Language Testing

edited by Brian Heaton

Language Testing

edited by Brian Heaton

Preface

The publication of this collection of articles on language testing comes at a very opportune time, as recent developments in communicative language teaching are now resulting in a widespread reappraisal of language tests and techniques. Not only have the many shortcomings of the structuralist and behaviourist approach to language teaching been strongly attacked, but the whole psychometric basis of language testing has been seriously questioned. Fierce arguments still rage over such established criteria as test validity and reliability. Consequently, it is now necessary to take stock of our present-day tests and examinations in English and to re-examine many of the basic premises so much cherished in the past. However, we should take care not to discard every well-tried and proven method of testing in our search for some magic formula or new technique which will enable us to solve our problems in the assessment of language used as communication. All too often in language testing, as in language teaching, there seems to be a tendency for many to jump on any current bandwagon, accepting half-formed theories and applying them hastily and uncritically. As most articles in this special issue of *MET* clearly demonstrate, considerable critical judgement is necessary in evaluating all the various types of tests and the principles on which they are based. It is always important to keep the best of established methods while, at the same time, seeking to develop where necessary new and more appropriate techniques to reflect the different emphases now being placed on language learning.

In the first article in the collection, **Carroll** points out that testing for any programme should be compatible with the ideas behind the teaching method used: hence communicative teaching programmes should be assessed by communicative tests. The main aspects of communicative tests treated by Carroll are: the test-curriculum relationship, a purposive test framework, test content and procedures, levels of performance and methods of analysing test data. **Davies** pursues this topic in discussing criteria for the evaluation of tests of EFL, and examines three kinds of validity with special reference to six examples of published tests in Britain. The question of validity and reliability is taken up again by **Underhill** in an article which identifies problems in assessing the productive skills of speaking and writing. Underhill uses the terms *direct, semi-direct* and *indirect* to classify tests of speaking and writing ranging from highly realistic tasks to unrealistic tasks. In the next article, **Fabian** reminds us that *studying* a language is vastly different from acquiring its communicative facility, and suggests that teachers can exert a beneficial influence on examining bodies by a more critical and creative participation in the design and construction of examinations.

However, the solutions to the many problems facing test constructors and administrators are neither as simple nor as straightforward as many might at first imagine. This is illustrated in an article by **Shephard**, who traces the development of the Cambridge EFL examinations from 1913, with their emphasis on the formal correctness of language use, literature and translation, up to a functional test currently under consideration, with its one-third oral component. Throughout the article, he shows the importance which the Cambridge Syndicate constantly attaches to public opinion, referring to questionnaires, correspondence, queries, and important public relations aspects of the Syndicate's work. **Seaton** also refers to the problems encountered in constructing examinations which will be administered on a large scale throughout the world. He describes the various difficulties met and overcome in the specification and design of the battery of tests set up and administered by the British Council and the University of Cambridge Local Examinations Syndicate.

Practical considerations in the use of progress tests by teachers and people directly concerned with the running of particular courses are touched upon by **Ward** in an article on the preparation and analysis of progress tests, while **Rogers** approaches this topic in a different way by providing several examples of ways in which test items can be devised so as to provide interest and even amusement on language courses.

It should be emphasised at this stage that a concentration solely on such aspects of communicative competence as authenticity, appropriacy and register is wrong if the testing of the grammatical system of the language is neglected and regarded as subordinate or inferior in any way. Language still consists of grammar, as **Rea** points out in her article on an alternative approach to the testing of grammatical competence within a model of language learning. She gives examples of ways in which the selection and production of a language form is determined not only by its grammatical correctness but also by its function within a given communicative area.

While it is always important to experiment with new testing techniques, it is also essential to attempt to develop existing techniques much further. After falling into disrepute in the 1960's, the long-established dictation test has been the subject of considerable research during the past decade. **Whitaker** examines the flexibility of dictation as a testing device and the various language skills involved, giving numerous practical suggestions for making dictation a relevant and realistic test. In a similar way, **Frantzis** touches on the skills involved in understanding spoken English, describing in detail the use of a radio news bulletin for improving such skills in a systematic way.

After discussing problems experienced by teachers in constructing tests of spoken language, Morrow gives some suggestions for devising appropriate realistic tasks, recommending the testing of students in groups. He then gives a number of criteria for evaluating spoken language before providing an example of the way in which these criteria are used in a basic level examination.

Following **Morrow's** article is a comprehensive account of cloze testing by **Johnson**, who questions several assumptions commonly held about cloze procedure. Johnson draws our attention to the intellectual, cultural and linguistic biases which may militate against attempts to measure a candidate's understanding of a particular text. He then proceeds to show how it is possible to identify much more precisely what each item in a cloze test is measuring, arguing that a random selection of items is not desirable. He adds that such a selection has in any case been largely responsible for the development of tests which are far too difficult and which do not discriminate sufficiently amongst candidates, especially as far as actual comprehension of the text is concerned. An article by **Nation** then deals with advanced reading tests and begins by examining multiple-choice items and reference-word items as techniques for assessing performance on advanced reading tests. These and other test items described in the article direct attention towards the structural features of a reading text and to analytical strategies. **Heaton** draws attention to the more communicative aspects of writing, first examining types of controlled composition before discussing briefly the classification of errors according to global/communicative errors and local/linguistic errors, and the implications of such a classification for the marking of compositions. He concludes by questioning the validity of composition tests which concentrate only on first drafts written within severe time constraints. This subsection is then concluded by **Boyle** with an article discussing various kinds of tests of language for students of literature. Examples of types of items testing reading, writing, listening and speaking are given in an attempt to foster an awareness among teachers of the relevance and educational values of the literature they teach.

