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

SECTION 2

SPECIFICATIONS FOR AN ENGLISH LANGUAGE TESTING SERVICE

Brendan J Carroll, The British Council, London

The Testing Problem

1 The present testing system, devised in the earlier half of the 1960's, was in its time a well-thought-out and convenient instrument. Over the years, however, there have been great changes both in the size of the placement problem and in approaches to language test development.

2 The number of applicants for training in Britain has grown out of all recognition over these years. At the same time, there has been an expansion in the range of courses of study required, with increasing emphasis on the applied technologies and on non-university courses and attachments which the earlier test had not been designed to accommodate. This increase in numbers reflects both an emphasis on manpower training schemes under aid programmes and the growing wealth of oil-producing countries in West Africa, South America and the Middle East.

3 Over this period, language teaching and testing methods have shifted their emphasis from atomistic language features, such as uncontextualised phonemic discriminations ('hit – pit') to broader features of linguistic communication. The trend now is, as exemplified in the present report, to postpone consideration of language realisations until the communicative needs of the users have been clearly determined, broadly-speaking a socio-linguistic approach.

4 The trends noted in the previous paragraph have also encouraged the development of programmes in English for Specific Purposes (ESP) so that fewer people are now engaged in devising tests and teaching programmes which aspire to meet equally well the needs of all users, regardless of the purposes for which they will need the language.

5 A recent breakdown of a large group of applicants for courses of study in Britain gives as the five most important categories:

Agriculture (including Fisheries, Timber, Vets.)
Engineering, Medicine (including Dentistry),
Economics (especially re Development) and
Public Administration.

Our problem is not just whether the present test can encompass the needs of these, and many other, diverse study courses, but whether **any** single test can do so. And we have adopted the hypothesis that the solution to our testing problem, and the way to improve the testing service, is through a process of diversification of test instruments to meet the diversity of the test situations.

6 The language test system so developed will have to provide information which will enable us to answer two important questions about any applicant – whether he is already likely to be able to meet the communicative demands of a given course of study or, alternatively, what would be the nature and duration of the course of language tuition he would need in order to reach the required competence level. In designing our testing service, then, we will need to specify the communicative demands of a variety of courses, of different levels, types and disciplines, and to devise workable instruments to measure how far applicants can meet those demands. We must, in doing so, effect a demonstrable improvement on the present system and ensure that the new test itself is capable of continual monitoring and improvement.

Compiling the Specification

1 Purpose of the Specification

Our purpose in compiling the specification is to build up profiles of the communicative needs of a number of students on study programmes in Britain in such a way that we will be able to identify common and specific areas of need upon which an appropriately diversified test design can be based. It is of crucial importance that at this stage our focus is on **the communicative demands the programmes make on the participants.** As we have already said, we will bring to bear on the test design important operational considerations affecting the administration of the test service, but it must be emphasised that such considerations, however pressing, will not make the communicative needs of the participants disappear. We would hardly be likely to achieve our aim of test improvement if we ignored a patently essential communicative need merely because it entailed practical problems.

2 The specification framework

Each specification will provide information about the communicative needs each participant will have in studying his programme and in living in an English-speaking community. The specification parameters are:

0 **Details of the participant;** a minimum amount of potentially relevant information about identity and language

1 **Purpose of Study;** establishing the type of English and the purpose for its use in the programme.

2 **Settings for English;** including both physical and psychosocial settings.

3 **Interactions involved;** identifying those with whom the participant will communicate in English, his position, role relationships and social relationships.

4 **Instrumentality;** the main activity areas — receptive/productive, spoken/written; the channels, face-to-face, print or radio for example.

5 **Dialects of English;** whether British or American English; which regional variety, both for production and reception. Any dialect variations regional, social or temporal.

6 **Proficiency Target Levels;** expressed on a scale from 1 (low) to 7 (high) related to the dimensions of text size, complexity, range and delicacy, and the speed and flexibility of handling it; tolerance conditions expressed on a scale from 1 (low) to 5 (high) related to tolerance of error, style, reference, repetition and hesitation.

7 **Communicative Events and Activities;** the description of what participants have to do, such as 'participating in a seminar' (event) and the parts of those events that assist skill selection later, such as 'putting forward one's point of view' (activity)

8 **Attitudinal Tones;** concerning *how* an activity is enacted; derived from an index of attitudinal tones - sets of antonymous continua such as 'formal-informal'.

9 **Language Skills;** a taxonomy of 54 skill categories, with their component skills, ranging from 'Discriminating sounds in isolated word forms — allophonic variants' to 'Transcoding information in speech/writing to diagrammatic display'.

10 **Micro-Functions;** as exemplified in sub-categories of function; units of meaning between the level of 'activities' and their linguistic realisations, such as the micro-functions of persuasion, advising, invitation.

Note: The specification data in Appendix A are arranged under the section headings, 0 to 10, as above.

3 Areas of specification

English Language Division staff members have prepared specifications of participants in each of the following six areas:

P1 Business Studies (HND)
P2 Agricultural Science (Post-Graduate)
P3 Social Survival (Academic)
P4 Civil Engineering (BSc)
P5 Laboratory Technician (Trainee)
P6 Medicine (FRCS)

Specifications P1, P4 and P6 are for fairly typical English for Academic Purposes (EAP) course participants. P3, Social Survival, relates to the social needs of the average student on an academic programme. P4, Laboratory Technician, is a good example of a sub-University trainee in a non-degree study atmosphere. P2, Agricultural Science, is an unusual but not impossible case where a student requires English almost entirely for the study of reference literature as, being on a two-way programme attachment, he mixes mainly with speakers of his own language or with English staff who speak his language.

