ELT documents

111- Issues in Language Testing

The British Council

ELT documents
111- Issues in Language Testing

Editors: J Charles Alderson
Arthur Hughes

The British Council
Central Information Service
English Language and Literature Division

ELT Documents is now including a correspondence section. Comments arising from articles in current issues will therefore be most welcome. Please address comments to ELSD, The British Council, 10 Spring Gardens, London SW1A 2BN.

ISBN 0 901618 51 9

CONTENTS

concern over these particular points. Although it was hoped to include responses from the authors of the original articles only one response was available at the time of going to press, that of Helmut Vollmer. Nevertheless, it is hoped that subsequent debate will include the responses and further thoughts of the other authors in the light of these discussions.

This is not a definitive volume on language testing – and it does not attempt to be such. What this book hopes to do is to encourage further debate, a critical or sceptical approach to claims made about 'progress' and 'theories', and to encourage practical research in important areas.

It has not been the intention of this Introduction to guide the reader through the discussions – that would have been presumptuous and unnecessary – but rather to set the scene for them. Thus there is here no summary of positions taken, arguments developed and issues raised. However, there is, after the three main sections, an Epilogue, and the reader is advised not to ignore this: it is intended, not to tell the reader what he has read, but to point the way forward in the ongoing debate about the assessment of language learning. 'Testing' should not and cannot be left to 'Testers': one of the most encouraging developments of the last decade is the involvement of more applied linguists in the area of assessment and evaluation. In a sense, there can be no Epilogue, because the debate is unfinished, and we hope that participation in the debate will grow. It is ultimately up to the reader to write his own 'Way Forward'.

Thanks are due to all Symposium participants, not only for their contributions, written and spoken, to the Symposium, but also for their help in preparing this volume. Thanks are also due to the Institute for English Language Education, Lancaster, for hosting the Symposium and contributing materially to the preparation of this book.

J Charles Alderson,
University of Lancaster

SECTION 1

COMMUNICATIVE LANGUAGE TESTING: REVOLUTION OR EVOLUTION?[1]

Keith Morrow, Bell School of Languages, Norwich

Introduction

Wilkins (1976) concludes with the observation that, 'we do not know how to establish the communicative proficiency of the learner' and expresses the hope that, 'while some people are experimenting with the notional syllabus as such, others should be attempting to develop the new testing techniques that should, ideally, accompany it' (*loc cit*). In the two years that have passed since the publication of this book, the author's hope on the one hand has been increasingly realised, and if his observation on the other is still valid, there are grounds for believing that it will not be so for much longer.

At the time of writing, it is probably true to say that there exists a considerable imbalance between the resources available to language teachers (at least in E F L) in terms of teaching materials, and those available in terms of testing and evaluation instruments. The former have not been slow to incorporate insights into syllabus design, and increasingly methodology, deriving from a view of language as communication; the latter still reflect, on the whole, ideas about language and how it should be tested which fail to take account of these recent developments in any systematic way.[2]

This situation does seem to be changing, however. A number of institutions and organisations have set up working parties to assess the feasibility of tests based on communicative criteria, and in some cases these have moved on to

---

[1] This article was first published in *The Communicative approach to language teaching* ed: C J Brumfit and K Johnson. Oxford University Press, 1979. Reprinted here by kind permission of Oxford University Press.

[2] Exceptions to this are the two oral examinations promoted by the Association of Recognised English Language Schools: The ARELS Certificate and the ARELS Diploma, as well as the Joint Matriculation Board's Test in English for Overseas Students. But without disrespect to these, I would claim that they do not meet in a rigorous way some of the criteria established later in this paper.

the design stage.[3] It therefore seems reasonable to expect that over the next five years new tests and examinations will become available which will aim to do precisely the job which Wilkins so recently held up as a challenge, ie to measure communicative proficiency.

This paper, then, will be concerned with the implications for test design and construction of the desire to measure communicative proficiency, and with the extent to which earlier testing procedures need to be reviewed and reconsidered in the light of this objective. But it is a polemical paper. The assumption which underlies it is that the measurement of communicative proficiency is a job worth doing, and the task is ultimately a feasible one.

The Vale of Tears

A wide range of language tests and examinations are currently in use but most belong to a few key types. Spolsky (1975) identifies three stages in the recent history of language testing: the pre-scientific, the psychometric-structuralist, and the psycholinguistic-sociolinguistic. We might characterise these in turn as the Garden of Eden, the Vale of Tears and the Promised Land, and different tests (indeed different parts of the same test) can usually be seen to relate to one or other of these stages. The historical perspective offered by Spolsky is extremely relevant to the concerns of this paper. While critiques of the 'prescientific' approach to testing are already familiar (Valette, 1967), it seems useful to take some time here to clarify the extent to which current developments relate to what has more immediately gone before through a critical look at some of the characteristics of psychometric-structuralist testing. The point of departure for this is Lado (1961).

Atomistic

A key feature of Lado's approach is the breaking down of the complexities of language into isolated segments. This influences both what is to be tested and how this testing should be carried out.

What is to be tested is revealed by a structural contrastive analysis between the target language and the learner's mother tongue. Structural here is not limited to grammatical structure – though this is of course important.

---

[3] My own work in this field has been sponsored by the Royal Society of Arts who have established a Working Party to re-design their range of examinations for foreign students. The English Language Testing Service of the British Council is developing communicative tests in the area of English for Academic Purposes, and a similar line is likely to be followed soon by the Associated Examining Board.

Contrastive analysis can be carried out of all the levels of structure (syntactic down to phonological) which the language theory encompasses, and test items can be constructed on the basis of them.

The same approach is adopted to the question of how to test. Discrete items are constructed, each of which ideally reveals the candidate's ability to handle one level of the language in terms of one of the four skills. It soon became recognised that it was in fact extremely difficult to construct 'pure' test items which were other than exceedingly trivial in nature, and thus many tests of this sort contain items which operate on more than one level of structure.

The clear advantage of this form of test construction is that it yields data which are easily quantifiable. But the problem is equally clearly that its measurement of language proficiency depends crucially upon the assumption that such proficiency is neatly quantifiable in this way. Indeed the general problem with Lado's approach, which attaches itself very firmly to certain very definite views about the nature of language, is that it crumbles like a house of cards as soon as the linguistic foundation on which it is constructed is attacked. This is not the place to develop a generalised linguistic attack, but one particular assumption is worth picking up, since it is so central to the issue under discussion.

An atomistic approach to test design depends utterly on the assumption that knowledge of the elements of a language is equivalent to knowledge of the language. Even if one adopts for the moment a purely grammatical view of what it is to know a language (cf Chomsky's definition in terms of the ability to formulate all and only the grammatical sentences in a language), then it seems fairly clear that a vital stage is missing from an atomistic analysis, viz the ability to synthesise. Knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language. Driving a car is a skill of a quite different order from that of performing in isolation the various movements of throttle, brake, clutch, gears and steering wheel.

Quantity v. Quality

In the previous section it was the linguistic basis of tests such as Lado's which was questioned. Let us now turn to the psychological implications. Following the behaviourist view of learning through habit formation, Lado's tests pose questions to elicit responses which show whether or not correct habits have been established. Correct responses are rewarded and negative ones punished in some way. Passing a test involves making a specified proportion of correct responses. Clearly language learning is viewed as a process of accretion.

An alternative view of the psychology of language learning would hold, however, that the answers to tests can, and should, be considered as more than simply right or wrong. In this view learners possess 'transitional competence' (Corder, 1975) which enables them to produce and use an 'interlanguage' (Selinker, 1972). Like the competence of a native speaker, this is an essentially dynamic concept and the role of the test is to show how far it has moved towards an approximation of a native speaker's system. Tests will thus be concerned with making the learner produce samples of his own 'interlanguage', based on his own norms of language production so that conclusions can be drawn from it. Tests of receptive skills will similarly be concerned with revealing the extent to which the candidate's processing abilities match those of a native speaker.

The clear implication of this is that the candidate's responses need to be assessed not quantitatively, but qualitatively. Tests should be designed to reveal not simply the number of items which are answered correctly, but to reveal the quality of the candidate's language performance. It is not safe to assume that a given score on the former necessarily allows conclusions to be drawn about the latter.

Reliability

One of the most significant features of psychometric tests as opposed to those of 'pre-scientific' days is the development of the twin concepts of reliability and validity.

The basis of the reliability claimed by Lado is objectivity. The rather obvious point has, however, not escaped observers (Pilliner, 1968; Robinson, 1973) that Lado's tests are objective only in terms of actual assessment. In terms of the evaluation of the numerical score yielded, and perhaps more importantly, in terms of the construction of the test itself, subjective factors play a large part.

It has been equally noted by observers that an insistence on testing procedures which can be objectively assessed has a number of implications for the data yielded. Robinson (*op cit*) identifies three areas of difference between testing procedures designed to yield data which can be objectively assessed and those which are open to subjective assessment.

1 The amount of language produced by the student. In an objective test, students may actually produce no language at all. Their role may be limited to selecting alternatives rather than producing language.

2 Thus the type of ability which is being tested is crucially different. In a subjective test the candidate's ability to produce language is a crucial factor; in an objective test the ability to recognise appropriate forms is sufficient.

3 The norms of language use are established on different grounds. In an objective test the candidate must base his responses upon the language of the examiner; in a subjective test, the norms may be his own, deriving from his own use of the language. Thus an objective test can reveal only differences and similarities between the language norms of the examiner and candidate; it can tell us nothing of the norms which the candidate himself would apply in a use situation.

The above factors lead to what Davies (1978) has called the reliability-validity 'tension'. Attempts to increase the reliability of tests have led test designers to take an over-restrictive view of what it is that they are testing.

Validity

The idea that language test designers should concern themselves with validity – in other words that they should ask themselves whether they are actually testing what they think they are testing, and whether what they think they are testing is what they ought to be testing – is clearly an attractive one. But unfortunately, because of the 'tension' referred to above, designers working within the tradition we are discussing seem to have been content with answers to these questions which are less than totally convincing.

Five types of validity which a language test may claim are traditionally identified (cf Davies, 1968).

Face	the test looks like a good one.
Content	the test accurately reflects the syllabus on which it is based.
Predictive	the test accurately predicts performance in some subsequent situation.
Concurrent	the test gives similar results to existing tests which have already been validated.
Construct	the test reflects accurately the principles of a valid theory of foreign language learning.

Statistical techniques for assessing validity in these terms have been developed to a high, and often esoteric level of sophistication. But unfortunately, with two exceptions (face, and possibly predictive) the types of validity outlined above are all ultimately circular. Starting from a certain set of assumptions

about the nature of language and language learning will lead to language tests which are perfectly valid in terms of these assumptions, but whose value must inevitably be called into question if the basic assumptions themselves are challenged. Thus a test which perfectly satisfies criteria of content, construct or concurrent validity may nonetheless fail to show in any interesting way how well a candidate can perform in or use the target language. This may occur quite simply if the construct of the language learning theory, and the content of the syllabus are themselves not related to this aim, or if the test is validated against other language tests which do not concern themselves with this objective. There is clearly no such thing in testing as 'absolute' validity. Validity exists only in terms of specified criteria, and if the criteria turn out to be the wrong ones, then validity claimed in terms of them turns out to be spurious. *Caveat emptor.*

Comments

This criticism, implicit and explicit, made in the preceding sections applies to a theory of testing which has hardly ever been realised in the extreme form in which Lado presented it. Certainly in the UK., a mixture of pragmatism and conservatism has ensured that much of the institutionalised testing of foreign languages owes as much to the 1920's as to the 1960's. This does not mean though, that there is anything chimerical about the ideas put forward by Lado. Their influence has been recognised by writers on language testing ever since the first publication of his book. But it is as representation of theory that the ideas are most significant. In practice, as Davies (1978) remarks, there is very often a gap between what Lado himself does and what he says he does.

But this gap is often of detail rather than principle. Even if the totality of Lado's views have been more often honoured in the breach than in the observance, the influence of his work has been tremendous. Of the ideas examined above, very few have failed to find implicit acceptance in the majority of 'theory-based' tests developed over the last fifteen years. The overriding importance of reliability (hence the ubiquitous multiple-choice), the acceptance of validity of a statistical rather than necessarily of a practical nature, the directly quantifiable modes of assessment – these are all ideas which have become common currency even among those who would reject many of the theories of language and language learning on which Lado based his approach.

