive descriptions of levels of performance to aid the examiners in making necessarily subjective judgements about the work of candidates' answers (see Appendix II for an example of such performance descriptors). Murphy (1979, p. 14) described the purpose of the marking scheme as follows:

1. To assist the Chief Examiner and those who will moderate the paper to check the content validity of the tasks which are being set.

2. To help the moderators to check that the demands made in the examination are appropriate and in accordance with stated aims and objectives.

3. To ensure that, where there is more than one examiner, each examiner marks in exactly the same way, awarding equal marks for equal levels of performance.

4. To ensure that each examiner marks consistently throughout the marking period.

The moderation of question papers and mark schemes

In assessing both speaking and writing it is essential that tasks and marking schemes are subjected to a rigorous process of moderation before they become operations. Again using Murphy as the informing source, examples of questions which might be asked of the tasks candidates are expected to perform are listed below:

1. Are the tasks set at an appropriate level of difficulty?

2. Will the paper discriminate adequately between the performance of candidates at different levels of attainment?

3. Does the test assess the full range of appropriate skills and abilities, as defined by the objectives of the examination?

4. Are the tasks unambiguous, giving a clear indication of what the examiner is asking, so that no candidate may take the task to mean something different?

5. Is there an excessive overlap in enabling skills or communicative tasks being assessed in different parts of the test?

6. Can the tasks be satisfactorily answered in the time allowed?

In addition the moderating panel might consider the format and the layout of the question

papers. This is important because a badly laid out question paper could be the cause of considerable problems for both candidates and examiners. Instructions to candidates need to be as clear and as concise as possible.

At the same time as the tasks are moderated the panel should consider the appropriateness of the marking scheme. Murphy (1979) again is a valuable informing source for drawing up a list of questions which might be asked of the marking scheme by a moderating panel. The following are examples of the types of questions moderators might address themselves to:

1. Does the mark scheme anticipate responses of a kind that candidates are likely to make?
2. Are the marks allocated to each task commensurate with the demands that task makes on the candidate?
3. Does the mark scheme indicate clearly the marks to be awarded for different parts of a question or the relative weighting of criteria that might be applicable?
4. Does the mark scheme allow for possible alternative answers?
5. Has the mark scheme reduced to the minimum possible, the amount of computational work which the examiner has to undertake to finalise a mark for a candidate's performance?
6. Does the marking scheme, by specifying performance criteria (see Appendix II), reduce as far as possible the element of subjective judgement that the examiner has to exercise in evaluating candidates' answers?
7. Are the abilities being rewarded those which the tasks are designed to assess?
8. Can the marking schemes be easily interpreted by a number of different examiners in a way which will ensure that all mark to the same standard?

The standardisation of marking

Even if examiners are provided with an ideal marking scheme, there might always be some who do not mark in exactly the way required. The purpose of standardisation procedures is to bring examiners into line, so that candidates' marks are affected as little as possible by the particular examiner who assesses them.

Examiners are normally requested to attend a standardisation meeting prior to commencing marking proper. Here the marking criteria are discussed to ensure the examiners understand the criteria that are to be applied. The marking scheme is examined closely and any difficulties in interpretation are ironed out. At this meeting the examiners have to conduct a number of assessments. In respect of an oral test this might involve listening to and/or watching audio tape or video tape recordings of candidates' performances on the test, at a number of different levels.

The examiners are asked to assess these performances and afterwards these are compared to see if they are applying the same marking standards. The aim is to identify any factors which might lead to unreliability in marking and to try and resolve these at the meeting.

In the case of new examiners there might also be extensive discussion with the chief examiner and his deputies about how to conduct an examination, how to make the candidate feel at ease, how to phrase questions, what questions to ask in an interview situation and how to react to candidates.

After the standardisation procedure examiners are allowed to assess candidates. In tests of writing it is possible to ask for sample scripts to be sent in to be re-marked by the chief examiner so that the process of ensuring reliability can be continued. In oral tests the chief examiner may sometimes sit in unannounced on a number of oral tests, observe how they are conducted and discuss the allocation of marks with the examiner subsequently. They might discuss how improvements could be made either in technique or marking.

Profiling ability in the productive and receptive skills

In terms of the criteria for assessing the productive modes of speaking and writing it seems that we are able to be far more explicit than is possible in the receptive skills. We can develop criteria of assessment, write behavioural descriptions of levels within each of the criteria and then apply these to samples of students' speech and writing. Because of the private, internalised nature of the reading and listening processes it does not seem possible to devise such explicit criteria by which candidates' proficiency in the receptive skills can be judged. Whereas we are able to assess candidates' productive ability directly, we can only take indirect measurements of what we label as listening or reading ability. We also have to make the assumption that the sum of the listening or reading skills being measured is equivalent to the whole of what might be described as proficiency in listening or reading.