It will be seen that a good range of levels and programme types has been included in our sample, although we do not pretend to have covered a representative range of the total population. We hope, however, to elicit from this participant sample, major design factors applicable to test development.

4 Specification data sources

Although it would be desirable to derive our data from comprehensive observational studies of the participants actually engaged on their courses, we decided that less time-consuming methods would be sufficient to assess the basic adequacy of our approach to test specification. **The ultimate validation of our methods would be in the effectiveness of the tests based on their results.** To ensure the best insights possible into this interdisciplinary problem we adopted the following procedures:

a Compilers

The compilers of the profiles were chosen according to their special interests and backgrounds. For example, the Business Studies specification involved two staff members one of whom had published a course in Business English, the other had a degree in Commerce and had lectured in Economics and Accountancy to adults. The Social Survival English profile

was compiled by a member of staff who was actually teaching the student concerned on a pre-sessional English course. The Medical profile was prepared by a staff member with considerable experience in teaching a University Medical English course and who had close family connections in Medicine.

b Contacts

All staff concerned made contact with institutions and/or individual lecturers in the disciplines concerned. The Laboratory Technician profile was compiled in discussion with our Technical Training colleagues and in contact with staff and members of a course currently being conducted for Laboratory Technicians. The Civil Engineering profile was prepared by an officer who had earlier done a study of Engineering courses and teaching methods in Britain who was advised by two colleagues in Education and Science Division with appropriate degrees and experience. It is intended that close and continual contacts of this kind will be maintained throughout the process of test development and validation.

c Documents

Continual reference was made to authentic documents in the disciplines such as college handbooks, course syllabuses and standard subject textbooks. We found the widely-used titles circulated under the Low-Priced Text Book Scheme to be of particular value in this respect. To exemplify the exacting demands of the programmes, we include in Appendix D the published requirements for a preparatory course in Civil Engineering.

In general, we believe our data collection methods represent a reasonable compromise between what would be theoretically perfect and what could be done in an acceptable time-scale with resources to hand.

Results of the Specification

We will now examine, in parallel, the results of the six specification studies with the purpose of identifying the essential communicative demands on all the participants. This examination should enable us to identify three levels of communicative demand — those common to all (or most) of the participants, those shared by some groups of participants and not by others, and those specific to an individual participant. In factorial terms we should obtain broad indications of the presence of general, group and specific factors. This information is essential if we are to make firmly-based recommendations about test diversification. Please note that it will not be possible to follow the discussion of results given below without constant reference to the appropriate sections of Appendix A.

0 Details of the Participant

Our purpose in personalising the profile is to focus the collection and interpretation of data on a real, or at least a putative, individual so as to counteract the natural but dangerous tendency to overgeneralise about communicative needs. We are in fact using a simple case-study approach to data collection. Now if we look at Appendix A at Spec. O, the Participant, we see details of our six participants P1 to P6 as regards age, nationality, language and standard of English. The Ps cover a range of countries and native languages, with a certain bias towards Muslim countries, their ages range from twenty to thirty, and their level of English is of Intermediate or Upper-Intermediate standard. It is worth considering at this stage to what extent our sample of course participants is, or needs to be, representative of the total population of candidates for our tests. In earlier approaches to testing, it would be considered necessary to ensure that the sample was representative of the population of candidates as a whole, and the statistics of probability would be used to measure the characteristics of that population; in other words the approach would be 'norm-referenced'.

In our present approach, however, we are starting from the specification of the communicative demands of target courses. Once these demands are defined, it is for us to decide whether a particular candidate has met them on the evidence of his test performance; it is not a matter of primary importance to us how performance characteristics are distributed throughout a population of applicants many of whom, we now know, are likely to be 'non-starters' about whom we are not required to make refined, or delicate, decisions. Our approach, then, is basically 'criterion-referenced' and our performance standards will derive from ongoing courses and their students. In our recommendations, we will put forward proposals which take into account the existence of these 'non-starters'.

1 Purpose of Study (Appendix A, Spec. 1)

We see from the information given that two of the participants are engaged in post-graduate study, two in undergraduate study and one in sub-university training. One of the specifications, P3, does not have a training focus. There is a fair range of disciplinary studies — Medicine, Agriculture, Business and Applied Technology. We are not, of course, centrally concerned with the disciplines as such but with the communicative demands their programmes make on the students, and their consequential communicative needs. It will be a matter of great interest to discover how far disciplinary domains coincide with or diverge from communicative domains.

2 Settings for English (Appendix A, Spec. 2)

It is immediately obvious that although there is a variety of programmes there is considerable uniformity in their physical settings. In all instances, we find the Lecture room, Seminar room and the Library or Study centre. There is a general need for practical or field work – on site, in industry or in the casualty ward. For the more technologically-oriented participants there is a need for work in the laboratory, workshop or operating theatre.

The Agricultural Science student, whom we have already discussed as the odd-man-out regarding study needs, will use his own language extensively except for reference reading and use English in a restricted range of settings. And all students, however retiring their nature, will be living in English-speaking communities with Social Survival requirements as outlined in the P3 profile.

The temporal settings indicate that, again with the exception of P2, English will be used many hours a day in term time and even extensively in vacations. It is salutary to realise how heavy this avalanche of language demands is for students who so often have had little practical experience of English as a communicative tool, who are confronted with new approaches to their subject and who come from a cultural background very different from, and even inimical to, their new environment.