Only in one area has a consistent alternative to Lado's views been argued, and that is the development of 'integrated' tests/test items[4] as opposed to Lado's arguments (at least in principle) in favour of 'pure' discrete items.[5] A clear statement of an 'integrated' position is made by Carroll (1968):

> '. . . since the use of language in ordinary situations call upon all these aspects [of language], we must further recognise that linguistic performance also involves the individual's capability of mobilizing his linguistic competences and performance abilities in an integrated way, ie in the understanding, speaking, reading or writing of connected discourse.'

This implies a view of language which runs directly counter to a key assumption which we have earlier examined in Lado's work. It denies the atomistic nature of language as a basis for language testing. To this extent, Carroll's contribution is extremely important, but even here it must be observed that in practical terms he was doing no more than providing a post-hoc rationalisation. For the purely practical reasons alluded to earlier, very few 'pure' items had found their way into tests; in a sense, Carroll was merely legitimising the existing situation.

Less casuistically, it must be observed that attempts to develop more revolutionary integrated tests (Oller, 1971, 1973) have left out of account a crucial element in the original formulation, viz. 'the use of language in ordinary situations'.

Both cloze and dictation are fundamentally tests of language competence. Both have their uses in determining the basic level of language proficiency of a given candidate. (More accurately, they enable the level of language proficiency to be assessed relative to that of other people who take exactly the same test under the same conditions.) Oller claims that both test basic language processing mechanisms (analysis by synthesis); both sample a wide range of structural and lexical items in a meaningful context. But neither

---

[4] Note that the word 'integrated' is used in different ways by different writers. For some it is possible to conceive of individual items which test integration of various elements of the language; for others the very isolation of separate items means that full integration is not being achieved.

[5] Earlier it was implied that Lado himself very rarely used items of a totally pure kind. See Davies (1978) for an interesting discussion of integrated v. discrete-point testing. Davies argues that they are at different ends of the same continuum rather than in different universes.

gives any convincing proof of the candidate's ability to actually use the language, to translate the competence (or lack of it) which he is demonstrating into actual performance 'in ordinary situations', ie actually using the language to read, write, speak or listen in ways and contexts which correspond to real life.

Adopting this 'use' criterion might lead us to consider precisely why neither discrete-point nor integrative tests of the type we have considered are able to meet it.

Let us look in a rather simple way at some of the features of language use which do not seem to be measured in conventional tests.

Interaction – Based: in the vast majority of cases, language in use is based on an interaction. Even cases such as letter writing, which may seem to be solitary activities, can be considered as weak forms of interaction in that they involve an addressee, whose expectations will be taken into account by the writer. These expectations will affect both the content of the message and the way in which it is expressed. A more characteristic form of interaction, however, is represented by face-to-face oral interaction which involves not only the modification of expression and content mentioned above but also an amalgam of receptive and productive skills. What is said **by** a speaker depends crucially on what is said **to** him.

Unpredictability: the apparently trivial observation that the development of an interaction is unpredictable is in fact extremely significant for the language user. The processing of unpredictable data in real time is a vital aspect of using language.

Context: any use of language will take place in a context, and the language forms which are appropriate will vary in accordance with this context. Thus a language user must be able to handle appropriacy in terms of:

context of situation	eg	physical environment role/status of participants attitude/formality
linguistic context	eg	textual cohesion

Purpose: a rather obvious feature of communication is that every utterance is made for a purpose. Thus a language user must be able to recognise why a certain remark has been addressed to him, and be able to encode appropriate utterances to achieve his own purposes.

Performance: What Chomsky (1965) described as 'competence', leaving out of account:

> 'such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic)'

has been the basis of most language tests. Such conditions may or may not be 'grammatically irrelevant', but they certainly exist. To this extent the idealised language presented in listening tests fails to measure the effectiveness of the candidate's strategies for receptive performance. Similarly, the demand for context-free language production fails to measure the extent to which features of the candidate's performance may in fact hamper communication.

Authenticity: a very obvious feature of authentic language should be noted in this context, ie with rare exceptions it is not simplified to take account of the linguistic level of the addressee. Thus measuring the ability of the candidate to, eg read a simplified text tells us nothing about his actual communicative ability, since an important feature of such ability is precisely the capacity to come to terms with what is unknown.

Behaviour-Based: the success or failure of an interaction is judged by its participants on the basis of behavioural outcomes. Strictly speaking no other criteria are relevant. This is an extreme view of the primacy of content over form in language and would probably be criticised by language teachers. Nevertheless, more emphasis needs to be placed in a communicative context on the notion of behaviour. A test of communication must take as its starting point the measurement of what the candidate can actually achieve through language. None of the tests we have considered have set themselves this task.

These then are some of the characteristics of language in use as communication which existing tests fail to measure or to take account of in a systematic way. Let us now turn to an examination of some of the implications of building them into the design specification for language tests.

The Promised Land

We can expect a test of communicative ability to have at least the following characteristics:

1 It will be criterion-referenced against the operational performance of a set of authentic language tasks. In other words it will set out to show whether or not (or how well) the candidate can perform a set of specified activities.

2 It will be crucially concerned to establish its own validity as a measure of those operations it claims to measure. Thus content, construct and predictive validity will be important, but concurrent validity with existing tests will not be necessarily significant.

3 It will rely on modes of assessment which are not directly quantitative, but which are instead qualitative. It may be possible or necessary to convert these into numerical scores, but the process is an indirect one and recognised as such.

4 Reliability, while clearly important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration, although it is recognised that in certain situations test formats which can be assessed mechanically will be advantageous. The limitations of such formats will be clearly spelt out, however.

Designing a test with these characteristics raises a number of interesting issues.

Performance Tests

Asking the question, 'What can this candidate do?' clearly implies a performance-based test. The idea that performance (rather than competence) is a legitimate area of concern for tests is actually quite a novel one and poses a number of problems, chiefly in terms of extrapolation and assessment. If one assesses a candidate's performance in terms of a particular task, what does one learn of his ability to perform other tasks? Unless ways of doing this in some effective way can be found, operational tests which are economical in terms of time are likely to run the risk of being trivial. Problems of assessment are equally fundamental. Performance is by its very nature an integrated phenomenon and any attempt to isolate and test discrete elements of it destroys the essential holism. Therefore a quantitative assessment procedure is necessarily impractical and some form of qualitative assessment must be found. This has obvious implications for reliability.

Given these problems, the question obviously arises as to whether communicative testing does necessarily involve performance tests. This seems to depend on what the purpose of the test is. If the purpose is proficiency testing, ie if one is asking how successful the candidate is likely to be as a user of the language in some general sense, then it seems to be incontrovertible that performance tests are necessary. The reasons for saying this should by now be clear, but at the risk of labouring the point let me re-state the principle that in language use the whole is bigger than the parts. No matter how sophisticated the analysis of the parts, no matter whether the parts are

isolated in terms of structures, lexis or functions, it is implausible to derive hard data about actual language performance from tests of control of these parts alone. However, if the test is to be used for diagnostic purposes rather than proficiency assessment, a rather different set of considerations may apply. In a diagnostic situation it may become important not simply to know the degree of skill which a candidate can bring to the performance of a particular global task, but also to find out precisely which of the communicative skills and elements of the language he has mastered. To the extent that these can be revealed by discrete-point tests and that the deficiencies so revealed might form the input to a teaching programme, this might be information worth having. (The form that such tests might take is discussed in Morrow, 1977.) But one more point must be made. It might be argued that discrete-point tests of the type under discussion are useful as achievement tests, ie to indicate the degree of success in assimilating the content of a language learning programme which is itself based on a communicative (notional) syllabus. This seems to me misguided. As a pedagogic device a notional syllabus may specify the elements which are to be mastered for communicative purposes. But there is little value in assimilating these elements if they cannot be integrated into meaningful language performance. Therefore discrete-point tests are of little worth in this context.

The clear implication of the preceding paragraphs is that by and large it is performance tests which are of most value in a communicative context. The very real problems of extrapolation and assessment raised at the beginning of this section therefore have to be faced. To what extent do they oblige us to compromise our principle?

Let us deal first with extrapolation. A model for the performance of global communicative tasks may show for any task the enabling skills which have to be mobilised to complete it. Such a model is implicit in Munby (1978) and has been refined for testing purposes by B J Carroll (1978). An example of the way this might work is as follows:

Global Task

Search text for specific information

Enabling Skills

eg Distinguish main point from supporting details
Understand text relations through grammatical cohesion devices
Understand relations within sentences
Understand conceptual meaning
Deduce meaning of unfamiliar lexis

The status of these enabling skills *vis-à-vis* competence:performance is interesting. They may be identified by an analysis of performance in operational terms, and thus they are clearly, ultimately performance-based. But at the same time, their application extends far beyond any one particular instance of performance and in this creativity they reflect an aspect of what is generally understood by competence. In this way they offer a possible approach to the problem of extrapolation.

An analysis of the global tasks in terms of which the candidate is to be assessed (see later) will usually yield a fairly consistent set of enabling skills. Assessment of ability in using these skills therefore yields data which are relevant across a broad spectrum of global tasks, and are not limited to a single instance of performance.

While assessment based on these skills strictly speaking offends against the performance criterion which we have established, it should be noted that the skills are themselves operational in that they derive from an analysis of task performance. It is important that the difference between discrete-point tests of these enabling skills and discrete-point tests of structural aspects of the language system is appreciated.

Clearly, though, there exists in tests of enabling skills a fundamental weakness which is reminiscent of the problem raised in connection with earlier structural tests, namely the relationship between the whole and the parts. It is conceivable that a candidate may prove quite capable of handling individual enabling skills, and yet prove quite incapable of mobilising them in a use situation or developing appropriate strategies to communicate effectively. Thus we seem to be forced back on tests of performance.

A working solution to this problem seems to be the development of tests which measure both overall performance in relation to a specified task, and the strategies and skills which have been used in achieving it. Written and spoken production can be assessed in terms of both these criteria. In task-based tests of listening and reading comprehension, however, it may be rather more difficult to see just how the global task has been completed. For example, in a test based on the global task exemplified above and which has the format of a number of true/false questions which the candidate has to answer by searching through a text, it is rather difficult to assess why a particular answer has been given and to deduce the skills and strategies employed. In such cases questions focusing on specific enabling skills do seem to be called for in order to provide the basis for convincing extrapolation.

If this question of the relationship between performance and the way it is achieved, and the testing strategy which it is legitimate to adopt in order to

measure it seems to have been dealt with at inordinate length in the context of this paper, this reflects my feeling that here is the central distinction between what has gone before and what is now being proposed.

Admitting the necessity for tests of performance immediately raises the problem of assessment. How does one judge production in ways which are not hopelessly subjective, and how does one set receptive tasks appropriate for different levels of language proficiency?

The answer seems to lie in the concept of an operational scale of attainment, in which different levels of proficiency are defined in terms of a set of performance criteria. The most interesting work I know of in this area has been carried out by B J Carroll (Carroll, 1977). In this, Carroll distinguishes different levels of performance by matching the candidate's performance with operational specifications which take account of the following parameters:

Size Complexity	of text which can be handled
Range	of, eg enabling skills, structures, functions which can be handled
Speed	at which language can be processed
Flexibility	Shown in dealing with changes of, eg topic
Accuracy Appropriacy	with which, eg enabling skills, structures, functions, can be handled
Independence	from reference sources and interlocutor
Repetition Hesitation	in processing text

These specifications (despite the difficulties of phrasing them to take account of this in the summary given) are related to both receptive and productive performance.

It may well be that these specifications need to be refined in practice, but they seem to offer a way of assessing the quality of performance at different levels in a way which combines face validity with at least potential reliability. This question of reliability is of course central. As yet there are no published data on the degree of marker reliability which can be achieved using a scheme of this sort, but informal experience suggests that standardisation meetings should enable fairly consistent scorings to be achieved. One important factor is obviously the form which these scores should take and the precise basis on which they should be arrived at.

It would be possible to use an analytic system whereby candidates' performance was marked in terms of each of the criteria in turn and these were then totalled to give a score. More attractive (to me at least) is a scheme whereby an overall impression mark is given with the marker instructed simply to base his impression on the specified criteria. Which of these will work better in practice remains to be seen, but the general point may be made that the first belongs to a quantitative, analytic tradition, the second to a qualitative, synthetic approach.

Content

We have so far considered some of the implications of a performance-based approach to testing, but have avoided the central issue: what performance? The general point to make in this connection is perhaps that there is no general answer.

One of the characteristic features of the communicative approach to language teaching is that it obliges us (or enables us) to make assumptions about the types of communication we will equip learners to handle. This applies equally to communicative testing.