Inter-dependence of verbal and non-verbal information both in the classroom and in the world outside suggests that visuals have a valuable role to play in language testing – provided that it can be ensured that candidates respond to language rather than find the meaning in any accompanying visuals. **McEldowney** shows how visuals compensate for fragmentary verbal comprehension questions on a text by helping to test objectively an awareness of the text content as a whole, without overt clues from the questions and eliminating the need for verbal production on the part of the candidate. McEldowney concludes by showing how visuals can also be used to test production skills, providing basic information for candidates in a non-verbal form and thereby avoiding situations in which prior knowledge of the subject is tested.

Godman deals with problems of assessing performance on academic subjects examined in the medium of English, providing specific examples of the types of difficulties encountered by examiners assessing science examination scripts written in English. In his article, Godman proposes three schemes for the scoring of such answers on a subjective scale, taking into account lexis, syntax, morphology and semantic content.

Pilliner next examines important aspects of the evaluation of language programmes, reminding us that student achievement is only one element in the process of evaluation. In showing that the fundamental purpose of evaluation is to produce information in order to make decisions about an educational programme, Pilliner cites the work carried out by Robert Stake, describing in detail the way in which his model for evaluation can be applied.

In a concise article with highly practical applications, **Chaplen** relates his experience of measuring student achievement in ESP programmes at Kuwait University. He shows how it is possible to convert raw scores by giving appropriate weighting to take account of the different components in a programme, and how a student's final grade can be calculated and decisions taken when several teachers are involved in the process of assessment.

It is only fitting that this collection of articles should end with some brief comments by Oller, whose research has contributed so much to the development of language testing over the past decade. Although maintaining that there appears to be a large general factor of language performance in all the various tests studied, Oller states that there is no basis to conclude that a single test such as a dictation or a cloze test is the best way to measure that general factor. He concludes that a multiplicity of testing methods concentrating on the kinds of language tasks which language users will be expected to perform makes the best language test in any given set of circumstances.

ACKNOWLEDGEMENTS

We wish to express our gratitude to the following for allowing the inclusion of copyright material.

SEAMEO Regional Language Centre for permission to reprint R. Keith Johnson's article *Questioning Some Assumptions About Cloze Testing*, which originally appeared in *Directions in Language Testing*, edited by John AS Read, Singapore University Press 1981; TESOL for permission to reprint Dr. John W. Oller's research notes *A Comment on Specific Variance Versus Global Variance in Certain EFL Tests*, appearing in TESOL Quarterly 14, 1980; the Joint Matriculation Board for permission to use two essay-type questions taken from their Test in English (Overseas), March 1968 and July 1968.

Table of Contents

Brendan J Carroll

Language Testing

Is there another way?

THE COMMUNICATIVE APPROACH

Over the last three years, I have been taking part in an evaluation consultancy for a Middle East teaching programme in which new communicative teaching materials are being devised to replace the structural materials in use there for many years. It is interesting to look back on an early report of the evaluation team which commented on the programme as follows:

> 'The communicative approach stands or falls by the degree of real life, or at least life-like, communication that is achieved in the foreign language in the classroom situation. In our observations of classes in progress, it was not often that pupils communicated naturally and unselfconsciously with one another. Perhaps this is not surprising and should not be expected at such an early stage in the project. However, it is worth stating that pupil-initiated communication should be one of the project's key aims, and the necessary provision should be made in the materials and teaching methodology that it comes about ..... We feel that there is still a tendency for teachers to talk too much and, correspondingly, for the pupils to talk too little and that a major aim of the teacher training programme should be to correct this imbalance. We noted that the teachers whose classes were visited dominated the flow of communication and allowed too little communication by and between pupils.'

'hese words are no doubt true and are often repeated by observers of classroom practice in many subjects and in many countries.

A second area of comment by the evaluation team which bears on our problem is the important question of *correctness* of children's language performance. The team mentions the tension between fluency of language use and accuracy of language usage, recommending a more sensitive and tolerant attitude to student error. Language accuracy and language fluency, it is maintained, should rightly have varying priority at various points in the teaching/learning sequence and the final aim should be that the learners perform both accurately *and* fluently to certain agreed levels.

In the programme we are discussing, there was an obstacle to achieving these enlightened aims of greater pupil activity and a balanced attitude to correctness in that the tests being used to measure the children's progress tended to be traditional ones such as those focussing on the accurate mastery of lexical and structural items by individual pupils. We thus had the position in which the children were being taught, as far as was possible, by one approach – a 'communicative' one – and being tested in terms of another approach – a 'structural' one. In truth, there were at the time no suitable, properly constructed tests for the programme, and this case is just one example of the dilemma we face at present. To test the accuracy of a learner's knowledge of lexical and grammatical patterns is a very different matter from testing the degree to which he has acquired the language so that he can use it in the communicative settings he is likely to face. The effectiveness of a person as a communicator will depend on a wide range of language and non-language skills, and an effective test will have to specify and assess them – a much more complex task than the assessment of his mastery of lexical and grammatical items, which can be much more easily pinned down and counted.