3 Interactions (Appendix A, Spec. 3)

The importance of interactions for our participants is shown in the variety of relationships recorded in the specifications. The most commonly-mentioned interactions, both within the programme and outside it, are:

Learner-instructor (and, for the Medical student, vice versa)
Professional-professional (in mixing with course staff and members)
Senior-junior (possibly re age, but more probably in the academic context)
Outsider-insider (as a foreigner, and as a newcomer to his groups)
Insider-insider (within national, student or academic groups)
Adult-adult (none of the P's has a major concern with children)
Man/woman-man/woman (in permutations)
Equal-equal (both socially and academically)

The largest range of interactions occurs in P6, the Medical participant. As a senior medical person, this participant is by turn lecturer, adviser, therapist and leader as well as having a student role. The Laboratory Technician, P5, will also occupy a non-student role and, as an older and more experienced

person, will be occupying non-student positions in preparation for his future duties as trainer and supervisor. It is thus important to realise that some of the trainees and scholars do not come to Britain in a humble role of tutelage but are likely to be put in positions of professional and personal leadership for which they must be linguistically fitted if they are not to suffer grave loss of face.

4 Instrumentality (See Appendix 1, Spec. 4)

We can see that both receptive and productive skills and spoken written media are required. We will see from the next section that the relative importance of the four communicative media (listening, speaking, reading and writing) will vary considerably from profile to profile.

The main channels are the conventional ones of face-to-face and print. With the increase in use of modern mechanical devices, we must also consider the use of sound and video tapes, audio and video cassettes, radio, television, the telephone and public address systems. This variety of channels contrasts with the very restricted range commonly used in language testing and suggests the possibility of widening the range of test presentations.

5 Dialect (Appendix 1, Spec. 5)

The common need is for contemporary English (Historical or Literary studies might have provided exceptions). The participants will need to understand varieties of standard British English and local varieties of English to be heard in their area of residence. They will be expected to produce intelligible and acceptable standard English varieties of their home region (eg West African), probably with a local accent (eg Northern Nigerian). The main basic requirement will be a certain flexibility in understanding a range of English accents and the ability to produce a variety of English intelligible to the other members and the staff of their own course.

6 Target Level (Appendix 1, Spec. 6)

In specifying the target level we need to know for the first dimension (size) the size of the text the participant will have to handle, for the second dimension (complexity), the complexity of the text, and so on for each of the six variables listed in Spec. 6. Each of these dimensions is assessed on a 7-point scale from very low (1) to very high (7) and derived from the purpose of study and the type of interaction for the participant.

The participants' situation may also allow various degrees of tolerance of error, stylistic failure, use of reference sources, repetition or re-reading and

hesitation or lack of fluency. This tolerance is assessed on a 5-point scale from low (1) to high (5) tolerance. It must be admitted that the assessments given by the compilers were subjective ones and we have not yet been able to calculate the reliability of the rating system. We must therefore not read too refined an interpretation into our analysis.

a Verbal Medium

For purposes of comparability we have used percentages (rather than a 1 to 7 scale) to express the averages of the dimension ratings in Spec. 6. For each participant we give the average percentage rating for each of the four verbal media: Listening, Reading, Speaking and Writing, as well as the averages for each row and column, in Table 1 below.

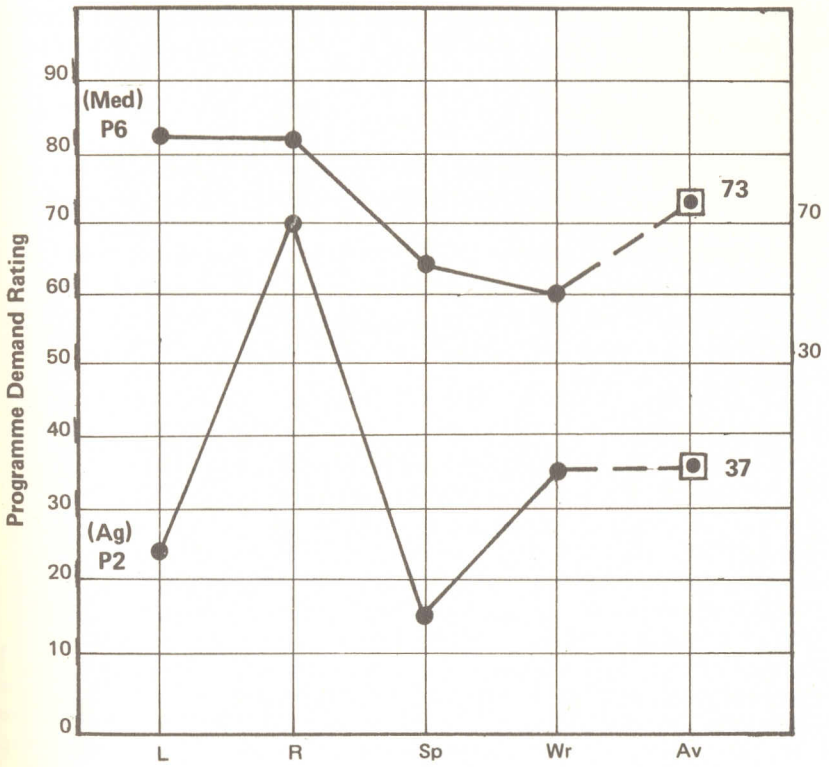
Table 1: Average Ratings % for Target Level Dimensions

Participant	Listening	Reading	Speaking	Writing	Average
P1 Business Studies	81	76	60	67	71
P2 Agric. Science	26	69	17	36	37
P3 Social Survival	74	60	50	14	50
P4 Engineering	81	79	52	57	67
P5 Lab. Technician	79	67	52	36	59
P6 Medicine	83	83	64	60	73
Overall averages	71	72	49	45	59

Even if we accept that the ratings in the table look more precise than they actually are, we can see very different types of profile for the various participants. The overall pattern of demand is for a high level for the receptive media (71 and 72) and a much lower level for productive media (49 and 45) indicating the fairly obvious fact that the participants play a responding rather than an initiatory role in the learning situation. The three EAP examples, P1, P4 and P6 have rather similar need profiles, with P6 (Medicine) having probably the most demanding one (average 73). Of the remaining three profiles, P2 (Agricultural Science) is the most remarkable, with a high demand only in reading and an overall average demand of only 37.