This means that there is unlikely to be, in communicative terms, a single overall test of language proficiency. What will be offered are tests of proficiency (at different levels) in terms of specified communicative criteria. There are three important implications in this. First, the concept of pass:fail loses much of its force; every candidate can be assessed in terms of what he can do. Of course some will be able to do more than others, and it may be decided for administrative reasons that a certain level of proficiency is necessary for the awarding of a particular certificate. But because of the operational nature of the test, even low scorers can be shown what they have achieved. Secondly, language performance can be differentially assessed in different communicative areas. The idea of 'profile reporting' whereby a candidate is given different scores on, eg speaking, reading, writing and listening tests is not new, but it is particularly attractive in an operational context where scores can be related to specific communicative objectives.

The third implication is perhaps the most far-reaching. The importance of specifying the communicative criteria in terms of which assessment is being offered means that examining bodies will have to draw up, and probably publish, specifications of the types of operation they intend to test, the content areas to which they will relate and the criteria which will be adopted in assessment. Only if this is done will the test be able to claim to know what it is measuring, and only in this way will the test be able to show meaningfully what a candidate can do.

The design of a communicative test can thus be seen as involving the answers to the following questions:

1 What are the performance operations we wish to test? These are arrived at by considering what sorts of things people actually use language for in the areas in which we are interested.

2 At what level of proficiency will we expect the candidate to perform these operations?

3 What are the enabling skills involved in performing these operations? Do we wish to test control of these separately?

4 What sort of content areas are we going to specify? This will affect both the types of operation and the types of 'text'[6] which are appropriate.

5 What sort of format will we adopt for the questions we set? It must be one which allows for both reliability and face validity as a test of language use.

Conclusion

The only conclusion which is necessary is to say that no conclusion is necessary. The rhetorical question posed by the title is merely rhetoric. After all it matters little if the developments I have tried to outline are actually evolutionary. But my own feeling is that those (eg Davies, 1978) who minimise the differences between different approaches to testing are adopting a viewpoint which is perhaps too comfortable; I think there is some blood to be spilt yet.

---

[6] Use of the term 'text' may mislead the casual reader into imagining that only the written language is under discussion. In fact the question of text type is relevant to both the written and the spoken language in both receptive and productive terms. In the written mode it is clear that types of text may be specified in terms such as 'genre' and 'topic' as belonging to a certain set in relation to which performance may be assessed; specifying spoken texts may be less easy, since the categories that should be applied in an analysis of types of talking are less well established. I am at present working in a framework which applies certain macro-functions (eg ideational, directive, interpersonal) to a model of interaction which differentiates between speaker-centred and listener-centred speech. It is hoped that this will allow us to specify clearly enough the different types of talking candidates will be expected to deal with. More problematical is the establishing of different role-relationships in an examination context and the possibility of testing the candidates' production of anything but rather formal stranger:stranger language. Simulation techniques, while widely used for pedagogic purposes, may offend against the authenticity of performance criterion we have established, though it is possible that those who are familiar with them may be able to compensate for this.

BIBLIOGRAPHY

CARROLL, J B
The psychology of language testing. In: DAVIES A, ed (1968), *qv.*

CHOMSKY, N
Aspects of the theory of syntax. MIT Press, 1965.

CORDER, S P
Error analysis, interlanguage and second language acquisition. Language teaching and linguistics abstracts, Vol. 8, no. 4, 1975.

DAVIES, A, ed
Language testing symposium. London: Oxford University Press, 1968.

DAVIES, A
Language testing. Language teaching and linguistics abstracts, Vol. 11, nos. 3/4, 1978.

MORROW, K
Techniques of evaluation for a notional syllabus. Royal Society of Arts (mimeo), 1977.

OLLER, J
Dictation as a device for testing foreign language proficiency. English language teaching journal, Vol. 25, no. 3, 1971.

OLLER, J
Cloze tests of second language proficiency and what they measure. Language learning, Vol. 23, no. 1, 1973.

PILLINER, A E G
Subjective and objective testing. In: DAVIES, A, ed (1968), *qv.*

ROBINSON, P
Oral expression tests. English language teaching, Vol. 25, nos. 2 - 3, 1973.

SELINKER, L
Interlanguage. International review of applied linguistics, Vol. 10, no. 3, 1972.

SPOLSKY, B
Language testing: art or science? Paper presented at the Fourth AILA International Congress, Stuttgart, 1975.

VALETTE, R M
Modern language testing: a handbook. Harcourt Brace & World, 1967.

WILKINS, D A
Notional syllabuses. Oxford University Press, 1976.

REACTION TO THE MORROW PAPER (1)

Cyril J Weir, Associated Examining Board

Three questions need to be answered by those professing adherence to this 'new wave' in language testing:

1 What is communicative testing?
2 Is it a job worth doing?
3 Is it feasible?

1 What is communicative testing?

There is a compelling need to achieve a wider consensus on the use of terminology in both the testing and teaching of language if epithets such as 'communicative' are to avoid becoming as debased as other terms such as 'structure' have in EFL metalanguage. Effort must be made to establish more explicitly what it is we are referring to, especially in our use of key terms such as 'competence' and 'performance', if we are to be more confident in the claims we make concerning what it is that we are testing.

Canale and Swain (1980) provide us with a useful starting point for a clarification of the terminology necessary for forming a more definite picture of the construct, communicative testing. They take communicative competence to include grammatical competence (knowledge of the rules of grammar), sociolinguistic competence (knowledge of the rules of use and rules of discourse) and strategic competence (knowledge of verbal and non-verbal communication strategies). In Morrow's paper a further distinction is stressed between communicative competence and communicative performance, the distinguishing feature of the latter being the fact that performance is the realisation of Canale and Swain's (1980) three competences and their interaction:

> '. . . in the actual production and comprehension of utterances under the general psychological constraints that are unique to performances.'

Morrow agrees with Canale and Swain (1980) that communicative language testing must be devoted not only to what the learner knows about the form of the language and about how to use it appropriately in contexts of use (**competence**), but must also consider the extent to which the learner is actually able to demonstrate this knowledge in a meaningful communicative

situation (**performance**) ie what he can do with the language, or as Rea (1978) puts it, his

> '. . . ability to communicate with ease and effect in specified sociolinguistic settings.'

It is held that the performance tasks candidates might be faced with in communicative tests should be representative of the type they might encounter in their own real world situation and would correspond to normal language use where an integration of communicative skills is required with little time to reflect on or monitor language input and output.

If we accept Morrow's distinction between tests of competence and performance and agree with him that the latter is now a legitimate area for concern in language testing, then this has quite far-reaching ramifications for future testing operations. For if we support the construct of performance based tests then in future far greater emphasis will be placed on the ability to communicate, and as Rea (1978) points out, language requirements will need to be expressed in functional terms and it will be necessary to provide operationally defined information on a candidate's test proficiency. Morrow raises the interesting possibility that in view of the importance of specifying the communicative criteria in terms of which assessment is being offered, public examining bodies would have to demonstrate that they know what it is that they are measuring by specifying the types of operation they intend to test and be able to show meaningfully in their assessment what a candidate could actually do with the language.

Morrow also points out that if the communicative point of view is adopted there would be no one overall test of language proficiency. Language would need to be taught and tested according to the specific needs of the learner; ie in terms of specified communicative criteria. Carroll (1980) makes reference to this:

> '. . . different patterns of communication will entail different configurations of language skill mastery and therefore a different course or test content.'

Through a system of profile reporting, a learner's performance could be differentially assessed in different communicative areas and the scores related to specific communicative objectives.

2 Is it a job worth doing?

Davies (1978) suggests that by the mid '70s, approaches to testing would seem to fall along a continuum which stretches from 'pure' discrete item tests at one end, to integrative tests such as cloze at the other. He takes the view

that in testing, as in teaching, there is a tension between the analytical on the one hand and the integrative on the other. For Davies:

> '. . . the most satisfactory view of language testing and the most useful kinds of language tests, are a combination of these two views, the analytical and the integrative.'

Morrow argues that this view pays insufficient regard to the importance of the productive and receptive processing of discourse arising out of the actual use of language in a social context with all the attendant performance constraints, eg processing in real time, unpredictability, the interaction-based nature of discourse, context, purpose and behavioural outcomes.

A similar view is taken by Kelly (1978) who puts forward a convincing argument that if the goal of applied linguistics is seen as the applied analysis of meaning, eg the recognition of the context-specific meaning of an utterance as distinct from its system-giving meaning, then we as applied linguists should be more interested in the development and measurement of ability to take part in specified communicative performance, the production of and comprehension of coherent discourse, rather than in linguistic competence. It is not, thus, a matter of whether candidates know, eg through summing the number of correct responses to a battery of discrete-point items in such restricted areas as morphology, syntax, lexis and phonology, but rather, to take the case of comprehension, whether they can use this knowledge in combination with other available evidence to recover the writer's or speaker's context-specific meaning. Morrow would seem justified in his view that if we are to assess proficiency, ie potential success in the use of the language in some general sense, it would be more valuable to test for a knowledge of and an ability to apply the rules and processes, by which these discrete elements are synthesized into an infinite number of grammatical sentences and then selected as being appropriate for a particular context, rather than simply test a knowledge of the elements alone.

In response to a feeling that discrete-point tests were in some ways inadequate indicators of language proficiency, the testing pendulum swung in favour of global tests in the 1970s, an approach to measurement that was in many ways contrary to the allegedly atomistic assumptions of the discrete-point testers. It is claimed by Oller (1979) that global integrative tests such as cloze and dictation go beyond the measurement of a limited part of language competence achieved by discrete-point tests and can measure the ability to integrate disparate language skills in ways which more closely approximate to the actual process of language use. He maintains that provided linguistic tests such as cloze require 'performance' under real life contraints, eg time, they are at least a guide to aptitude and potential for communication even if not tests of communication itself.

Kelly (1978) is not entirely satisfied by this argument and although he admits that to the extent that:

> '. . . they require testees to operate at many different levels simultaneously, as in authentic communication, global tests of the indirect kind have a greater initial plausibility than discrete items . . . and certainly more than those items which are both discrete and indirect, such as multiple-choice tests of syntax.'

he argues that:

> 'only a direct test which simulates as closely as possible authentic communication tasks of interest to the tester can have a first order validity ie one derived from some model of communicative interaction.'

Even if it were decided that indirect tests such as cloze were valid in some sort of derived fashion, it still remains that performing on a cloze test is not the same sort of activity as reading.

This is a point taken up by Morrow who argues that indirect integrative tests, though global in that they require candidates to exhibit simultaneous control over many different aspects of the language system and often of other aspects of verbal interaction as well, do not necessarily measure the ability to communicate in a foreign language. Morrow correctly emphasises that though indirect measures of language abilities claim extremely high standards of reliability and validity as established by statistical techniques, the claim to validity remains suspect.

Morrow's advocacy of more direct, performance-based tests of actual communication has not escaped criticism though. One argument voiced is that communication is not co-terminous with language and a lot of communication is non-linguistic. In any case, the conditions for actual real-life communication are not replicable in a test situation which appears to be by necessity artificial and idealised and, to use Davies's phrase (1978), Morrow is perhaps fruitlessly pursuing 'the chimera of authenticity'.

Morrow is also understandably less than explicit with regard to the nature and extent of the behavioural outcomes we might be interested in testing and the enabling skills which contribute to their realisation. Whereas we might come nearer to specifying the latter as our knowledge of the field grows, the possibility of ever specifying 'communicative performance', of developing a grammar of language in use, is surely beyond us given the unbounded nature of the surface realisations.

Reservations must also be expressed concerning Morrow's use of the phrase 'performance tests'. A test which seeks to establish how the learner performs in a single situation, because this is the only situation in which the learner will have to use the target language, (a very unlikely state of affairs) could be considered a performance test. A performance test is a test which samples behaviours in a single setting with no intention of generalising beyond that setting. Any other type of test is bound to concern itself with competence for the very act of generalising beyond the setting actually tested implies some statement about abilities to use and/or knowledge. In view of this it would perhaps be more accurate if instead of talking in terms of testing performance ability we merely claimed to be evaluating samples of performance, in certain specific contexts of use created under particular test constraints, for what they could tell us about a candidate's underlying competence.

Though a knowledge of the elements of a language might well be a necessary prerequisite to language use, it is difficult to see how any extension of a structuralist language framework could accommodate the testing of communicative skills in the sense Morrow is using the term. Further, a framework such as Lado's might allow us to infer a student's knowledge which might be adequate, perhaps, for diagnostic/ordering purposes, but is it adequate for predicting the ability of a student to use language in any communicative situation?