It was at one time widely believed that a person's language competence could be adequately tapped by requiring him to respond to a string of separate multiple-choice items, sometimes as many as 200 in one sitting. Later, it was hoped that the assessment needs could be met by presenting testees with the task of filling in randomly-selected spaces in a written or spoken text. We will examine more fully later the value of such techniques for testing language, all we will say now is: would that measuring communicative performance were so easily done! To approach the problem methodically, I would like now to examine major problems raised by the current structural-objective approach to language testing under five headings.

1. The counting of bits.
2. The four-skill model.
3. The place of correctness.
4. The role of purpose.
5. Justification by correlation.

Later on, I will put forward five considerations for broadening the approach to testing.

1. The counting of bits

If language performance is to be described by means of numbers, it would be most helpful if such performance could be broken down into small, discrete parts which could be easily judged as to correctness or incorrectness. Then the bits could be put together in a test to provide a numerical score, such as 35 out of 50. The tasks are unambiguous, the marking introduces no element of capriciousness and a person's final score is clear for all to see. Here is an example of such a test item:

> Yesterday I (A. be B. were C. was) very tired. (*Instruction – choose the correct option in the brackets.*)

Clearly, option C is a good candidate for choice according to standard English usage. A test made up of a chain of such items could be accurately and objectively marked, possibly by mechanical scanning methods.

So far so good. But the testing problem has not yet been completely solved. For one thing, it is often very difficult to decide if a sentence is correct or incorrect without knowing the context in which it was said. There are, for instance, certain communities in which *I be* and *I were* are accepted forms. For another, an utterance can be quite flatly incorrect from any formal point of view, and yet be perfectly intelligible to the ordinary listener. The French speaker who says 'I have been in London since three days' is certainly incorrect in his English usage, but there will be little disagreement among his listeners about the length of time he has been in London. Much will depend on the amount of tolerance we are prepared to extend to such a speaker. Finally, the suitability of any utterance will be closely related to the relationship between the speakers: whether it is very close and informal, or distant and formal, and so on; and what would be quite proper in one circumstance could be quite offensive in another.

Who, then, was our sample sentence spoken by, and who was the listener? Who are the mysterious 'she's', 'he's' and 'John's' of these pseudo-utterances? Where are they? When was 'yesterday'? The plain fact is that, from the point of view of the tester, it just doesn't matter; these are not real utterances at all, but just vehicles for providing lexico-grammatical traps for the unwary.

All this is just to say that an independent, decontextualised sentence is a very tenuous basis for making accurate judgements about a person's mastery of a language. It may well be that such snippets will allow us to examine certain limited details of language performance, but I believe that any adequate test must consciously encompass in its design wider strategies and purposes of language use. Adding up the separate bits of performance, however many, cannot tell us the whole story.

2. The four-skill model

One readily intelligible model of language description, and one widely used for many years, is that of the four skills of listening, speaking, reading and writing, certainly observable aspects of linguistic performance. It is tempting, therefore, to specify our test content in terms of these skills – productive/receptive and oral/graphic. We could, then, devise separate tests of listening, speaking, reading and writing, and thus map a person's language competencies. The trouble is that language is interactive, so that there is interplay between speaking and listening, between reading and writing, and so on. The total communicative situation cannot be adequately described by these separate categories. Furthermore, any test tasks must elicit some responses. Even an objective listening test required the testee to record some response – verbal or graphic; a speaking test must be a response to some verbal or written instructions. To resort to 'objective' techniques of ticking alternatives must trivialise the interaction, debasing one side of what is an essentially double-sided process.

The bases of four-skill assessment – usually spelt out in terms of pronunciation, spelling, vocabulary and grammar – have also been deficient insofar as they rely on features of formal usage rather than on effective use, and thus can be faulted on the grounds of the criticism we have already made of discrete-item, objective-type tests.

3. The place of correctness

We mentioned at the beginning of the chapter the matter of the *correctness* of children's language performance. There are few more emotive educational topics. In the one camp are the purists for whom any mistake – in spelling, grammar, pronunciation, punctuation – is regarded as a personal affront. To them the learning process boils down to the rooting out of errors, accompanied by indignant letters to *The Times*: 'What is our education system (or our country) coming to?' they cry. In the other camp are the permissive ones who have little time for rules, and who see any attempt to insist on their observance to be an assault on the liberty of the individual and his right to free expression. Merely to describe these two extreme positions in this way is to imply that somewhere between them lies a sensible attitude to correctness; the ultimate aim is to produce students who can perform both accurately and fluently to certain agreed levels of performance, and within agreed levels of tolerance. It is one thing to differ about ultimate aims regarding accuracy, it is quite another to differ about emphases given to accuracy and fluency whilst pupils are still struggling towards those ends.

What features other than formal accuracy might one consider in building up assessment criteria? It seems to me that we can work at three levels, macro-scopic, meta-scopic and micro-scopic along a spectrum from the broad effectiveness of the message in given settings, through the strategies used in achieving this level of effect, to the linguistic minutiae of spelling and pronunciation through which the strategies are realised. This three-level hierarchy, arbitrary though it may be, can provide a framework for a comprehensive assessment of performance. Below are listed a number of factors subsumed under each of the three levels.