We will show, in Table 2 below, a graphic representation of the two extreme profiles P6 and P2 to illustrate perhaps the most significant conclusion to be obtained from the present report, namely that the pattern of demands of the various programmes can be very different both overall and for the individual verbal media. Admittedly we have, for illustrative purposes, chosen the two extreme cases but the same considerations, in less extreme form, will apply to the other profiles.

Table 2: Comparison of Medical (P6) and Agricultural (P2) profiles



The first point to note is that the profiles are not level, but subject to considerable rise and fall across the scale, indicating that the average demand rating should not be used unqualified as an estimate of the difficulty of a programme. In the above case, a student with a level profile of 50 (assuming that we have devised comparable profile calibrations for both

programme demands and student competence)would be above the average rating of 37 for P2, but would be below standard on his reading rating. A student with a level profile of 70 would be above the level of all demands for P2 but still fall short in listening and reading modes for P6. The important point we wish to make, and to which we will return later in the report, is **that in making placement decisions we must match the profile of programme demands with the profile of candidate performance.** This conclusion is extremely significant in that we can now achieve our object of improving our test system not only by improving the precision and relevance of the tests themselves (the centre of our negotiations so far) but also by clarifying and making more precise the communicative demands of the various programmes.

b Tolerance Conditions

We will not go into such detail in our analysis of the ratings for tolerance conditions because indications are in the same direction as those reached in the previous section.

The different programmes have their own respective patterns of tolerance level and the tolerance ratings are negatively correlated with the level of demand; in other words high demand on performance goes with low tolerance, and vice versa.

One conclusion from the tolerance conditions analysis is that the least tolerance is, broadly speaking, extended to language errors and the most to deficiencies in style, recourse to reference sources and to repetition. We can thus conclude that correctness of language usage — lexis, grammar, spelling, punctuation, etc — is by no means an unimportant component of communicative competence in study programmes, although, as we already observed, this correctness should be judged in a communicative context; the higher level skills of scanning, evaluation and logical deduction, for example, cannot be exercised in a linguistic vacuum. This is a consider-- ation that enthusiasts for the communicative approach have been in danger of forgetting.

Apart from the ratings of tolerance we have been considering, there is one important polarity which placement agencies have been very familiar with and which derives from the autonomy of British educational institutions and their departments. This is that it is for the Head of a Department to decide whether or not an applicant is to be accepted on a programme. At one extreme we may have a post-graduate course in Medicine which is already over-subscribed and whose Head is naturally concerned with retaining very high standards of competence if only because the students' decisions will

often be a matter of life and death. At the other extreme, we may have the Head of a Science Department in a College of Further Education whose students come almost wholly from overseas and whose staff would be courting redundancy if they rejected applicants because they had language problems.

It is clear that for the former type of department, our testing and tuition must be such as to guarantee that the students have reached the required level of communicative competence before they embark on their course of study. In the latter type, whilst it will still be necessary to establish programme demands and student competence levels, there will be much more scope for concurrent language tuition and, no doubt, for the provision of special bridging courses in which attention can be given both to the improvement of language and to subject skills.

These considerations reinforce our earlier conclusion about the need to match course demands and student competence levels. A clear, intelligible system for presenting the two kinds of information should therefore be available so that Heads of Departments will have to hand a convenient instrument for making placement decisions.

7 Events and Activities (Appendix A, Spec. 7)

Events are what the participants have to do by virtue of the training programme they have undertaken. A typical event would be 'Attending a lecture in the main subject area', and this event could be broken down into component activities such as:

'listening for overall comprehension'
'making notes on main points of lecture',
and 'asking questions for clarification'.

From the topics treated in the events are derived the significant lexical items and lexical sets to be used on academic programmes. It should be noted, however, that language realisations are not derived directly from these activities but via skills and socio-semantic units described later.

The events and activities recorded in Spec. 7 reinforce the information about settings already discussed. The main study focuses are lectures, seminars/ tutorials, reference study, report writing, laboratory work, and practical work in industry, on field projects and in hospitals. The extent to which Social Survival English should play a part in the assessment process has been the subject of some controversy. On the one hand, trainees in Britain will need some mastery of the kind of English used in social interactions; on the

other hand, as the language formulae are heavily culture-bound, it may be unreasonable to expect candidates to be familiar with them in the way that they could be expected to be with the type of discourse used in their own subject areas. We are on the point of completing a new profile, P7, based on 'English for International Use', which may provide a compromise in this area of Social English.

8 Attitudinal Tone Index (Appendix A, Spec. 8)

The communication units derived from the specified activities (and referred to again in our next section on micro-functions) are marked for attitudinal tone. It is the expression and recognition of attitudes which often pose to non-native speakers their greatest problem, and is usually the area of language training which is the most neglected. In our specification, no less than forty-three attitudinal tone continua are recorded. We list below thirteen of these tones which we judge to be most important partly in view of their frequency of occurrence:

Pleasant-unpleasant
Cautious-incautious (p)
Caring-indifferent
Formal-informal(p)
Grateful-ungrateful(p)
Honest-dishonest(p)
Disinterested-biased
Respectful-disrespectful
Approving-disapproving(p)
Inducive-dissuasive(p)
Certain-uncertain(p)
Intelligent-unintelligent
Assenting-dissenting(p)

The participants are expected to recognise manifestations of all these tones and to be able to produce those marked (p).