I do not feel we are yet in a position to give any definite answer to the question 'Is communicative testing a job worth doing?'. Though I would accept that linguistic competence must be an essential part of communicative competence, the way in which they relate to each other or either relates to communicative performance has in no sense been clearly established by empirical research. There is a good deal of work that needs to be done in comparing results obtained from linguistically based tests with those which sample communicative performance before one can make any positive statements about the former being a sufficient indication of likely ability in the latter or in real-life situations.

Before any realistic comparisons are possible, reliable, effective, as well as valid, methods for **establishing** and **testing** relevant communicative tasks and enabling skills need to be devised and investigated. This raises the last of the three questions posed at the start of this paper: 'How feasible is communicative testing?'. A satisfactory standard of test reliability is essential because communicative tests, to be considered valid, must first be proven reliable. Rea (1978) argues that simply because tests which assess language as communication cannot automatically claim high standards of reliability in the same way that discrete item tests are able to, this should not be accepted as a justification for continued reliance on measures with very suspect validity.

Rather, we should first be attempting to obtain more reliable measures of communicative abilities if we are to make sensible statements about their feasibility.

3 Is it feasible?

Corder (1973) noted:

> 'The more ambitious we are in testing the communicative competence of a learner, the more administratively costly, subjective and unreliable the results are.'

Because communicative tests will involve us to a far greater extent in the assessment of actual written and oral communication, doubts have been expressed concerning time, expenditure, ease of construction, scoring, requirements in terms of skilled manpower and equipment, in fact, about the practicability of a communicative test in all its manifestations. To add to these problems we still lack a systematic description of the language code in use in meaningful situations and a comprehensive account of language as a system of communication.

For Kelly (1978) the possibility of devising a construct-valid proficiency test, ie one that measures ability to communicate in the target language, is dependent on the prior existence of:

> '. . . appropriate objectives for the test to measure.'

Advocates of communicative tests seem to be arguing that it is only necessary to select certain representative communication tasks as we do not use the same language for all possible communication purposes. In the case of proficiency tests, these tasks are seen as inherent in the nature of the communication situation for which candidates are being assessed. Caution, however, would demand that we wait until empirical evidence is available before making such confident statements concerning the identification of these tasks as only by first examining the feasibility of establishing suitable objectives through research into real people coping with real situations, will we have any basis for investigating the claims that might be made for selecting a representative sample of operational tasks to assess performance ability. Even if it were possible to establish suitable objectives, ie successfully identify tasks and underlying constituent enabling skills, then we would still have to meet the further criticism that the more authentic the language task we test, the more difficult it is to measure reliably. If, as Morrow suggests, we seek to construct simulated communication tasks which closely resemble those a candidate would face in real life and which make realistic demands on him in

terms of language performance behaviours, then we will certainly encounter problems especially in the areas of extrapolation and assessment.

Kelly (1978) observed that any kind of test is an exercise in sampling and from this sample an attempt is made to infer students' capabilities in relation to their performance in general:

> 'That is, of all that a student is expected to know and/or do as a result of his course of study (in an achievement test) or that the position requires (in the case of a proficiency test), a test measures students only on a selected sample. The reliability of a test in this conception is the extent to which the score on the test is a stable indication of candidates' ability in relation to the wider universe of knowledge, performance, etc., that are of interest.'

He points out that even if there is available a clear set of communication tasks:

> '. . . the number of different communication problems a candidate will have to solve in the real world conditions is as great as the permutations and combinations produced by the values of the variables in the sorts of messages, contexts of situation and performance conditions that may be encountered.'

Thus on the basis of performance, on a particular item, one ought to be circumspect, to say the least, in drawing conclusions about a candidate's ability to handle similar communication tasks.

In order to make stable predictions of student performance in relation to the indefinitely large universe of tasks, it thus seems necessary to sample candidates' performances on as large a number of tasks as is possible, which conflicts immediately with the demands of test efficiency. The larger the sample, and the more realistic the test items, the longer the test will have to be.

In the case of conventional language tests aimed at measuring mastery of the language code, extrapolation would seem to pose few problems. The grammatical and phonological systems of a language are finite and manageable and the lexical resources can be delimited. The infinite number of sentences in a language are made up of a finite number of elements and thus tests of the mastery of these elements are extremely powerful from a predictive point of view. Thus, we might tend to agree with Davies (1978):

> '. . . what remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning . . . Grammar is far more powerful in terms of generalisability than any other language feature.'

However, Kelly (1978) puts forward an interesting argument against this viewpoint. It is not known, for example, how crucial a complete mastery of English verb morphology is to the overall objective of being able to communicate in English, or how serious a disability it is not to know the second conditional. We thus have:

> '. . . no reliable knowledge of the relative functional importance of the various structures in a language.'

Given this failing, it would seem impossible to make any claims about what students should be able to do in a language on the basis of scores on a discrete-point test of syntax. The construct, ability to communicate in the language, involves more than a mere manipulation of certain syntactic patterns with a certain lexical content. In consequence, it seems we still need to devise measuring instruments which can assess communicative ability in some more meaningful way.

As a way out of the extrapolation quandary, Kelly (1978) suggests a two-stage approach to the task of devising a test that represents a possible compromise between the conflicting demands of the criteria of validity, reliability and efficiency.

> 'The first stage involves the development of a direct test that is maximally valid and reliable, and hence inefficient. The second stage calls for the development of efficient, hence indirect, tests of high validity. The validity of the indirect tests is to be determined by reference to the first battery of direct tasks.'

As far as large-scale proficiency testing is concerned, another suggestion that has been made is that we should focus attention on language use in individual and specified situations while retaining, for purposes of extrapolation, tests of the candidate's ability to handle that aspect of language which obviously is generalisable to all language use situations, namely the grammatical and phonological systems. The hard line Morrow has adopted in the article under consideration makes it unlikely that he would contemplate either of these suggestions and would continue to argue for the use of pure direct performance-based tests.

Morrow's argument is that a model (as yet unrealised) for the performance of global communicative tasks may show, for any task, the enabling skills which have to be mobilised to complete it. He argues that assessment of ability in using these skills would yield data which are relevant across a broad spectrum of global tasks, and are not limited to a single instance of performance, though in practice these are by no means as easy to specify as precisely as he assumes nor are there any guidelines available for assessing their relative importance for the successful completion of a particular communicative operation, let alone their relative weighting across a spectrum of tasks. He is also aware that there exists in tests of enabling skills a fundamental weakness in the relationship between the whole and the parts, as a candidate may prove quite capable of handling individual enabling skills and be incapable of mobilising them in a use situation or developing appropriate strategies to communicate effectively.

In practice it is by no means easy even to identify those enabling skills which might be said together to contribute towards the successful completion of a communicative task. Morrow would appear to assume that we are not only able to establish these enabling skills, but also able to describe the relationship that exists between the part and the whole in a fairly accurate manner (in this case, how 'separate' enabling skills contribute to the communicative task). He would seem to assume that there is a prescribed formula:

possession and use of enabling skills X+Y+Z $=$ successful completion of communicative task

whereas it would seem likely that the added presence of a further skill or the absence of a named skill might still result in successful completion of the task in hand.

The second main problem area for Morrow is that of assessment. Given that performance is an integrated phenomenon, a quantitative assessment procedure would seem to be invalid so some form of qualitative assessment must be found. This has obvious implications for reliability. A criticism often made is that it is not possible to assess production qualitatively in ways which are not hopelessly subjective. For Morrow, the answer seems to lie in the concept of an operational scale of attainment, in which different levels of proficiency are defined in terms of a set of performance criteria. B J Carroll (op. cit. and 1978a and this volume) distinguishes different levels of performance by matching the candidate's performance with operational specifications which take account of parameters such as:

size, complexity, range, speed, flexibility, accuracy, appropriacy, independence, repetition and hesitation.

Morrow, as Carroll, advocates the use of a qualitative-synthetic approach, a form of banded mark scheme (see Caroll, this volume, for examples of this type of rating scheme) where an overall impression mark is awarded on the basis of specified criteria in preference to any analytic scheme. It is quite likely that the operational parameters of B J Carroll (op. cit.) eg size, complexity, range, accuracy, appropriacy, etc., will be subject to amendment in practice and in some cases even omission, but as Morrow argues in the article under review:

> '. . . they seem to offer a way of assessing the quality of performance at different levels in a way which combines face validity with at least potential reliability.'

There are no published data on the degree of marker reliability which can be achieved using a scheme of this sort, but Morrow's experience with the new R S A examination and the vast experience of G C E boards in the impression-based marking of essays suggests that standardisation meetings should enable **fairly** consistent scorings to be achieved, or at least as consistent as those achieved by analytical marking procedures.

Perhaps the point that should be made in answer to the question 'Is it feasible?' is that once again we do not yet know the answer. Until we have actually sought to confront the problems in practice, I feel it would be wrong to condemn communicative testing out of hand. What is needed is empirical research into the feasibility of establishing communicative tests, plus a comparison of the results that can be obtained through these procedures with those that are provided by discrete-point and indirect integrative measures.

BIBLIOGRAPHY

CANALE, M and SWAIN, M
Theoretical bases of communicative approaches to second language teaching and testing. In: Journal of applied linguistics, 1, 1, 1-47, 1980.

CARROLL, B J
An English language testing service: specifications. London: British Council, 1978 and this volume.

CARROLL, B J
Guidelines for the development of communicative tests. London: Royal Society of Arts, 1978a.

CARROLL, B J
Testing communicative performance: an interim study. Pergamon, 1980.

COOPER, R L
An elaborated testing model. In: Language learning (special issue) 3: Problems in foreign language testing, 57-72, 1968.

CORDER, S P
Introducing applied linguistics. London: Penguin, 1973.

DAVIES, A *ed.*
Language testing symposium: a psycholinguistic approach. London: Oxford University Press, 1968.

DAVIES, A
Language testing: survey article. In: Language teaching and linguistics abstracts, 2 3/4; part 1: 145-159; part 2: 215-231; 1978.

FITZPATRICK, R and MORRISON, E J
Performance and product evaluation. In: THORNDIKE, R L, ed. Educational measurement. 2nd ed. Washington, DG: American Council on Education, 1971.

HYMES, D H
On communicative competence. In: PRIDE AND HOLMES, eds, Sociolinguistics. Harmondsworth: Penguin, 1972, pp 269-293 (excerpts from the paper published 1971 by University of Pennsylvania Press, Philadelphia).

KELLY, R
On the construct validation of comprehension tests: an exercise in applied linguistics. PhD. University of Queensland, 1978.

MORROW, K
Techniques of evaluation for a notional syllabus. London: Royal Society of Arts, 1977.

MORROW, K
Testing: revolution or evolution. In: JOHNSON, K and BRUMFIT, C, eds. The communicative approach to language teaching. London: Oxford University Press, 1979 and this volume.

OLLER, J W
Language tests at school. London: Longman, 1979.

REA, P M
Assessing language as communication. In: MALS journal (new series: 3). University of Birmingham: Department of English, 1978.

ROYAL SOCIETY OF ARTS
Examinations in the communicative use of English as a foreign language: specifications and specimen papers. London: Royal Society of Arts, 1980.

SPOLSKY, B
Language testing: the problem of validation. In: TESOL quarterly, 2, 88-94, 1968.

WIDDOWSON, H G
Teaching language as communication. London: Oxford University Press, 1978.

REACTION TO THE MORROW PAPER (2)

Alan Moller, The British Council, London

Morrow's article is an important contribution to the discussion of communicative language testing. Some of the content, however, is marred by a somewhat emotional tone, although Morrow admits at the end that the title is rhetorical. The effect on the reader who is not informed about language testing could be misleading. The case for communicative language testing may well be stated forthrightly and with conviction, but talk of 'revolution' and 'spilling of blood' implies a crusading spirit which is not appropriate. The most traditional forms of language examining, and indeed of examining in most subjects, have been the viva and the dissertation or essay, both basic forms of communication. Reintroduction of these forms of examining, with some modifications, can hardly be termed revolutionary. What is new is the organisation of these traditional tasks. The nature of the task is more clearly specified, there is a more rigorous approach to the assessing of the language produced, and the label given to this process is new. More suitable titles for this discussion might be 'language testing: the communicative dimension', or 'communicative language testing: a re-awakening'.

Work in this area is recent and falls within the compass of what Spolsky (1975) termed the psycholinguistic-sociolinguistic phase of language testing. However, it is perhaps time to identify a fourth phase in language testing, closely linked to the third, the sociolinguistic-communicative phase.

As is often the case with discussion of communicative competence, communicative performance, and now communicative testing, no definition is given! But the characteristics identified by Morrow give some indication as to what might be included in definitions. It would seem that the general purpose of communicative tests is to establish first whether communication is taking place and secondly the degree of acceptability of the communication. This implies making judgements on the effectiveness and the quality of the communication observed.