Macro-scopic	*Meta-scopic*	*Micro-scopic*
purpose	strategies	spelling
message	interaction	pronunciation
setting	text	grammar
effectiveness	appropriacy	vocabulary
affect	style	correctness
	fluency	
extra linguistic	size of text	intra-sentential
performance	range of skills	
pragmatics	flexibility	
	cohesion	
	coherence	
(Communicative value)	inter-sentential	(Signification)

In one of our recently-framed writing assessment scales, a number of performance levels were spelt out in terms of this hierarchy. The macro-scopic features, referred to as *Message*, included clarity of presentation, coverage of points, validity of conclusion and the overall 'flow' of the work. The meta-scopic features, referred to as *Text*, included format and lay-out, coherence of theme, use of cohesive devices, appropriacy of style and neatness of appearance. And the micro-scopic features, referred to as *Language*, concerned the control of grammar, suitability of vocabulary, accuracy of spelling and intelligibility of handwriting.

Few would deny that such features as those listed are very pertinent to the measuring of language performance. To confine our assessment criteria to formal correctness would subtract greatly from the breadth and depth of our assessment. To sum up on accuracy, then, we can say that it is an important criterion, but by itself cannot form the basis for an adequate judgement of performance – it is necessary but insufficient as a criterion.

4. The role of purpose

I believe that any measurement of language competence which lacks a detailed and systematic specification of the purposes for which, and the contexts in which, the language is to be used by the person concerned is highly suspect. At the best, ignorance of such purposes and contexts will risk a waste of time and resources; at the worst, it will pervert the very object of our testing. And yet there are those who, believing that 'language is one', think it can be adequately tested by a single set of procedures and with a common content regardless of the specific aspirations of the individual testee; it is immaterial to them whether he requires to use the language to practise Medicine, to teach Architecture, to fly Concorde or to sell Jelly Babies.

But there must be *some* purpose behind a person's choice of the language to be learnt. If he is learning English, why is he paying his money? Why is he devoting considerable time and energy to the task? Why has he chosen English and not Telegu or Spanish? There must be some reasons for his actions; these reasons must be specifiable in some terms, however broad; and his purpose can thus be taken into account in the test content and procedures. Even if one were to accept the monolithic, unitary theory of language, there would still be massive practical reasons for devising context-sensitive testing, and surely no educationist would hope to earn extra points for deliberately ignoring discoverable factors which must be highly significant for his pupils.

5. Justification by correlation

In this section we will, unfortunately, have to turn to a discussion of statistical issues – not usually a popular topic, but certainly at the centre of any discussion of measurement. Even to those not interested in statistics, the following section could be useful.

The most commonly quoted statistic in discussing the analysis of abilities is the correlation coefficient. A high degree of positive correlation between traits is shown by a coefficient of +1.0, a high degree of negative correlation is shown by a coefficient of −1.0, and absence of correlation would be indicated by a co-efficient in the region of 0.0. In practice, however, human traits and abilities tend to be related positively, that is, a person good at one task (say, Maths) tends to be good at another (say, Essay Writing), and many correlations fall in the area from +0.4 to +0.9 or thereabouts.

Using such correlation coefficients, we may make a statement about language abilities going something like this: 'In our sample of testees, scores on Test A (Writing) correlate highly with scores on Test B (Reading). There is therefore no need to have a special test of Writing because performance on the Reading test will tell us what we want to know about the person's language competence and, moreover, the test of Reading can be marked much more objectively'. This argument is also extended to matrices of correlation coefficients which, when factor analysed, produce factor patterns of the abilities being studied.

To me, this argument is highly suspect, not only because of its practical, educational dangers, but also on theoretical grounds. When I was a teacher, I knew that Sarah Williams, aged 10, was likely to be at the top of the class, or near it, in Arithmetic, Geography, History and English. Had we introduced classes in Industrial Relations or Atomic Physics, I have little doubt that Sarah would have been at or near the top of the class in those subjects as well. By the same token, John Bennett would, in every test (except swimming), have consistently trailed behind the rest of the class. To conclude from such observations that it was thus necessary only to set one or two tests and even to teach only one or two subjects, such as Reading and General Intelligence, in order to foster and measure the children's ability would, I am convinced, be the height of educational naivety. If one were to cut out most courses in Medicine on such 'correlation' grounds, we would finish up with a singularly dangerous generation of doctors. It may well be that for rough, snap decisions we can reduce our testing to the lowest common denominator of elements but, if we want our tests to make accurate and sensitive decisions about our learners, we can ill afford to rely on this type of correlational thinking. We must decide *a priori* on educational grounds what our test must contain, and use statistics merely as an ancillary process to check whether the tests are doing what we wished them to do.

It is not always appreciated that correlations depend on co-occurences of variance in the responses of the sample being tested. If there is little variance in one or more traits, then their intercorrelations are most likely to be small. Furthermore, even if a correlation is sizeable, it is extraordinarily difficult to trace the precise reason for its being so. It is often found, for example, that it is similarities in testing method between the two tests rather than any relationship between the two traits which lie behind the correlation. Even with the use of factor analysis techniques, there are many ways in which a factor pattern can be explained. It is only recently that sophisticated methods of analysing factor patterns have been used in language studies (see Palmer, A. and Bachman, L., 1981). If these techniques are not used, unless every precaution is taken to identify sources of trait and method variance and unless discriminant as well as convergent features are taken into account, it is very easy to leap to unjustified conclusions from our correlationally handled data.