9 Language Skills (Appendix A, Spec. 9)

The activities listed in Spec. 7 may also be realised in terms of language skills contained in the fifty-four skill categories of our model and listed as a taxonomy in Appendix A. For practical purposes of test development, this area of specification is of the greatest importance. We have recorded for each skill any profile which refers at least once to that skill.

On the assumption that any skill recorded for 4, 5 or all of the profiles is likely, because of the heterogeneity of our participants, to be of a general, or non-disciplinary, nature and the skill category to be of broad significance, we mark such skills with an asterisk below. We also list other skills categories for which there are skills with 3 occurrences as well as a small number whose absence would give inconsistency to our list.

List of Essential Language Skill Categories

Skill Category	Abbreviated Title
4	Articulating sounds in connected speech.
7/8	Recognising and manipulating stress variations in connected speech.
9/10	Recognising and manipulating stress for information, emphasis and contrast.
11/12	Understanding and producing neutral intonation patterns.
13/14	Interpreting and expressing attitudinal meaning through intonation.
15	Interpreting attitudinal meaning through pitch, pause and tempo.
17/18*	Recognising and manipulating the script.
20/21*	Understanding and expressing explicit information.
24/25*	Understanding and expressing conceptional meaning.
26/27*	Understanding and expressing communicative value.
19	Deducing meaning of unfamiliar lexical items.
22*	Understanding information not explicitly stated.
28/29*	Understanding relations within the sentence.
30/31	Understanding and expressing relations through lexical cohesion devices.
32/33*	Understanding and expressing relations through grammatical cohesion devices.
35*	Recognising indicators in discourse.
37/38	Identifying and indicating main point of discourse.
39*	Distinguishing main idea from supporting details.
40/41*	Extracting salient points of text.
43*	Reduction of text.
44*	Basic techniques of text layout and presentation.
45	Skimming a text.
46	Scanning a text.
47/48*	Initiating and maintaining a discourse.
51/52*	Transcoding information (diagram/language)

If a test were devised using the skill categories marked with an asterisk, it would cover the main language skill needs of all types of participant. In framing the test items we would refer to the Target Level indices and the topic areas provided by the specifications. The skills covered in the categories between 4 and 15, which we might call the lower-level skills, tend to be related to profiles P3, P5 and P6, indeed 84% of occurrences in these categories occur in respect of those three profiles indicating the existence of an EOP (English for Occupational Purposes) group factor. Further analysis of the factor pattern suggested by the Language Skill analysis is of the highest importance and is to be found in Section 3 below.

10 Micro-Functions (Appendix A, Spec. 10)

The use of the term 'function' is currently a matter of extended debate, and for a detailed discussion of its use in the present document one must refer to J Munby's thesis. For present purposes, however, we will define the micro-function as representing an inter-level between events (with their component activities) and their linguistic realisation. When we have specified an event and its component activities, we are not yet in a position to generate language realisations. This process can be carried out via the selected language skills categorised in Spec. 9 with particular reference to skill categories 26 and 27 related to the communicative value (or function) of an utterance; or it may be done by selecting the appropriate micro-functions from Spec. 10 (affirmation, certainty, negation, etc) and marking them for attitudinal tone from the index given in Spec. 8.

We suggest that none of the micro-functions in the 7 categories given in Spec. 10 are to be ignored. It may be best in practice to base test items on a good coverage of the important skill taxonomy items suggested in Spec. 9 and to support them with relevant socio-semantic units derived from the list of Micro-functions marked with appropriate items from the index of Attitudinal Tones, the latter half of the process being particularly relevant to the less academic communicative activities.

This suggested procedure can be checked for its value during the test development phase.

Implications for Test Design

1 The various conclusions arising from the analysis of our sample specifications have now to be drawn together so that specific proposals for test design and development can be made. It will be prudent first to reiterate our reservations about the data:

a The six participant types we have selected do not purport to be a representative sample of the levels and disciplines of the total testee population.

b The field work so far done depends too much on the subjective judgements of the compilers and too little on close, extended observation of learning situations.

c The reliability of the target level ratings cannot be vouched for and they should only be used to support broad conclusions.

In spite of these reservations, however, we should not forget that the present approach to test design via the detailed specification of communicative needs is a breakthrough, and a considerable advance on the traditional approach to test design based either on purely linguistic categories (vocabulary, structure), on the convenience of particular test types (cloze, multiple-choice,) discrimination of phonemes or on hybrids of language categories and communicative tasks (reading comprehension, interviews) supported by norm-referenced statistics of probability. It is not that any of the above features are irrelevant, it is just that they do not operate in a coherent communicative framework.

2 Range of Communicative Demands

On studying the various profiles, one is struck by the very wide range of communicative demands the programmes make on the participants. This wide range — of skills, topics, channels, verbal media, interactions and functional categories — exists even in apparently the most simple programmes. We are bound to conclude that conventional tests are covering too narrow a range of communicative and language requirements; this fact may explain the disappointing results which validation studies of language testing so often produce.