The deficiencies of the structuralist method of language teaching and of that phase of language testing are well rehearsed, and Morrow need not have devoted so much space to it. He was right to point out J B Carroll's (1968) underlining of the integrated skills of listening, speaking, reading and writing.

But he has failed to point out that although integrated texts were presented to students, and although students were often asked to produce short

integrated texts, the items themselves were normally discrete, focusing on structural or lexical features. While agreeing that the primacy of contrastive analysis as a basis of language tests is no longer acceptable, we must beware of implying or insisting that the primacy of language as communication is the sole basis for language proficiency tests.

Discussions on language testing normally touch on two key questions. Morrow's concern with language as communication and his failure to define communicative language testing ensure that reaction to his article bring these questions to the fore:

1 What is language, and what is language performance?
2 What is to be tested?

In answer to these questions we might propose the following definition of communicative language testing:

> an assessment of the ability to use one or more of the phonological, syntactic and semantic systems of the language
>
> 1 so as to communicate ideas and information to another speaker/reader in such a way that the intended meaning of the message communicated is received and understood, and
>
> 2 so as to receive and understand the meaning of a message communicated by another speaker/writer that the speaker/writer intended to convey.

This assessment will involve judging the quality of the message, the quality of the expression and of its transmission, and the quality of its reception in its transmission.

Morrow has commented on discrete item (atomistic) tests and integrated (competence) tests and concluded that neither type 'gives any convincing proof of the candidate's ability to actually use the language'. Seven features of language use 'which do not seem to be measured in conventional tests' are then examined. If by conventional tests is meant discrete item and integrated tests, it is true that certain features may not be measured. It is equally questionable whether some of these features are even measured in so-called communicative tests. Does the measurement of a subject's performance include measuring the purpose of the text, its authenticity or its unpredictability, for example? It would seem to me that the claim is being

made that these features are not present in the test task in conventional tests. Even this claim is not entirely accurate.

It is helpful to examine the characteristics put forward by Morrow individually. **Purpose of text** The implication that every utterance produced in a communicative test is purposeful may not always be so. In many tests candidates may participate in communication and make statements which fulfil no other purpose than to follow the rules of what is likely to be an artificial situation. There is apparent purpose to the text being uttered, but the text may genuinely be no more purposeful than the texts presented in discrete and integrative test tasks. **Context** There are few items, even in discrete item tests, that are devoid of context. Communicative tests may attempt to make the context more plausible. **Performance** is not wholly absent from integrative tests, although it may be limited. Perhaps what is meant is production. **Interaction** Many conventional reading and listening tests are not based on interaction between the candidate and another speaker/hearer, but the candidate does **interact** with the text both in cloze and dictation. **Authenticity** This notion has been questioned elsewhere by Davies (1980) and seems to me to need careful definition. Language generated in a communicative test may be authentic only insofar as it is authentic to the context of a language test. It may be no more authentic – in the sense of resembling real life communication outside the test room – than many a reading comprehension passage. **Unpredictability** It is certain that unpredictability can occur naturally and can be built into tests of oral interaction. This feature would seem to be accounted for most satisfactorily in communicative language tests as would certain **behaviour** as the outcome of communicative test tasks.

Thus there are only two features of language use which are likely to occur only in communicative language tests. The absence or presence of seven characteristics in different types of test is shown more clearly in the table below. Column D refers to discrete item testing, column I to integrative tests and column C to communicative tests. Absence of a characteristic is indicated by X and presence by ✓.

There is, however, an important difference in the role of the candidate in the various kinds of tests. In the discrete and integrative tests the candidate is an **outsider.** The text of the test is imposed on him. He has to respond and interact in the ways set down. But in communicative performance tests the candidate is an **insider,** acting in and shaping the communication, producing the text together with the person with whom he is interacting.

Characteristics	D	I	C
Purpose of text	x	✓	✓
Context	(✓)	✓	✓
Performance	x	✓(limited)	✓
Interaction	x	✓	✓
Authenticity	?	?	?
Unpredictability	x	x	✓
Behaviour-based	x	x	✓

There may be little new in the subject's actual performance in communicative language tests. The main differences between traditional (pre-scientific) and communicative tests will lie more in the content of the tests and the way in which student performance is assessed. The content of the tests will be specified in terms of linguistic tasks and not in terms of linguistic items. Tests will be constructed in accordance with specifications and not simply to conform to formats of previous tests. Criteria for assessment will also be specified to replace simple numerical or grading scales which frequently do not make it clear what the points on the scale stand for. Certain criteria at different levels of performance will be worked out incorporating agreed parameters. These criteria may well take the form of a set of descriptions.

Another way of comparing communicative language testing with other types of tests is by considering the relative importance of the roles of the test constructor, the subject (or candidate) and the assessor in each of the phases of language testing identified by Spolsky – the pre-scientific, the psychometric-structuralist, and the psycholinguistic-sociolinguistic (competence) phases. The table below summarises these roles. The type of test is given on the left, column T refers to the role of the test constructor, column S to the role of the student, and column A to the role of the assessor. A ✓ indicates the importance of the role, (✓) indicates minor importance, and ( ) no importance.

Test type	T	S	A
Pre-scientific	(✓)	✓	✓
Psych/Struct	✓	(✓)	( )
Psych/Socio	(✓)	✓	(✓)
Communicative	✓	✓	✓

This table suggests that whereas in the pre-scientific and psycholinguistic/sociolinguistic (competence) tests the role of the test constructor (T) in setting questions and choosing texts is not important in the sense of being neither arduous, complex nor lengthy, his role is much more dominant in the psychometric/structuralist tests and communicative tests. In the psychometric/structuralist tests the work of the test constructor is all important, the task of the subject (S) is essentially to recognise or select, and in the majority of tests of this type marking is objective with therefore no role for the assessor (A). In the psycholinguistic/sociolinguistic tests, as defined, the main role is assumed by the subject who interacts with the text in his task of restoring it to its original or to an acceptable form. Communicative tests, however, are exacting at all stages, and the test constructor may well participate in the oral interaction with the subject and seek to introduce new tasks or different features of language use during the live interaction. His main preoccupations will be to set performance (global) tasks that will incorporate the language skills, microskills (enabling skills) and content that have been specified in order to provoke the subject to generate appropriate communication. The subject will seek to impress the assessor by carrying out the communication effectively and by responding to unpredictable shifts in the communication, and to new topics and new tasks. The assessor is confronted with communication that is unpredictable and of varying quality on which he must impose his pre-determined scale of criteria and reach a conclusion.

Morrow is right to point out that communicative language performance will be criterion-referenced as opposed to norm-referenced. The definition of these criteria is one of the major factors in the establishment of the validity of such tests. The relevance and consistency of these criteria are crucial and lead naturally to the question of the reliability of such tests.

It will be seen from the above table that communicative tests, in common with pre-scientific tests, put a lot of responsibility on the assessor in the

testing process. The subjectivity of the assessment gives rise to the problem of the reliability of such tests. Morrow touches on this problem, but it is not sufficient to say that it will simply be subordinate to face validity. Some further statement needs to be made. Careful specification of the tasks to be performed and careful specification of criteria for assessment are essential steps in the process of reducing the unreliability of this type of test. In the final analysis it may well be necessary to accept lower than normally accepted levels of reliability.

It has not been the intention of this reaction to Morrow's paper to consider in detail the points he has made but rather to use many of his observations as points of departure in an attempt to establish what communicative language performance might be, what it is that is being tested, and how valid assessments might be arrived at. It has been suggested that communicative language performance relates to the transmission and understanding of particular meanings in particular contexts and that what is being tested is the quality and effectiveness of the performance observed. Since this performance is highly subjective on the part of the subject and since the assessment must also be subjective, the reliability and validity of such tests will not be easy to establish. Careful specification of test tasks and assessment criteria would seem to be essential, but comparisons with other forms of language testing suggest that communicative testing places a heavier burden on test constructor, candidate and assessor. This does not mean that achievement of valid tests is impossible but implies more careful training of constructors and assessors and close monitoring of all phases of the testing process. Experience with ELTS[1] to date supports this contention.

There is a tendency when discussing new developments in language teaching and testing to throw out previous 'orthodoxies' and replace them with the latest one. Morrow's article has repeated the observation that good performance on a large number of discrete items in structuralist tests does not necessarily add up to ability to integrate them in effective language use. In discussing enabling skills the same problem of relating the parts to the whole has been observed. Communicative language testing seems to me to be primarily concerned with presenting subjects with integrated texts with which to interact, and with presenting them with sets of integrated tasks which will lead them to produce integrated spoken or written 'text'. As such the focus would seem to be predominantly on the whole rather than on the parts.

[1] English Language Testing Service administered jointly by the British Council and the University of Cambridge Local Examinations Syndicate.

Morrow suggests that the purpose of communicative testing may be proficiency testing. Later he suggests that proficiency tests will be specified in terms of communicative criteria. It is clear that communicative testing does test certain aspects of proficiency. But it is important to be aware that testing language proficiency does not amount just to communicative testing. Communicative language performance is clearly an element in, or a dimension of, language proficiency. But language competence is also an important dimension of language proficiency and cannot be ignored. It will also have to be tested in one or more of the many ways that have been researched during the past 30 years. Ignoring this dimension is as serious an omission as ignoring the re-awakening of traditional language testing in a communicative setting. Communicative language testing need not mean spilling the rather thin blood of present day language testing but could even enrich it!

BIBLIOGRAPHY

CARROLL, J B
The psychology of language testing. In: DAVIES, A, ed. Language testing symposium: a psycholinguistic approach. London: Oxford University Press, 1968.

DAVIES, A
John Oller and the restoration of the test. Paper presented at the Second International IUS Symposium, Darmstadt, May 1980.

SPOLSKY, B
Language testing: art or science? Paper presented at the Fourth AILA International Congress, Stuttgart, 1975.

REACTION TO THE MORROW PAPER (3)

J Charles Alderson, University of Lancaster

One of the main problems I seem to have with this paper is that I am not sure what it is about. The title implies a discussion of the issue of whether communicative language testing is fundamentally different from 'traditional' language testing, and the conclusion suggests the same when it says that the differences between the two approaches are really quite considerable. However, I agree with Morrow himself that this hardly matters: what would seem to be important is the precise nature of these differences and in particular the precise nature of communicative language tests. I am not sure that the paper does this, or even sets out to do so. The paper fails to identify traditional language tests despite frequent reference to them. Of course, an unknown or unidentified bogeyman is easy to attack, since the truth or accuracy of the attack cannot be ascertained. This is the not unfamiliar straw man syndrome. However, this opposition between traditional and communicative tests may not be the theme of the paper, since Morrow states 'this paper will be concerned with the implications for test design and construction of the desire to measure communicative proficiency' and later it is claimed that the paper has outlined 'some of the characteristics of language in use as communication which existing tests fail to measure or to take account of in a systematic way' and will examine 'some of the implications of building them into the design specification for language tests'. Note that 'existing tests' are not identified, so that it is difficult to evaluate the claim. The second footnote of the paper leads one to expect that criteria will be established for the design of communicative tests, by its criticism of the ARELS and JMB tests for not meeting 'in a rigorous way' such criteria. Unfortunately, this most interesting area remains undeveloped, since it is never clear what the criteria for the construction of communicative tests are, or how the JMB and ARELS tests fail to meet such criteria. Morrow goes on to say that working parties have been established to 'assess the feasibility of tests based on communicative criteria' but tantalisingly does not specify what these criteria are or might be. I wonder whether this is not the basic problem with the paper, namely that criteria are promised but not established. The importance of such criteria is that they would allow one not only to attempt to construct communicative language tests, but also to judge the feasibility or success of such attempts. Although the article goes on to talk about 'features of language use', 'characteristics of a test of communicative ability' and 'answers to questions', none of these amounts to an explicit statement of criteria, although, conceivably, such might be derived by implication from the criticisms of 'traditional' language tests. And indeed, later on we do appear to be back with the apparent topic of the paper, 'the central distinction between