Though an attempt can be made to specify what each item is testing in these receptive areas, the candidates' responses can only be judged right or wrong. It does not seem possible to establish levels of attainment on individual receptive skills items, however explicit we can be about what an individual item is testing. Whereas in writing or speaking tests we are presented with something more tangible to make qualitative judgements about, it is more difficult to see at what stage the process might have broken down in items testing the receptive skills or to employ anything other than a dichotomous rating scale. We cannot normally say how near a candidate came to getting an item right in assessment of receptive skills and the more discrete the item the more this must be the case.

Where the intention is to present a profile of proficiency in the macro-skills of reading, listening, speaking and writing, the differences in the way we are able to assess these will obviously affect the manner in which we are able to decide how overall grades are to be awarded in each macro-skill (see Angoff, 1971; Cresswell, 1983; Houston, 1983; Weir, 1983a; and Zieky and Livingston, 1977). The process is likely to be that much easier for the productive skills where explicit criteria can be applied directly to a concrete, integrated product.

4.5 Integrated tests

Relatively little is known about integrated tests, and even less is reported in the literature,

with the exceptions of preliminary work in this area by Weir (1983a), Emmings (1986) and Low (1986). Low describes a variety of possible tests along a continuum of directness. At one end there are tests such as TEEP where there is an attempt to include tasks which involve an integration of different macro skills. Weir (1983a) describes the integrated activity in TEEP, where listening and/or reading texts, as well as providing a stimulus for more discrete testing of skills, are also employed as the stimulus material for a writing test (see Appendix I). At the other end of the continuum there are attempts to develop a 'story line' through the whole of the test so that not only are the components of the test related thematically through the content but there is also a development in the content itself throughout the test (see Low, 1986).

These tests are able to tap performance ability more directly by the provision of realistic contexts which simulate target situations more closely. It is argued (see Emmings, 1986) that they offer a better means of assessing the traits underlying language proficiency than is possible if the skills are tested in relative isolation from each other. Low (1986) points to the context effects in real life where the ability to perform for example in conversation 'depends in large measure on the skill with which one picks up information from past discussions and uses it as an aid to formulating new strategy.'

Emmings (1986) investigated the incorporation of contextual developments through integrated testing procedures and compared the reliability and validity of these with more discrete proficiency measures taken from an RSA CUEFL test. While recognising the importance of context for the development of an interaction, he sought to establish whether the enhanced validity of a story line development through sub-tests is negated by the dangers of 'muddied measurement' adversely affecting candidate profiling in an integrated test of this type. He found that:

> in principle, the adoption of clear content criteria can produce a clear reflection of the aims of the test items in the factorial structures and that therefore some indications of testees' abilities to process text can be provided by integrated tests.

He advises caution in employing a purely integrated approach but argues that the findings with regard to validity and reliability were encouraging.

Low (1986) offers a detailed rationale for the use of an integrated approach and makes a number of suggestions for the use of story lines and other developing contexts in tests. He also draws attention to a number of problems involved in the design of integrated tests, such as lower reliability, the possibility of content bias, greater difficulty in design and implementation, and their radical nature as compared to the curricula in operation in certain countries.

Advantages

1. On the grounds of authenticity, or approximations to it, integrated tasks demand consideration. If tests are to simulate reality, as closely as is possible, recognition of the integrated nature of activities in certain contexts is necessary. In academic life, for example, students have to read a variety of texts and process spoken discourse in lectures and seminars in preparation for writing tasks, and work is often discussed with others before a final version is presented for assessment.

2. Increasingly the importance of context in language use is recognised and with it there has been a demand that test tasks should be improved in this respect (see Stansfield, 1986). Low's (1986) advocacy of employing a developing context through the use of a story line across tasks is an attempt to provide for this desire for increased validity.

Disadvantages

1. The question arises as to how integrated a test can ever be; how close can it mirror reality? If Low's (1986) examples of integrated tests are examined carefully, even the most direct of these evidence a good deal of idealisation. The chimera of full authenticity is just that.
2. The dangers of 'muddied measurement' are often referred to in the literature. This relates to a concern about the local independence of items or of tasks within a test battery. The feeling is that performance on one item should not interfere with performance on a subsequent item. In terms of tasks, it is felt, for example, that performance on a writing task should not be dependent on successful performance in coping with prior reading and listening tasks. The profiling of abilities may be problematic if there is difficulty in determining where the process has broken down. If there is a need for separate skills profiling, more discrete test tasks may be required.

What is clear is that communicative tests will be more difficult to construct than traditional measures. As well as the need for greater explicitness about what it is that one is trying to test there are serious problems in successfully realising communicative specifications in a test form and in devising suitable assessment procedures. As with all new departures integrated communicative tests are at present difficult to construct, complex to take, difficult to mark and difficult to report results on.

Too little is known at present as to whether the advantage of enhanced validity gained by using an integrated format is outweighed by the possible contaminating influence on test scores. It may well be that the latter can be avoided through careful design but at present any claims for integrated tests being the panacea for testing within a communicative paradigm must be tentative. In the customary fashion we can only point to the need for future research to shed light on this.