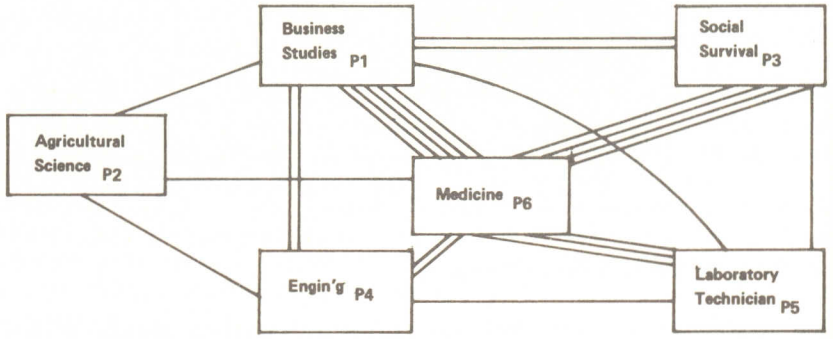
3 Common and specific factors

We have used the taxonomy of Language Skills to study the pattern of relationships existing between the various disciplines. Using the data of Appendix A, Spec. 9, we have recorded for each skill category all co-occurrences of all Ps; in pairs, in threes, in fours, in fives, and those skills recorded in all six P's or for only one P. The data give us indices of the amount of communicative overlap between the various disciplinary programmes which we assume to indicate similarities of demand between them. We illustrate our findings in Table 3 in the shape of a network, the number of lines indicating the strength of the relationship between any two programmes; to keep the diagram intelligible we have omitted small or negligible relationships.

The main network feature is a clearly-defined star-pattern with Medicine (P6) strongly related to Business Studies (P1) and to Social Survival (P3), and fairly strongly to Laboratory Technician (P5) and Engineering (P4).

The second main network feature is the isolated position of Agricultural Science (P2).

Table 3: Language Skill Network



The third network feature is the position of Business Studies (P1) as itself the centre of a subsidiary cluster related to all the other Ps and as a satellite of P6.

The conclusion we draw from these relationships is a perfectly clear one, **that Language Skill requirement patterns cut right across disciplinary boundaries;** indeed, in this study, we find the smallest communicative relationships between disciplines which seem to have the most in common, eg Engineering and Technician, both in the applied technology field.

We have not carried out such detailed work on other specification areas but a rapid check on overlap of attitudinal tones suggests a similar sort of conclusion about communicative features and disciplinary boundaries.

This finding has important implications for test design, but still leaves us with a major unsolved problem. Even if the Medical and Business Studies programmes we have considered are highly correlated communicatively, it still remains that the spoken and written discourse of the two disciplines are very different indeed; their linguistic and diagrammatic realisations have very different appearances. Can we then test different disciplines with identical test material, selected to test their common communicative requirements? Or will we, in doing so, use over-generalised language/diagram realisations which may favour candidates in one particular discipline or, worse still, be equally irrelevant to all the disciplines? We are not yet in a position to answer these questions, so we propose to continue in a pragmatic fashion by preparing tests in different disciplinary areas and by paying particular attention in test data analysis to assessing any benefits, in improved test effectiveness, which can be related to diversification on a disciplinary basis.

Pending a full statistical analysis of future test results, we put forward a tentative assessment of the factor pattern underlying our network diagram in Table 3:

Factor I: 'general' factor, accounting for a sizeable proportion (perhaps half) of the variance, representing the common communicative requirements and characteristics (intelligence, motivation, academic aptitude) of all participants.

Factor II: an 'Academic Study' factor reflecting the ability to use the communication/language skills necessary for handling academic discourse of a relatively neutral attitudinal nature.

Factor III: a 'Personal Relationships' factor representing non-study relationships with contacts in field or clinical work.

Factors IV + : Specific or small-group factors representing the special additional requirements of odd-man-out programmes.

4 Testing versus Matching;

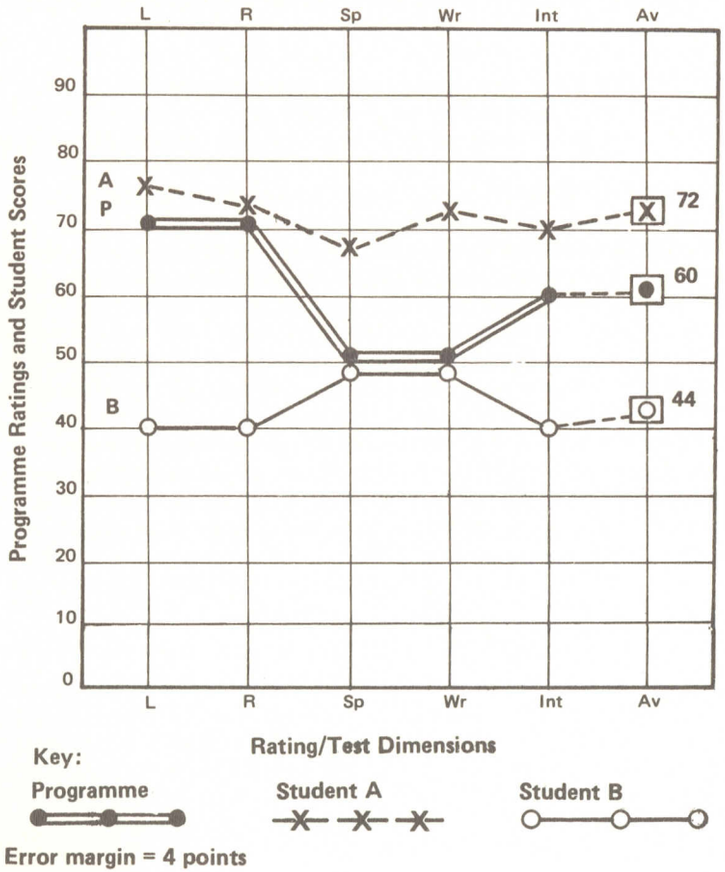
It will be remembered that earlier (in Section 6. a) we reached a conclusion of the greatest operational significance, that considerable improvement in placement efficiency could be achieved not only by improving the tests themselves but also by matching the competencies of the candidates with the communicative demands of the programmes, on a profile basis. This close integration cannot be achieved if the testing service is seen as an autonomous, separately - conducted operation in the manner of a periodically-set Proficiency examination. Nor will test efficiency be improved if tests are based mainly on formal language considerations divorced from programme communicative requirements. The closer the involvement of the receiving institutions and placement agencies in the assessment process, the more likely they will be to conduct an efficient placement service.