what has gone before and what is now being proposed' and this is stated as being 'the relationship between performance and the way it is achieved and the testing strategy which it is legitimate to adopt in order to measure it'. My confusion may stem from two sources: the already mentioned failure of Morrow's clearly to identify exactly which tests are being attacked as 'traditional', allied with a failure to define terms like 'communicative proficiency', 'language competence', 'performance test', 'behavioural outcome', and so on; and on the other hand, my feeling that it is not necessary to draw unsubstantiated and inevitably over-simplified distinctions between past and present practice in language testing in order to explore the important issue of how to test communicative proficiency however that might be defined. It is, I think, important to bear in mind that Morrow is probably talking about proficiency testing – tests designed by examination bodies, or for organisations like the British Council – rather than about classroom tests. It is unlikely that the latter have been consistently guilty of placing too much importance on reliability, or accepting 'validity of a statistical rather than a practical nature', or of confining itself to 'the directly quantifiable modes of assessment', as he suggests. But even within the confines of proficiency testing, I fear Morrow overstates his case. He claims, for example, that the traditional 'measurement of language proficiency depends crucially on the assumption that (language) proficiency is neatly quantifiable in this way' (ie atomistically). I wonder whether traditional language testing 'crucially' depends on this assumption, in which case one might very well reject it, or whether the fact is not something more sensible, namely that such quantification is actually possible, unlike other, perhaps more direct and indeed desirable 'measurements' and that such quantitative measures at least give **some** indications, in an **indirect** manner, of some aspect of language proficiency. It seems that such an interpretation would not then rule out the value of **qualitative** measurement, even within traditional testing theory. The same point recurs when Morrow claims that an atomistic approach depends utterly on the assumption that knowledge of the parts equals knowledge of the whole. Do we know or believe that such **is** the assumption (in which case, Morrow is probably correct) or do we believe that the traditional testing position is one of assuming that we can **infer** the knowledge of the whole from the knowledge of the parts? Perhaps this is another example of the straw man syndrome. Similarly with the analogy with car driving which, although commonplace, is actually misleading. Nobody would wish to claim that a knowledge of the isolated elements of the integrated skill is sufficient for use, just as nobody would wish to claim that knowing how to manipulate the throttle, brake, clutch and so on of a car amounts to driving a car. The real issue is whether such knowledge, and in particular the knowledge of words, and of structure is **necessary**, and if necessary whether such knowledge is precisely specifiable and therefore testable. Even Carroll's 'clear statement of an "integrated" position'

recognises the need for both integration **and** atomism: one cannot interpret his (oft-quoted) remarks to mean that Carroll was against atomism merely because on its own he felt it to be insufficient. Morrow wishes to add the 'ability to synthesise'' to the ability to analyse language, but it seems important to examine in more detail precisely what such an ability is. Leaving aside conceivably equally important factors like the ability to operate under pressure of time, emotion, society and the like, the synthetic ability would seem worthy of much more treatment than it gets from Morrow in this paper. The nature or indeed existence of enabling skills, which we look at in more detail later, would perhaps qualify as part of such an examination.

Another charge levelled against (unidentified) traditional testing is that it views language learning as a 'process of accretion'. Now, if this were true, one would probably wish to condemn such an aberration, but is it? Does it follow from an atomistic approach to language that one views the **process** of learning as an accretion? This does not necessarily follow from the notion that the **product** of language learning is a series of items (among other things). Be that as it may, the alternative view of language learning that Morrow presents is not in fact an alternative, since by the same reasoning inter-languages can be acquired through accretion. No different view of the language learning process is necessarily implied, as far as I can see, by the notion of inter-language, which can be translated as one or more intermediate products on the road to proficiency.

Incidentally, contrary to what Morrow states, a 'structural/contrastive analysis' does not appear to follow necessarily from an atomistic approach although it is probably impossible without such an approach. It does not make sense to rule out contrastive analysis as the background for, or one of the inputs to, all test construction: presumably its usefulness depends on the test's purpose, and contrastive analysis may very well be useful for diagnostic tests.

Morrow's coyness when it comes to identifying actual examples of traditional testing, makes it extremely difficult to evaluate his claims, particularly for communicative language testing. In particular, he claims that there are seven features of language use that are not taken account of in 'conventional tests'. Now these features of language use are undeniable, and it is helpful to have them listed in this paper, but I doubt very much whether 'conventional tests' do not measure them. Of course, the question of how one knows or establishes whether they do or do not is of central importance, both for traditional tests and for communicative tests, since the issue is one of validation. If one uses the same technique that Morrow himself employs in the discussion of cloze and dictation, (that is, face validity) then it is almost certainly just not true that conventional tests took no account of

unpredictability, interaction, context, purpose, performance and so on. Of course, the crucial question, whatever the historical truth, is how will the 'new types of test' take account of these seven features 'systematically'? The question is evaded, as is the issue of the exhaustiveness of the list: ought we not perhaps consider an extension of the list of features to account more fully for the nature of language use, and include other features like deviance, and negotiated meaning, or the frequent existence of mutually conflicting interpretations of communicative interactions, and then examine the consequences in testing terms of such a list?

The assertion that conventional tests fail to account for the seven features of language use is not the only unsubstantiated claim that is made in the paper, and some of the claims seem central to the argument. 'The demand for context-free language production fails to measure the extent to which features of the candidate's performance may in fact hamper communication' – the fact is that we simply do not know whether this is true or not, or indeed, how to investigate it: what criteria shall we use to measure the hampering of communication? Traditional tests are criticised implicitly for using simplified texts rather than 'authentic' texts and tasks, yet the statement that 'the ability of the candidate to, eg read a simplified text tells nothing about his actual communicative ability', is merely an assertion, and will remain as such until we can measure 'actual communicative ability', by which time, of course, we would presumably not dream of asking someone to read a simplified text instead of being directly measured for his communicative ability. (A related point is whether simplification actually makes processing easier, which Morrow appears to think it does. The evidence is at best ambiguous).

The demand for 'authenticity' is itself not unproblematic. What are 'authentic language tasks ' in a language test? Does not the very fact that the setting is one of assessment disauthenticate most 'language tests'? Are there not some language tasks which are authentic in a language test, which would be inauthentic outside that domain? I find the authenticity argument somewhat sterile since it seems to assume that the domains of language teaching and language testing do not have their own set of specifications for authentic language use which are distinct from the specifications of other domains. Thus 'What is this? – It's a pencil' is authentic language teaching language, and so on. If one does not accept this, then authentic tasks are in principle impossible in a language testing situation, and communicative language testing is in principle impossible. A related problem, possibly caused by lack of definitions results from Morrow's statement that 'the success or failure of an interaction is judged by its participants on the basis of behavioural outcomes. Strictly speaking, no other criteria are relevant'. Without a definition of behavioural outcomes, this is hard to evaluate, but on the face of things, I can only assume that this refers to certain limited language functions like the

directive function. How can phatic or poetic uses of language be judged on behavioural outcomes? And why should behaviour on a language test be judged only in those terms? This presumably relates to the notion of performance test, but this term also remains undefined: what are the essential characteristics of a performance test? How is such a test to be validated? Against what? Behavioural outcomes? What would a performance test of listening look like that is different from the sorts of tests we already have? What, incidentally, would a nonintegrated test of listening be?

The question of test validation is central to any discussion of (proficiency) testing. In communicative tests, the main means of validation would appear to be content or construct validation, but without clear specification of the constructs, this is just not possible. A good example of the problems faced by the theory, and practice, is the issue of enabling skills. The paper implies that we already know the relation of such skills to performances ('An analysis of the global tasks in terms of which the candidate is to be assessed . . . will usually yield a fairly consistent set of enabling skills'), but in fact we know very little of the contribution made to any particular event by any one skill or even set of skills, and very little of the way in which such 'enabling skills' can be said to 'enable'. Even if we knew that such enabling skills existed, we would presumably need to know their relative importance overall, or even in one global task. And even if we knew this, we would still be faced with the likelihood that any one individual can plausibly do without (ie not call upon or not master) one, or a range, of the enabling skills, and still perform the task adequately: this supposition is at least as reasonable as the one that Morrow makes, and subject to the same requirement of verification. How either assertion might be verified is central to the problem of validation, and no solution appears obvious. The same point would appear to apply to the parameters of B J Carroll: to what extent, if at all, are the actual values of these parameters of size, range, accuracy, appropriacy and the like, actually specifiable for any one communicative event? If the values are not specifiable in terms of some notion of the ideal performance (a requirement of criterion-reference testing, which is what Morrow claims– and it remains a claim – communicative testing to be) then what is the use of such parameters? The question is complicated by this notion of the ideal (or optimal) performance: whose performance, which performance is criterial? Morrow implies in the paper that we are to compare non-native speakers' performance with those of native speakers ('Tests of receptive skills will similarly be concerned with revealing the extent to which the candidate's processing abilities match those of a native speaker'). How are we to compare the performance of the two groups (natives and non-natives)? Which native speakers are we to take? Are all native speakers to be assumed to be able to perform ideally on communicative tests? We know native speakers differ in at least some communicative abilities (reading, oracy, fluency) – how can they be

compared with non-natives? This aspect of the criteria question is simply ignored: how are we to judge performances on our tests? Tests, after all, are not merely elicitation devices for getting at samples of language behaviour, but assessment procedures: 'Tests will, thus, be concerned with making the learner produce samples of his own interlanguage based on his own norms of language production so that conclusions can be drawn from it' (Morrow, this volume p. 12). What sort of conclusions will be drawn and why? The questions are not asked.

How are we to evaluate communicative language tests? What criteria are we to use to help us construct them, or to help us determine their validity? It has been suggested that Morrow does not provide us with any explicit statements on this. However, some criteria are surely possible, unrelated to any particular view of language or language use in the sense of being determined by such a view; the criteria are statable in the form of questions one might pose of a test: in a sense they are meta-criteria, and the validity of the answers depends on the validity of the related theories. The questions one should ask of language tests (of any sort, not only proficiency tests), when judging them, when discussing the issue of test validity – does the test measure what it claims to measure? – can be divided into four areas: the test's view of language, the test's view of the learner, the test's view of learning and background knowledge:

What is the test's view of language?

What is 'knowing a language' in the test's terms?
Does the test view language as a set of isolated, separable items?
Does performance on the test reflect performance in the real world?
Do the testees have to **do** things with language?
Does the test measure the ability to function within a specified set of sociolinguistic domains?
Is the test based on a model of communication?
Does the test relate to the sociolinguistic variables that affect the use of language in communication?
(eg Does the test measure the learner's ability to recognise the effect of, and produce appropriate language for:
the setting of a communication?
the topic of a communication?
the function of a communication?
the modality of a communication?
the presuppositions in a communication?
the degree of formality of a communication?
the roles of participants in a communication?

the status of participants in a communication?
the attitudes of participants in a communication?)

Does the test take account of the fact that communication:
is interaction-based?
is unpredictable?
takes place under pressure of time?
takes place in a context?
takes place for a purpose?
is behaviour-based?
is not necessarily totally dependent on language?
that is,
are student reactions predictable?
are complex language skills measured?
is the situation real?
is the ability to **interpret** original messages measured?
is the ability to **produce** original messages measured?
is the **creative** element of language use tapped?
is the testee's **participation** required?

What is 'meaning' according to the test?
static, residing in words?
variable, according to context?
negotiable, depending on all the factors in the interaction?
Does the test recognise that language is redundant?
Is the language sample of the test biassed?, ie inauthentic, unusual.
Does the test cover **relevant** aspects of language skills?

What is the test's view of the learner?

Does the test confine itself to the lower part of a hierarchy of skills?
Does the test make demands on the cognitive skills (knowledge of the world, understanding, reasoning)?
Does the test involve the affects of the learner especially as in interpersonal behaviour?
Is the test appropriate for the proposed testees in terms of their knowledge, affects, skills?
Does the test take account of the learner's expectations?
ie his definition of his needs?
his notion of what it is to know a language?
Does the test allow different types of learners to show their abilities equally, or is it biassed in favour of one type of learner?
How would native speakers perform on the test?

What is the test's view of language learning?

Does the test assume that language learning is equivalent to gaining control over linguistic problems?
Is the test congruent with the aims and practices of the language teaching?
ie is the language being tested in the way it is taught?
are the tests appropriate both to the target performance of the course **and** to the competence which is assumed/known to underlie or enable that performance?
is the weighting (balance) of subtests appropriate to the language teaching?

Background knowledge?

Are extraneous variables – culture, subject-specific knowledge – involved in the test? Can they be excluded?
Does the test favour one type of knowledge?
Should the test have 'neutral' content? Is this possible?
Can content be separated from language?
What if the learner knows what to say, but does not know how to say it?
If we are to measure communication, which includes **ideational** knowledge, then should not the subject specialist also be involved in a 'language' test?