One particularly disconcerting assertion is one which says that such and-such a correlation is 'a good one'. But how can we tell what is a good, or high, correlation and what is a poor, or low, one? Is, for example, a correlation of 0.60 high, medium or low? And what about 0.85, or 0.52? There are three basic ways of answering these questions:

One: to ascertain whether the coefficient is statistically significant and unlikely to be due to chance. A coefficient of 0.60 with 100 subjects would be highly significant in that the odds against it having occured by chance are better than 100 to 1. Clearly, then, there is almost certainly something behind a correlation of this kind.

Two: to calculate the percentage of shared variance between the two variables. This is done by squaring the correlation, so that the 0.60 correlation indicates that 36% of the variance is shared. By this criterion, the correlation is looking decidedly shaky as we have now left unexplained no less than 64% of the variance.

Three: to estimate the forecasting efficiency indicated by the correlation. Using Kelley's coefficient of alienation, we find that a correlation of 0.60 has a forecasting efficiency of only 20%; that is, we are only 20% better off than we would have been if guided by pure chance. (Indeed, even a 0.90 correlation is only 56% better off than pure chance.) Our 0.60 correlation is now looking most decidedly chancy.

All this goes to show that it would be wise to treat apparently 'high' correlations with the greatest caution. A correlation is useful for comparative exploratory studies, but too flimsy a basis for absolute value assessments. And, if the individual correlations themselves are open to question, so much more so are the elaborate factor analyses based on them. This is not to say that producing correlational statistics and analyses is necessarily a fruitless operation; correlational approaches have their contribution to make in exploring the nature of linguistic-communicative transactions. By all means, let us use them whilst recognising their limitations, but let us not give up the more radical probing of the social, psychological and linguistic entities involved in inter-personal communication. There is a place for descriptive and observational studies as well as quantitative, experimental ones.

We have now discussed five problem areas associated with traditional language testing and reach the following conclusions:

a What we need to do is to look at the whole field of communication in devising our tests and not restrict our task to the counting up of easily-devised and easily-assessed bits of language performance.

b We need to give systematic consideration to purpose and strategy in our task design and not just rest with the simplistic framework of four de-contextualised skills.

c Our important assessment feature of accuracy, or 'correctness', must be put in the context of a range of broad and context-specific criteria.

d We need a direct study of the dynamics of linguistic-communicative behaviour and should not just rely on the circumstantial and *post hoc* evidence of correlational statistics.

The question now is – if a new emphasis is needed – what should it be, and will it produce workable tests? It is all very well to list the alleged defects of a current method of setting about things; it is also expected that the critic outline an alternative approach and show that it has the potential of producing better results. *Is* there another way?

THE COMMUNICATIVE ALTERNATIVE IN TESTING

The main aim of the new approach must be to widen the basis of our tests from a narrow grammatico/statistical focus towards a broader, multi-disciplinary and multi-level approach which can yet maintain essential features of measurement, always remembering that language testing is much too important to be left to grammarians and statisticians! The way we achieve this broadening, or opening up, is outlined below under five headings – to help provide answers in the problem areas already discussed:

6. The test-curriculum relationship
7. A purposive test framework
8. The test content and procedures
9. Levels of performance
10. Methods of data analysis

6. The test-curriculum relationship

The close relationship of tests with other elements of the curriculum has already been touched on. One way of ensuring their relevance to each other is to see that they both come from a common source, such as a specification of linguistic-communicative needs; thus the programmes of the curriculum aim to meet those needs, and the tests aim to indicate how far the needs are being satisfied. In this way, there is no 'general proficiency' test, as this title would imply lack of specification of the language skills actually needed. Nor would there be a separate 'achievement test' element, such as when a test springs exclusively from a syllabus, because the spelt-out needs are there for direct use in both syllabus and test design.

Without some spelling out of language-communication needs, and of related student aspirations, it is all too easy to base the language programmes on existing tests or examinations, which are usually described in the vaguest of terms, so that students spend their time either on past examination papers or on a course book designed to help students to pass the examination with the minimum of effort. A recently published book of tests, for example, contains strings of items (disguised here for security reasons) such as:

(a) Her mother died when she was young, so she was . . . . . by her aunt.
(4 options presented, the accepted one being 'brought up')
(b) The lion fell into the . . . . . set for it.
(c) When he . . . . . the age of 65, he will retire.
(d) I stood in a . . . . . for half an hour to get my theatre tickets.
(e) Roses are the . . . . . flowers in our garden.
(f) You could tell from his big feet that he . . . . . his father. (*etc*.)

The hapless examinee, had he not been conditioned to such a sequence, might well ask who these mysterious 'he's', 'I's' and 'she's' are, and how he can plausibly be expected to consider in one breath a dead mother, a retiring employee, an unfortunate lion, a slow-moving ticket queue, a rose garden and a boy with big feet. Apart from the inherent absurdity of this juxtaposition of topics, the testee has to struggle through masses of options which are either inappropriate or quite incorrect in the context of the sentence.

Nor is the suspension of disbelief much less in demand with those so-called integrative tests which ask the testee to insert suitable fillers in texts from which words have been eliminated at random. Although much in vogue at present, and justified on correlational grounds, such tests are only a little more credible than the separate-item test described above. Who, for example, would write a book, report or letter omitting every 7th word, expecting the omissions to be remedied by the unfortunate reader?