5 A framework for measurement

We have already established the value of comparing, or matching, candidate performance with programme demands. What we now need is a common scale upon which we can plot in a comparable fashion, the profiles which express significant dimensions of the two types of assessment. This framework should be intelligible to the non-specialist staff who have to make day-to-day decisions about the placement of thousands of applicants. We give in Table 4 an illustration of such a framework.

Let us suppose we are rating programme demands, and testing student performance, on six dimensions – listening, reading, speaking, writing, integrated skills and the average of all these scores. We show, in the framework, profiles for the programme (P) and for two students (A) and (B). To allow for rating and scoring unreliability we assume a margin of error of 4 points which can be visualised as a grey area 4 points above or below P. Our placement officer is asked to make the classic decisions for Students A and B – whether they are acceptable as they stand or, alternatively, what type of language tuition they may require before acceptance. This task, which in normal cases he should find a good deal easier than filling in his Income Tax return, is done by reference to the respective profiles.

Table 4: Matching programme demands and student proficiency



Student A, even allowing for any errors of measurement, is significantly above the profile, P, in all dimensions and he can be recommended for acceptance (in respect of his communicative competence) without qualification. The chances of his having language problems on his course of study are very small indeed.

Student B, however, is significantly below the Programme Rating in 3 areas, listening, reading and integrated skills; and just below, although not significantly so, in speaking and writing. He will therefore require language tuition before his course begins. A decision then has to be made about the nature and duration of his tuition. As his main deficiencies are in the receptive media and in integrated skills, some emphasis on those areas will be recommended. The extent of his deficiency can be counted in terms of bands, ie 3 bands each for L and R and 2 bands for Int, or 8 bands in all. Let us assume an average tuitional requirement of 25 hours per band, then we will recommend 200 hours of language tuition. The bases for such estimates can be made more precise in the light of experience.

Such a matching system would not only improve our placement process but could also effect considerable economies in pre-course tuition – an extremely expensive activity – because we would now have much more precise guidance about the nature and duration of the tuition than we could have obtained by comparing a student's average score with a vague estimate of course requirements, a hit-or-miss procedure which runs the risk of providing over-tuition for certain students and under-tuition for others.

6 Syllabus Implications

In preparing the test content specifications for our participants, we have at the same time been specifying essential parts of the syllabus content specification for teaching purposes because we cannot specify test requirements in a curricular vacuum. This double result is, however, a fortunate one for our Testing Service as we now have ready to hand a tuitional blueprint to supplement the placement system. The detailed work on specification, then, has not been nugatory but has added a significant dimension to the operational resources of the testing/tuition service overall.

7 Test Format

In our preparatory work, we have had no difficulty in devising test types to measure the varied communicative features revealed in the specifications, indeed the range of activities brought up has been a valuable stimulus to test development. It is not the devising of test formats which has been the

problem, but the making of an operational framework in which to deploy them. We will in our proposals give an outline of a test format which we consider relevant, but we emphasise that the central point of this report is the specification of communicative needs and demands and that discussion of test formats should not by-pass the crucial area of specification.

Operational Requirements

In this section, we will focus our attention on the operational requirements of overseas representations and training/scholarships departments but we must remember that they are working basically on behalf of the British institutions of all kinds, Universities, Colleges and Research Institutes, who actually receive the students. Here are the main operational requirements:

1 Tests must be readily available at all times of the year. Several representatives have said that to arrange fixed dates for test applications (say three or four times a year) would introduce intolerable delays in the manpower training cycle.

2 Results of the tests must be available within days or even hours of their administration to candidates. One senior representative for example has said that if he has to wait for more than a week for results he will not be able to use the Test Service.

3 Clear guidance must be available to assist staff in interpreting test results for placement and/or tuition purposes.

4 In certain countries there are large numbers of candidates (estimates vary between 50% and 80%) who have no reasonable chance of achieving any kind of satisfactory pass performance. A rapid screening device for identifying such candidates is urgently needed.

5 Most representatives are keen to see an improvement in the efficiency of the testing service but wish to achieve this with the minimum of increase to their administrative load.

6 The cost of testing is a sensitive issue. Considerable opposition to a proposed fee of £10 plus local costs has been demonstrated. Different regions of the world vary considerably in their reactions to price increases.

7 Security of tests is important, particularly as versions of the present test are known to have been compromised. This does not mean that every test has to be a completely new one, but that alternative versions should be available, and old versions should be replaced, at a rather faster rate than they are at present.

8 In small representations or where professional ELT resources are not available, the application, marking and interpretation of tests may require external assistance on a regional or central basis.

9 Areas with large Direct English Teaching operations have considerable resources available for testing.

10 There will always be unusual or specially urgent demands for testing not catered for within any broadly applicable test framework. Exceptions must be allowed for.

Overall, the variety of requirements of 70 or 80 representations and up to 120 countries demands a flexible (even untidy) approach to language assessment if a large and complex manpower programme is to maintain its operational momentum.

Recommendations for a Language Testing Service

1 We now put forward for consideration a number of recommendations concerning the design and development of the testing service. In framing the recommendations, we have aimed to give balanced consideration to the findings of our specification analyses, to the practical constraints upon those who have to operate the service and to commonsense considerations about what is feasible in present circumstances.