Many of these questions derive from Morrow himself although they are not confined to this source. In a sense, they form the unspoken criteria promised but not given in this paper. The paper is really about the relationship between theories of language, language use and language learning, and tests of language knowledge, language proficiency and language use. Morrow's final set of five questions can be seen as pointing the way to such detailed questions as above. The paper and in particular this final set of five questions, is very useful for the way in which directions are suggested for future research. Indeed, the only way in which we will ever get answers to the questions posed by Morrow is by carrying out research, and for a considerable period.

Summary

It seems to me that the Morrow article contains many important points.

1 It correctly emphasises the need for testing to catch up with language teaching.

2 It implicitly suggests ways in which testing might help teaching, through the specification of language use, for example. One of the advantages of a 'testing approach' is that it forces explicitness.

3 Morrow is right to avoid continua and clines, and to argue polemically. To say that everything is really part of the same thing appears to me to be unhelpful: what is interesting is where the differences lie. Thus it is helpful to set up dichotomies, provided, naturally, that the part of the dichotomy one is putting forward is not merely a negative attack on straw men.

4 The view of language use that Morrow puts forward seems to be essentially correct, and fruitful of further hypotheses and research. He may, however, rather underestimate the **dynamic** and negotiated nature of communication.

5 He is correct to see tests as embodiments of theories, or views, of the nature of language and of language learning. This aspect of test design seems to be neglected elsewhere. As he points out, if the theory is wrong, then the validity of the test is zero.

6 The problem and importance of extrapolation and assessment are rightly stressed.

7 On the whole, he is right to criticise the past's search for maximum reliability, and to point out the circularity of most validities.

However, I feel that the paper deals rather inadequately or not at all with a number of important issues.

1 How are the seven (or more) features of language use to be taken account of in communicative language tests?

2 It is important to distinguish between the problem of what language is to be sampled, and how that sample is to be judged.

3 What is the status of the enabling skills? How are they to be adequately measured?

4 The nature of language proficiency is left vague. Is proficiency something a native speaker has and a non-native has to acquire? Does the non-native already possess such proficiency which is merely realised in another language, but which is readily transferable, once one has 'cracked the code'? What is **successful** communication? On what basis are judgements to be made? Who judges, and why? What about the effect of non-linguistic elements like personality, motivation, awareness, and the like on successful outcomes? To what extent is this a purely language problem? To what extent should tests of 'communicative proficiency' be language tests?

5 What is the purpose of the test? Is there not an intimate relation between test purpose, test content and test format which is barely touched upon here? How, precisely, would test content and format be affected by test purpose?

The advantage of testing is that it forces explicitness: the test is an operationalisation of one's theory of language, language use and language learning. Testing is the testing ground for any approach to teaching. If we cannot get the tests our theories seem to require, then we have probably not got our theories right (unless, of course, the theory implies the impossibility of testing). Why has there apparently been such a failure to develop tests consistent with theories of communicative language use?

REPORT OF THE DISCUSSION ON COMMUNICATIVE LANGUAGE TESTING

J Charles Alderson, University of Lancaster

The most important question to be asked of any test, and communicative language tests are no exception, is what is it measuring? The question that arose in the discussions as to whether what communicative language tests are testing is actually anything different from what has been tested before is a subsidiary and less important issue: although there was a general suspicion that nothing new was being tested in communicative language tests, less agreement was reached on what such tests actually measure.

It is not important that communicative language tests look different from other types of test: what is important is that they measure what one wishes to measure. (There may, however, be good political or administrative reasons why 'communicative language tests' should look different: if they relate to an innovative curriculum which itself appears to be different, a measure of achievement on that curriculum which looked like traditional measures might engender disbelief in either the validity of the measure or the virtues of the new curriculum). However, even though the difference between communicative language tests and other tests may be relatively less important, one reason for comparing the different types of tests is to understand why communicative language testing has developed, and what it is that such tests appear to be measuring.

There would appear to be a variety of dimensions of language in use that existing language tests do not tap. It was generally agreed that existing tests may be unsatisfactory to the extent that they do not cover psycholinguistic abilities, (like enabling skills), or features of language (like unpredictability) which it may be important for students to be exposed to or tested upon. Such features or dimensions derive from two possible sources: either from our theories of language use – that is, our developing theories of the use of language for and in communication generate the dimensions which are to be operationalised in language tests; or they derive from 'real-life': from observations of the world around us at a pre-theoretical, ie descriptive stage.

Attempts to improve existing language tests from the first perspective – that of theory – are attempts to improve the construct validity of the tests; attempts to improve tests from the second perspective, that of mirroring reality in a more adequate fashion, are attempts to improve content validity.

There is a potential conflict between these two validities, in that a theory-derived test may look very different from a real-life derived test. For example, one's theory may include the notion of the transmission of information as being an important component of communication, of language in use. One might then construct a test to measure the quality of such transmission. Upshur's (1971) oral test, for example, is an attempt to do just this, and strives for construct validity. However, it may not look like a real-life situation. When do real people look at a set of four pictures and try to guess which one another person is describing? Tests striving for content validity could constitute job samples, that is, replications of reality, and would therefore inevitably be performance-based. The question is whether tests are mirrors of reality, or **constructed instruments** from a theory of what language is, what language processing and producing are, what language learning is.

In our discussion we were in no doubt that an awareness of the existence of other dimensions has increased in recent years, partly from work in psycholinguistics and sociolinguistics, partly from dissatisfaction with existing tests (either because they do not look right, or because they are thought not to give the results that are required).

However, one evaluates any theory, presumably, by its operationalisation. If operational definitions are not possible, then the theory is poorly stated or inadequate. It is not clear to what extent such operationalisations have been achieved in the construction of communicative language tests, and the view was expressed that possibly the fault lies, not with testers, but with the theories: if they do not permit adequate definitions in test terms, they are not adequate theories. Should one, however, wait for the development of adequate theories of language in use before proceeding with the development of communicative language tests? It was generally felt that this would be inappropriate, especially if it is the case, as seems likely, that a complete theory of communication will not be developed for a very, very long time.

One claimed advantage of communicative tests, or perhaps more accurately performance tests, is that they do not rely on adequate theory for their validity. They do not, for example, make assumptions about the status of competence in a Chomskyan sense, and its relation to performance — its predictive relationship to what people can actually do — because such tests aim to **measure** what people can do. If one is interested in whether students can perform adequately (adequacy being undefined for the moment) at a cocktail party, 'all' one has to do is to put that student into a cocktail party and see how he fares. The obvious problems with this are that it may not always be possible to put the student into a cocktail party (especially if there are several thousand students involved), and the fact that the performance is

being assessed may actually change the nature of the performance. One solution to the first problem is to simulate the cocktail party in some way, but that raises problems of authenticity, which relate to the second problem, that of the relationship between the performance and its assessment. Inevitably, any test is in danger of affecting performance if the testee is aware that he is being tested. To that extent, it is impossible for a test to be 'authentic' in the sense of mirroring reality. Of course, tests are themselves authentic situations, and anything that happens in a testing situation, must be authentic in its own terms: the problem comes when one tries to relate that testing situation to some other communicative situation. In a sense, the argument about authenticity is trivial in that it merely states that language use varies from situation to situation. The feeling was expressed that the pursuit of authenticity in our language tests is the pursuit of a chimera: it is simply unobtainable because they are language **tests.**

It was argued that the only interest in authenticity in tests is in the gathering of genuine data (ie data that has occurred) as part of test input. Tests have been developed based upon genuine data, where a real conversation has been recorded, transcribed, and re-recorded using actors reading from the transcription, at least partly in order to ensure good sound quality of the final test. Such practice may be authentic and justified within a testing context, although it probably runs counter to the original reason for gathering data.

Since one cannot, *a priori,* replicate in a test situation what the students will have to face in 'real-life', it was argued that what we should be doing is looking at students' performances on tasks defined according to criterial features, (for example the dimensions mentioned by Morrow like 'unpredictability') and then extrapolate to the outside world. Thus our tasks may not be authentic in the other-world sense, but they have value and validity because we are tapping dimensions, or abilities, which other tests do not tap.

Another, weightier problem than 'authenticity' that was discussed, is that of sampling. If one is interested in students' abilities to perform in cocktail parties, and one somehow measures that ability in one cocktail party, how does one know that in another cocktail party the student will perform similarly? The cocktail party chosen may not have been an adequate sample. This is particularly a problem when we are unable to be as specific about what we want students to be able to do as in this example. If our goals are to measure students' abilities to use language communicatively or to use English in a variety of situations, how are we to decide which tasks to give students in our tests which will adequately represent those goals?

If we are not interested in the students' ability to perform in a situation, but in situations A to Z, then how can we be sure that X is an adequate sample of A – Z. Might not situation B or M be more adequate?

This problem assumes that we are interested in prediction. The question being asked in the debate about sampling is – can we predict from performance on one task to performance on another task or series of tasks? Testing, in other words, is about predicting some criterion behaviour. The assumption of communicative testing, which is an assumption until evidence is produced to justify the notion, is that the only way to predict criterion behaviour is to set up (real) performance tasks. The question is whether one has to put people in to a particular situation in order to find out how they would perform in that situation. The view was expressed that there may be in communicative testing a danger of confusing the 'how' of predicting something, with the 'what' of the prediction. Communicative testing appears to try to bring together the manner and the content (or the test and the criterion) in an arguably unnecessary or indeed impossible manner: the communicative testing argument seems to be that instead of giving somebody a driving test, you put him into a car, and see if he hits the wall. Such assumptions about the need for performance tests need considerable research activity to support them: the discovery of the best predictor (driving test or performance) of the criterion (hitting the wall or not) is an empirical issue.

It may be that the sampling problem is also an empirical issue: in order to find out whether performance on task X is the best predictor of performance on tasks A to Z, one might give subjects a vast array of tasks to perform, and see which is the best predictor. However, predictive validity is not the only type of validity in which we are interested, as we have already seen.

In particular, the traditional proficiency test argument ignores the dimensions of face or content validity. One might argue, from the perspective of predictive validity, that what one is testing does not matter, provided that it predicts the criterion behaviour (performance in a cocktail party). If the best predictor of such behaviour is the size of one's boots, then what one must do is measure students' boots. This argument confuses causality with concomitant variation (students might change the size of boots they are wearing in order to pass the test, but still be unable to perform well in cocktail parties), and generally takes no account of issues like face or content validity.

It was generally agreed that the prior problem in both the sampling debate and the prediction debate, would seem to be that of defining what one wishes to assess, what performance one wishes to sample or predict. First one needs to define what it is that students have to do with language in a specific situation, or series of situations. The danger is that in specifying

communicative performance, one might end up describing an impossible variety of situations, which one cannot encompass for testing purposes.

The value of communicative language testing, and the difficulty, is that it represents an attempt to do precisely that: to define the criterion one is trying to sample or predict. Traditionally, proficiency testing at least has been concerned to find the best predictor of a criterion: the argument has run that the best proficiency test is the one which best predicts future behaviour. Thus one might claim that test X is valid because it predicts performance in a cocktail party. The crucial question surely is: what does one know about behaviour in a cocktail party? Gaining that knowledge was felt to be of paramount importance, since it represents the ultimate test. Thus one has to define what it means to perform well in a cocktail party. Once one has described this, one has produced a specification, a set of guidelines, for the construction of the test. Discovering the best predictor of this, or the most adequate sample, is of secondary importance. Thus it may be that the issue of extrapolation is not (yet) of crucial importance: even if we cannot generalise from performance in one situation to performance in a variety of situations, if we can say something about performance in **one** situation, then we have made progress, and if we can say something important about performance in the target situation, so much the better. Ultimately, after all, the student will have to perform, despite the statistical evidence of the relationship between predictor and predicted, or the theorised relationship between competence and performance.

The discussion focussed on what communicative language tests should do or should look like. What is the nature of the tasks which students are given? What makes them different from existing tests, and which features of language use do they take account of? What, for instance, does a communicative test of reading or listening look like? Presumably, a communicative test of reading would be, for example, a set of instructions leading to a behavioural outcome, linguistic or otherwise. The problem with this is that a satisfactory outcome may be reached without 'adequate' linguistic performance. It is possible to devise a vast variety of different tasks: what are the dimensions that must be included to qualify as 'communicative'? A claimed virtue of communicative testing is that it is more explicit about what it is trying to measure than existing tests are: in reading it may result in increased specificity of text type, or type of reading required, although this is not exclusive to communicative testing. This specification may result in an atomistic analysis of behaviours, which, paradoxically, may not be desirable in communicative tests. An interesting result of this consideration is the idea that the so-called dichotomy of communicative testing versus existing tests may be separate from, and unrelated to the (equally arguable) posited dichotomy between discrete-point and integrative tests. In this case, discrete-point communicative tests of reading would be perfectly feasible and justifiable.