If we are keen to see that our tests, in their content and tasks, are meaningful in themselves and can be seen to reflect features of the settings in which the testee will operate, we will, I fear, have to look further than these barren, eduationally-destructive techniques.

7. A purposive test framework

To meet learners' needs and to ensure a benign relationship between the test and other elements of the curriculum, a justifiable approach is to make a clear and detailed statement of the purposes and settings of language use, and of the skills and functions to be called on, and from this statement to generate the language, content and tasks which a comprehensive test will have to encompass. All these design operations must be carried out without losing sight of the specified purposes for learning and using the language.

The framework for approaching our test design may be adapted from such models as those of the Council of Europe (see Van Ek, 1975) and J. Munby (see Munby, J., 1978), to name but two. One of the problems of these models is that they are extraordinarily detailed and lengthy, and it is difficult to fit even a fraction of the specified elements into a test of anything like an acceptable length. Thus, a decision has to be made at an early stage in test design as to how accurate and how delicate, or detailed, the instrument must be. If we required a quick decision for rating students in broad categories for programme placement purposes, if any misplacements caused by the tests could be remedied easily and if the applicants had a wide range of purposes for learning the language, then a fairly short general test of competence could well fill the bill. If, however, the decisions to be made are very refined ones, if the chances of remedying them are small, and if the applicants' language needs are very job-specific, then a detailed needs specification, careful test development and an allocation of time and resources for test application will be called for.

It will, therefore, be for the organisers of the testing operation to decide what scale of delicacy, what tolerance of mistakes and what resources shall be devoted to that operation. The object is to devise tests which will give the necessary answers more economically.

To assist in making a reasonably exhaustive test needs analysis, we have prepared working papers under the following headings which we illustrate by excerpts from an actual specification for a test in English for Computer Programmers.

JOB: Computer programming (English for occupational purposes)

1. Main Events, with their Activities

I Attending classes in principal subjects
- listening to lectures
- observing demonstrations
- taking notes for further study
- questioning and discussing lectures/ demonstrations

II Making field visits to computer centres
- understanding functions of computer centre
- touring various divisions
- observing sample running programmes
- questioning programmers in centre
- reporting conclusions of visits
- discussing results in groups

III Studying at home and in library
- carrying out reading assignment
- doing written exercises
- writing reports and critiques
- preparing for classwork and examinations

IV Carrying out practical computer assignments
- selecting appropriate guidance information
- testing and amending each phase
- general run of programme

Note: the *events* are the main focusses of work during the course, the *activities* are component parts of the events. Further specification is done in terms of each event.

2. Language skills (for Event I)

50 - 1 and 2*. Transcoding information in diagrammatic display involving conversion of diagrams/tables/graphs/ into speech or writing.

48 - 1, 2 and 3. Maintaining discourse; how to respond, to continue and adapt as a result of feedback.

40 Extracting salient points to summarise: the whole text, a specific idea or topic, the underlying point of the text.

39 Distinguishing the main idea from supporting details by differentiating: the whole from its parts, fact from opinion, and a proposition from its argument, (and so on for this and the other events).

*The skills believed to be critical to each event are specified along the lines above, in this case in terms of the Munby model (as described in Munby, J.L., 1978 and Carroll, B.J., 1980). The numbers are those in the Munby taxonomy of language skills.

3. Language functions

To appreciate the functions:
certainty, probability, conjecture, intention, obligation, evaluation; to inform, report, agree, endorse, prove, assume, ratify, conclude, generalise, demonstrate, explain, classify, define, exemplify.
(Taken from the categories of function as described in the two references above.)

4. Topic areas

As in computer programmer training syllabuses:
- special English for computers
- introduction to computer hardware
- advanced Mathematics
- systems analysis
- COBOL
- management
- human relations.

Text books and instruction manuals will contain the details of the above topics and provide appropriate samples of the language likely to be needed. These topic areas may be general for the programme or allocated to particular events if appropriate.

5. Medium

The events are now categorised according to the medium involved, i.e. listening, speaking, reading, writing and mixed media.

6. Target performance level

On a 1-9 Band system, the level of performance for each medium is specified. In this case, a minimum target level of Band 7 in all media was established (i.e. 'Good user' on a continuum from 'Expert user' to 'Non-user' – see below).

7. Channel

Selected from the following channels:
Face-to-face, phone, radio, TV, film, tape, groups, public address, print, telex, print-out.

8. Other parameters

- *socio-cultural*: (roles and relationships)
  learner — instructor
  insider — insider
  native speaker — non-native speaker
  professional — professional
- *attitudinal tones*
  assenting — dissenting
  cautious — incautious
  formal — informal
  friendly — unfriendly
  inducive — dissuasive
  respectful — disrespectful
  patient — impatient
  certain — uncertain
- *dialect*
  Understand British English dialects including RP or near RP accents.
  Produce standard English dialect of own area with appropriate regional accent (but generally intelligible to colleagues).

By compiling the above compendium of communicative needs, the tester will have gained considerable insights into the demands of the computer programmers' course, its content, the various activities to be undertaken and the language skills and functions which will have to be mastered to the appropriate target level.