The requirement that one analyse situations or target performance in order to establish criteria would appear also to demand an atomistic approach. In order to produce a communicative test, one must, presumably, either sample, or analyse and test. As has been seen, the problem with sampling is that it is difficult to do. However, it would appear that without a prior analysis of performance or tasks, one would have no basis for sampling. Thus, at some level, analysis is essential for communicative testing.

Most communicative testing has been concerned not with reading and listening, but with tests of oral and written production, which have been largely neglected in recent years because of the inherent problem of their reliability. The communicative test of oral production *par excellence* is often said to be the interview (a traditional form of test!). In an interview, the tester can probe and force the students to produce language, based on an inventory of questions and prompts. Typically, he does not work from a list of structures, since, in a communicative test situation, there is no need to think in terms of structural complexity. Interviewers do not deliberately manipulate structures to see if candidates can comprehend or produce them.

One of the dimensions of language in use that was discussed in more detail was that of unpredictability. The argument is that language use is unpredictable, and therefore so should our tests be. To what extent are interviews unpredictable? The interviewer has a set of possible prompts and questions and it would appear that the interview must be highly predictable. However, from the testee's point of view it is considerably less so (he presumably does not know what questions will be asked). What would a test that incorporated the dimensions of unpredictability look like? It would presumably not be a set of question-answer routines (although as was suggested this is less predictable for student than examiner): to what extent are 'unpredictable' tests possible for writing rather than speaking? If, in speaking tests, one requirement is that the responses, and indeed the initiations, should be unpredictable for the examiner, as participant in the interaction, then the question arises of the difficulty of participating in as well as evaluating an interaction that is 'unpredictable'. A common solution to this not unfamiliar problem is to have an interviewer and an observer in the same interview, where the observer is the examiner. This, however, raises the issue of outsider views: is it possible for an outsider to interpret interactions, especially ones which are supposed to be unpredictable? If they are unpredictable what does/can the observer look for? Can criteria be established to allow the assessment of something about whose nature we know little in advance? In any case, different observers will inevitably have different interpretations of events and their quality. This raised the familiar problem in testing: the issue of subjectivity. To what extent in

communicative testing is 'objectivity' of assessment attainable, if desirable? It was argued that objectivity is never possible in judgements about language related performance, and that one should simply aim to pool subjective judgements. This does not mean that everybody should agree on one judgement (score), but that judgements are averaged. There is considerable evidence to show that any four judges, who may disagree with each other, will agree **as a group** with any other four judges of a performance. (It was pointed out that it is, however, necessary for markers to agree on their terms of reference, on what their bands, or ranges of scores, are meant to signify: this can be achieved by means of a script or tape library).

Communicative testing has resulted in a focus, not only on the tasks of a test, but also upon the criteria used for assessing performance on those tasks. In particular the British Council has been involved in developing scales and criteria for assessment, which cover areas like appropriacy, amount of communication, content, establishment of communication, and so on. Judges are typically asked, in a non-impression scheme, to rate performances on several dimensions (thought to be relevant to the quality of language in use). One would expect, and indeed one gets, differential performance on different dimensions (such that it is possible to get, say, a three for appropriacy and a five for content), and it is undesirable to add scores on the separate dimensions together in order to arrive at some global assessment, because individual differences will be hidden in such a procedure: what is required is the reporting of some sort of profile. However, the question was raised of the independence of such dimensions, if not in reality, then at least in the ability of judges to rate independently. Cross contamination is quite likely, and only avoidable, if at all, by having different judges rate performances on different dimensions (such that one judge, for example, might rate on appropriacy, whilst another rates on amount of communication). The value of such a procedure would need to be established by empirical research. A problem related to the question of whether the grades given on particular scales actually represent performance on the stated dimension rather than some other dimension, is the question of whether communicative language tests are actually measuring language performance as subsumable under language in use, or whether they are measuring variables that might be said to be extraneous, non-language related. What, for example, is one to conclude about the performance of somebody who, when asked his opinion on a particular topic, does not volunteer anything because he does not have an opinion? Or what is one to make of the shy or introverted student on, say, a discussion test? Particularly in the area of EFL, it is quite likely that there will be cultural differences among testees as to what is acceptable behaviour on performance tasks, which might influence the amount and the quality of the 'behavioural outcome'? What is one to make of that? Must one accept the fact that the measures are not pure measures, on the grounds that 'life is like

that', ie people with different cultural backgrounds or personality or cognitive styles will suffer in the real world as well as on our tests?

The point was made that laymen have for a long time expected of language tests that they test language: indeed, such has been the requirement by sponsors of language tests, like the British Council, or the General Medical Council, namely that **only** language should be tested, and 'irrelevant' variables like personality, knowledge of subject matter, opinions and the like, be left out of language tests. To the present-day applied linguist, this looks like a naive oversimplification of the relationship between language and personality, language and thought, language and culture and one might well claim that it is in practice impossible to separate language from these other areas. Yet, since lay people hold such (strong) views on the matter, testers ignore them at their peril.

A further expectation, particularly of sponsors, is that native speakers should do well, even (within the bounds of reliability) perfectly on a language test. Traditionally, proficiency tests were partially validated by reference to native-speaker (perfect) performance. Communicative language tests in particular, though not exclusively, raise the issue of whether native speakers **can** do the task satisfactorily. Which native speakers is one talking about – educated? uneducated? certain professional groups rather than others? Which language is one a native speaker of – English? Medical English? The English used to write inflammatory articles on medical topics in the popular science press in a British context? Are we talking about native speakers who are (the equivalent of) first year under-graduate science students, or eminent and experienced neuro-surgeons? If a native speaker performs poorly on a task, is that because he is the wrong native speaker? Because he lacks the skill or the language? Because he is too clever? One problem that was mentioned with some native speakers on language tests is simply that they are too good: they see ambiguities and difficulties on certain test items that non-native speakers do not see: native speakers can often create plausible contexts for apparently incorrect responses.

Talk, within the field of communicative language testing, of behavioural outcomes, suggests that greatest importance is attached to the product of a communicative interaction. Considerable discussion took place, however, on the question as to whether in communicative language testing, or language testing in general, we need to know how individuals reach their result. Presumably for diagnostic purposes, information on the process is essential, in order to plan some sort of pedagogic treatment or intervention, but is it important to know how results were achieved, for other purposes? Proficiency testing might only be interested in the product, not the process, in which case one might argue that testing enabling skills is inappropriate,

because they belong to process. Indeed it was argued that enabling skills may vary from individual to individual, and certain of them may not be used by one person on one occasion to reach a given product, in the performing of a particular task. If one is only interested in the product, then behavioural outcomes are sufficient. If one is interested in knowing whether somebody can cross London, one simply measures whether they get across London, and does not worry about whether they used a map, used Arabic to consult more knowledgeable informants, or followed the written instructions in English that we as test designers had expected them to follow. What is important in this view is whether testees cross London, rather than whether they crossed in some prescribed manner (since in any event in 'real life' it is unlikely that they would follow such prescriptions). It was felt in any case, salutary to make the point that we are ignorant of how people achieve their ends, and that this is impossible to predict, on present knowledge at least, since different individuals will do it in different ways, or even the same individuals will do it differently on different occasions.

Does one need a breakdown of Process in order to construct a valid test task? To validate a test *vis-a-vis* its theory, one would appear to need a breakdown of possible performances on that task. Otherwise, one only has the final outcome for validation purposes. And one does not normally know whether a test is valid simply because people have 'passed' it. However, if one wishes to extrapolate, then one has presumably to talk about underlying skills (ie Process – how people go about doing the task) unless the sampling solution is accepted: 'If you can understand that lecture, then you will be able to understand lectures'. How one understood the lecture, or rather how one arrived at one's understanding of the lecture, is unimportant in this view. Traditional proficiency tests, it was pointed out in the discussion, are not intended to tell one anything at all about students' processes and problems: they 'simply' seek to answer the layman's question: 'Does this man speak English?'

Although the debate about communicative language tests focussed upon the question of what is being measured, it was felt to be impossible to determine what is being measured independently of considerations of how a measure will be validated. In other words, one anticipates the question – 'how do you know?' – as a response to an assertion that a test is a measure of X. How, with communicative language tests, do we know if we have measured what we claim to measure? How can we improve our communicative tests? When designing a new test one must know what one thinks represents an advance and an improvement over existing tests, and there must be some notion of how one can evaluate that, how one can confirm one's suspicion. It was generally agreed as unfortunate that in the world of communicative language testing, there is rather little discussion of how to validate and evaluate such tests, or how they might have been evaluated in the past. One is certainly not

absolved from the responsibility of stating one's criteria for validation (not just validity) by the (apparent) absence of other valid tests with which to compare one's own. The argument that one cannot validate a test because there are no other valid tests in existence does not stand up since it appeals only to concurrent validity. One problem with concurrent validation that was touched upon is the problem of interpretation of correlations. If the 'communicative' language test correlates highly with (invalid) discrete point tests, then is this evidence for the invalidity of the test, or for the existence of one general language proficiency. If one observes the (desired) low correlation, does this mean that the test is valid or that it is simply measuring something different, or measuring the same thing rather badly, because of unreliability?

Of course, one way of improving a test is to see what people think is wrong with the existing instrument, for particular purposes, and then see if the new test does the job better. A frequent complaint about proficiency tests is that they fail to identify students who subsequently have problems in their fields of study: they let into institutions students who should have been kept out. Ignoring the fact that test use and test construction are partly separate matters, one might say that such a proficiency test is failing to do its job because it fails to tap relevant skills. The problem is defining those relevant skills. To find out if one's new test is better, one might see how many students passing it actually had problems, (ignoring the difficulties caused by the fact that students who fail are not normally admitted). The problem with this sort of predictive validity is the time factor: one would expect and hope that the correlation between test performance and subsequent problems would decrease as other factors intervene over time, until in the end there would be no correlation. One can see that the extrapolation problem is in fact a validation problem, which relates to the problems of prediction (including the relationship with time factors) common to all language tests, communicative or otherwise. The point about communicative tests is that they make clearer the need to break the circularity of most validation procedures (the circularity consists of correlating with another test or measure) by appealing to outside criteria, because, precisely, of the claim that communicative tests are measures of language in use, 'real' language tests. However, appeal to ideology is not sufficient evidence for accepting the validity of a test. One needs empirical evidence to back up assertions of validity and claims that performance on one task relates to performance on other tasks.

One way of validating tests is to relate them closely to the language teaching that has preceded them. It is at times claimed that communicative language tests are more valid because they relate better to current trends in teaching than do other types of test. There may, however, be good arguments for tests

not being in line with teaching (despite the washback effect) because tests can be used as a means of evaluating the teaching; of validating the teaching. If one wishes to know not whether what has been taught has been learnt, but rather whether the right things have been taught, then one needs a test unrelated to the teaching: one needs a proficiency test rather than an achievement test. Thus test purpose should have an effect on test content and form.

Most arguments in favour of communicative language tests are concerned with the validity problem. However, validity is inevitably tied up with reliability: an unreliable test cannot be valid (although an invalid test can be reliable). If one concentrates on validity to the exclusion of reliability, it was pointed out, one needs to ask whether one is **measuring** anything, since measurement is quantification, and with quantification comes the need for reliability. There was general agreement that communicative language tests need to concentrate on improving their reliability. It was argued by some that this means taking the traditional 'pre-scientific' tests, and making them more reliable. One way of improving both validity and reliability of tests is to specify more closely both content and the criteria for assessment. It was felt to be still an open question as to whether communicative language tests have succeeded in doing this, to result in more adequate and successful tests.

One of the problems of communicative language tests is the problem of language in use: it is infinitely variable, being different for different individuals at different points in time. Systematisation (in terms of a theory or a description) seems highly unlikely, and yet desirable for test construction. Language, on the other hand, and more particularly grammar, is relatively systematisable, and therefore usable. In addition, although it may be claimed that communicative language tests are more valid because they relate to students' needs, such validity is relative, since it must depend upon the level of abstraction: what two engineers have in common may be different from what an engineer and a waiter have in common. Inevitably tests are about and for groups of people, not individuals. Levels of abstraction are likely to be higher rather than lower: but it was argued that if one abstracts far enough from a situation or task, one reaches grammar, which is what language learners will need whatever they are going to use the language for, and grammar is the level of language most amenable to systematic description (and therefore it was suggested, incorporation in tests). However, it was generally agreed that linguistic competence can only be a part of communicative competence: and that although one cannot ignore 'grammar' in communicative language tests, one cannot rely exclusively on it. The problem lay in defining precisely what else there is to test