8. The test content and procedures

The events and activities already specified will provide the basis for the balance of tasks contained in the actual test, and the topic areas should point to the main semantic fields to be covered. Similarly, the other parameters of the eight-fold specification will give guidance about skills, functions, tones, dialect and so on. Of course, it will not be necessary, or possible, to provide in any one test a complete coverage of all the specified features but, having selected the main events to be replicated in the test, we will be able to include key features implied by these events.

Test items themselves will fall broadly into three main categories:

Open-ended items for which the testee is allowed a fair measure of latitude in carrying out the task, his performance being assessed on a graded scale probably with accompanying actual samples of different levels of performance.

Closed-ended items, where the testee selects from a given set of responses the one he considers the most appropriate, from a 'Yes–No' dichotomy to anything up to five possible options.

Restricted-response items, which allow a response to be composed by the testee, but on very restricted grounds. Probably the answer will consist of one or two words or, at the most, of a short sentence.

The most effective test instrument will contain a good balance of the above three types of item, allowing the authenticity of the open-ended tasks to be supported by the objectivity (in scoring) of the closed-ended items, bearing in mind the serious limitations of such 'objective' items discussed in earlier sections.

The computer programming test derived from the above specifications included tests of:

Listening Comprehension, using lecture texts with multiple-choice items.

Group discussion, with individual roles provided by guide cards; performances to be assessed by observers according to a graded scale.

Study skills, in which a range of information was provided both in ordinary English and in computer language (with explanations), using multiple-choice items.

Report Writing, responding to written and oral presentations of information; rated by assessors according to a graded scale, with examples.

We have found that, by these test design and development techniques, we can gain insight into the language needs of specialist groups of testees and give ourselves an opportunity of selecting an authentic basis for devising test tasks and items.

9. Levels of performance

For centuries, examiners have made use of scales, or grades, of performance describing successive levels of behaviour in the particular areas being examined. We have, for example, examinations in Pianoforte which range from Beginners to Advanced Level specifying as many as 12 levels of piano piece and the kinds of performance an examiner is to expect when awarding grades. Although basically subjective, this examining procedure is surprisingly reliable and has a good backwash effect on the piano-playing of the pupils. To attempt to improve reliability by breaking up a piece of music into small countable bits would, of course, lead to all sorts of musical absurdity, as the unit of performance *is* the piano piece, which must be judged as an artistic whole.

In language assessment, probably the most widely-used scale has been that of the American Foreign Service Institute (FSI) oral interview, which has proved to result in reliable, consistent judgements whilst retaining some of the naturalness of oral interaction. In the same tradition, we have devised scales of all sorts of communicative activity described at several levels according to the main parameters of level, activity and descriptor. The *levels* range from 1 to 9. The *activities* are, in effect, important communication focusses as described in our category of event/activity above, and can be based on the 'four skills' categorisation if appropriate. The *descriptor* consists of a label (e.g. 'non-writer' through to 'expert writer') followed by a brief thumbnail sketch of the critical features of a typical performance at that level, which can be elaborated in terms of up to a dozen performance criteria for detailed design purposes. The scales are usually accompanied by photostats or tape recordings of performances judged to be representative of each level. For objective, scored tests, conversion tables are provided to convert raw scores into levels on the 1 to 9 band system.

Initial figures for reliability and validity for such tests compare very favourably with those for current widely-used language tests and, more important, authenticity and testee motivation are much more in evidence.

10. Methods of data analysis

A perennial problem in measuring human behaviour is how to describe the target behaviour unequivocally, and to assess accurately how far an individual reaches, or falls short of, that level. The demands are for valid, systematic description, for accurate observation of responses, and for reliable measurement and data analysis procedures. In the main, language-communication tests have suffered on all three counts, description having too often been vague or based on limited, linguistic, features of performance, observation methods being such as to destroy the very communication processes under scrutiny, and data handling being focussed on counting the bits with analysis by correlational means. Thus the measuring devices have tended to demean the very processes they were trying to measure causing the results, however precise in appearance, to be of very little significance.

In the absence of direct measurement based on carefully spelt out, relevant criterion behaviour, much reliance has been placed on sampling and probability theory, the scores obtained from selected samples of performers being arranged in order.

Crucial importance is given to measures of central tendency (mean, median, mode) and dispersion (standard deviation, range, quartile and mean deviations). Norms of performance are established on internal criteria, a 'good' score being expressed in terms of standard deviations above the mean, a poor by standard deviations below the mean; or by deciles or centiles. Thus the most accurate information obtained is about where in the population any individual lies – in the top 10%, or at the 50% point, and so on. What are so often lacking in such approaches are detailed descriptions of test content and of the exact nature of a 'good' and a 'bad' performance. Briefly, the basis of performance assessment has been the relative performance of individuals in a sample rather than on links with pre-determined behavioural criteria. We thus have the constant danger of vague and circular reasoning behind our testing.

It would be wrong, however, to decry these 'norm-referenced' techniques because, as methods of exploration, they can give us many initial insights into behavioural problems. For instance, some years ago it began to be suspected that cigarette smoking and lung cancer were related. What doctors wanted to discover was the actual causes behind the observed correlations of cancer and smoking. So there came a time when correlational, circumstantial evidence had to be reinforced by more precise, and direct, techniques for describing, observing and measuring the related phenomena. In short, it is crucial that we move as soon as possible to a more authoritative test basis and do not remain shackled in the permanent relativity of norm-referencing.

CASE STUDIES IN COMMUNICATIVE TESTING

There are now interesting developments in the direction of communicative testing in various parts of the world, and I will mention some which I have been connected with. Others are no doubt being reported on as I write.

The Royal Society of Arts examinations in the communicative use of English. (RSA London, 1980).
The English Language Testing Service set up by the British Council and the University of Cambridge Local Examinations Syndicate, (descriptive Handbook in preparation, 1981).
The 'Crescent' English Course and the Mid-East projects associated with it. (O'Neill, T. and Snow, P., Oxford University Press, 1979 onwards).
British Council Course in Testing, testing for communicative programmes, Course 040 (report in preparation by the British Council, 1981).
The Pergamon English Tests in both general and specific areas of English, (in production by Pergamon Press, Oxford starting in 1981).

It is suggested that the above reports be referred to and, when early versions of the tests are released from security, the actual tests be studied. It would be a rash person who would claim to have resolved the incompatibilities between the terms 'test' and 'communication', but at least we can claim to have made genuine efforts to produce tests which contain identifiable communicative features.

Pertinent references

Canale, M. and Swain, M. (1980), 'Theoretical bases of communication approaches to second language teaching and testing', *Applied Linguistics 1.1.*

Carroll, B.J. (1978), *An English language testing service: specifications*, The British Council, London.

Carroll, B.J. (1980), *Testing Communicative Performance*, Pergamon Press, Oxford.

Carroll, B.J. (1980), *Communicative tests for communicative programmes*. Paper given at TESOL Convention, Detroit, 1981.

Fishman, J. and Cooper, R.L. (1978), 'The socio-linguistic foundations of language testing'. In Spolsky, B. (Ed.), *Advances in language testing series No. 2*, Washington D.C. Center for Applied Linguistics.

Morrow, K.E. (1977), *Techniques of evaluation for a notional syllabus*, Centre for Applied Language Studies, University of Reading. Study commissioned by the Royal Society of Arts.

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

Palmer, A.S. and Bachman, L.F. (1980), *Basic concerns in Test validation*. Paper for RELC Seminar of April 1980 (to appear in proceedings). More recent developments of this project were reported in TESOL, '81, Detroit.

Alan Davies

Criteria for evaluation of tests of English as a Foreign Language

In this paper I want to make some general comments about test choice and test construction, and then to discuss six examples of (British) published tests referring particularly to use in an English as a Foreign Language testing situation.

I plan to begin by making a series of very large but, I hope, gnomic generalisations. Even if nothing else in this paper sticks, I hope that one or two of these statements will.

1. Most linguistics is normative.
2. Language teaching samples by guesswork.
3. Language testing levels are intuitive.
4. Criterion and norm referenced tests are the same thing.
5. Language teaching operations need language level systems.
6. Language level systems need external validation.

Let me examine each of these briefly.

1. Most linguistics is normative

The rules of grammar, the observations of language acquisition and discourse analysis, the selections for language teaching and testing, all present language as if it were unvaried and thus they are all a kind of necessary idealisation: they all set a limit on what the language is (and by definition isn't). Typically, all such descriptions, all such norms, hold up internally, but, as we shall see, what is urgently needed is some measure of external validity. Here, paradoxically, it is easier for the more applied parts of linguistics in language teaching and language testing to provide such buttressing for the hypotheses of descriptive linguistics.

2. Language teaching samples by guesswork

This is even more of an exaggeration, but it is likely that, until we know far more about second language acquisition, the construction of a syllabus and a textbook will be more art than science. It is not surprising that so many language teaching materials should resemble one another. What is noteworthy for language testing is that *first*, the achievement test cannot improve on the syllabus: the test simply reports, it doesn't teach; *second*, and more optimistically, the test can be used as a check on the guesswork. I do, of course, admit that in the sense in which I have used guesswork, language proficiency tests are also guesswork; but it does make sense, I suggest, for new courses, syllabuses, etc. to be piloted – which they are – and tested – which they are only rarely. Testing and teaching have the same interests if not the same purposes.

3. Language testing levels are intuitive

I am thinking here of the ease with which examiners allot test papers to categories (High, Mid, Low, etc) and then, commonly, convert into some numerical score. It is part of the professional judgement of the teacher to grade thus. Even when we provide the apparatus of an itemised test which distributes students widely, it is still necessary to decide the *meaning* of a particular score or band. Sometimes this is done internally (or historically): such and such a score will allow 50, 60, 70, 80% to pass (or fail) and that's what is wanted or how it's always done. Or it may be possible (it is a pity it isn't done more often) to make use of expectancy tables and an external criterion of some kind, and show the effectiveness of a particular cut off – this is using a norm referenced test for criterion referenced purposes.

4. Criterion and norm referenced tests are the same thing

Again an overstatement. But what I mean is what I have just explained. *Intuitively* examiners do work to a Pass/Fail (etc.) criterion, i.e. we always do work in a criterion referenced way. Imposing a series of equal interval scores (1–10, 1–100, etc.) just extends the curve and distributes types of Fail and Pass. So from this point of view a norm referenced test is a use of a criterion referenced test just as we have noted the reverse. They do the same job, in the one case emphasising the distribution, in the other the cut off.

5. Language teaching operations need language level systems

What criterion referenced test uses may have, but what language level systems lack, is a criterion or