Recommendation 1 — Test Phases

That a two-level testing pattern be adopted with the following phases:

Phase A A broad-span, easily-administered screening test in listening and reading skills, covering in a non-disciplinary manner the receptive Language Skill categories 20, 24 and 26, (information handling, conceptual meaning and communicative value) and Skills 30, 32, 37, 39 and 40.

Phase B A modular test pattern covering the communication skills appropriate to about 6 major disciplinary areas with sizeable numbers of candidates. These disciplinary tests should be supplemented by an Academic Communication Skills test designed for applicants who are not certain about their course of study, who are not adequately catered for in the existing disciplinary modules or are undertaking inter-disciplinary studies.

Recommendation 2 — Marking

That Phase A be marked in an objective manner and capable of being applied, marked and interpreted locally by non-specialist staff. That Phase B should be marked in as objective a manner as possible but may contain features requiring trained assistance for application and assessment.

Recommendation 3 — Interpretation

That the principle of matching students to course demands be accepted and a profile framework be devised to facilitate interpretation of test results.

Recommendation 4 — Development

That a test development team be specially trained in the use of specification techniques and the devising of tests derived from them and to prepare two parallel versions of a Phase A test and one version of a test for each of the Phase B areas.

2 A Sample Testing Pattern

Before a firm test pattern can be devised, decisions on the recommendations above will have to be made and the number and content of modular areas will have to be ascertained. We put forward a 'shadow' test pattern, subject to modification, as follows:

Phase A. Reading Test (approx 50 minutes)

1 **Test of conceptual meaning skills** in Skill Category 24 and relations within sentence, Skill 28. (50 items, m/choice, discrete)

2 **Test of communicative value,** Skill 26, and Lexical and Grammatical cohesion devices, Skills 30 and 32. (50 items, modified m/choice cloze type)

3 **Understanding of information,** Skill 20, with component of Attitudinal Tone input (Spec. 8) and Communicative Value, Skill 26 (and Spec 10) (30 m/choice items based on texts)

Listening Test (approx 30 minutes)

1 **Recognition of shapes,** diagrams and pictures from taped descriptions, testing conceptual meaning, Skill 24. (30 multiple-choice items)

2 **Recognition of short sentences** from taped descriptions testing conceptual meaning, Skill 24 and function, communicative value, Skill 26. (30 multiple-choice items)

3 **Comprehension of a lecturette** of about 3 minutes, test of recognition of facts, Skill 20 and identifying main point as in Skills 37, 39 and 40 (20 multiple-choice items)

Phase B Modular Tests (approx 100 minutes)

[Possible areas:— Agriculture, Medicine, Science, Technology, Administration, Education; plus General Academic test based on English for academic and international use]

1 **Reading Study Skills test;** of Skills numbered between 22 and 52, especially the starred skills, based on information booklet on topic area. (40 multiple-choice items with same accepted alternatives for all modules to facilitate marking)

2 **Writing Skills test;** problem-solving, descriptive and reference skill writing based on information booklet. (Subjective rating according to scale and with photo'ed samples of examples at different levels)

3 **Structured Interview;** in cases where there is high demand and low tolerance for speech skills. (Subjective rating on detailed scale and based on information booklet. Cassette samples of different levels available)

Time Limits. As tolerance for time/fluency is fairly high, it is recommended that time limits should be fairly generous and allow the normal student to complete most of the items. Overseas, a good deal of testing will be confined to Phase A (Reading Test) and perhaps A (Listening Test) and a comparatively small number may do all parts. In UK, the interest will probably shift to Phase B especially for University entrance purposes.

APPENDIX A

Specification of Communicative Needs

	The Participant	P1. Business	P2. Agriculture
Spec. 0	**Age** **Nationality** **Language** **English Std**	**20's** **Nigerian** **Hausa** **Intermediate**	**20's** **Venezuelan** **Spanish** **Elementary**
Spec. 1	Purpose of study		
	Course	HND Business Studies Polytechnic	Post Graduate Agricultural Studies University (English for Reference)
	Study Areas	Business Studies: Economics, Law, Business Accounts, Statistics Marketing, Purchasing	Agriculture: Cattle breeding, Animal husbandry, Physiology
	General Area	Social Sciences	Biological Sciences
Spec. 2	Setting for English		
	Physical	Lecture room Tutorial room Library Factories Business offices	Lecture rooms Laboratories Library Bookshop
	Temporal	Full-time in term, plus vacations, Av: 10 hours per day	In English classes In term-time 10 hours per week Less in vacation
Spec. 3	Interactions	*Learner-instructor *Outsider-insider Non-professional-professional *Non-native-native *Insider-insider *Adult-adult	Learner-instructor Non-native-native Insider-insider Adult-adult *Professional-professional

Note: Interactions recorded three or more times are marked with an asterisk

P3. Social	P4. Engineering	P5. Technician	P6. Medicine
20's **Turkish** **Turkish** **Upper Intermed.**	**20's** **Sudanese** **Arabic** **Intermediate**	**30** **Ethiopian** **Amharic** **Intermediate**	**26** **Saudi** **Arabic** **Upper Intermed.**
Academic Studies at University - (Social purpose)	BSc in Civil Engineering University	Experience as Medical Lab. Technician Hospital/College	Post Graduate studies in Medicine for FRCS. Teaching Hospital
not specified; social survival for specific study area	Engineering: all branches (gen) Maths, Electrical Science, Thermo-fluids, Mechanics, Surveying, Project Finance & appraisal	Medic Lab Techniques: Physical Sciences Biological Sciences Para-medical Workshop practice	Medical Studies: Anatomy, Surgery, General Medicine, Consultancy & Casualty work
	Engineering Science	Mixed Technology	Medicine
On campus, Canteens, cafes offices, Houses, Places of Entertainment Sports places	Lecture halls Workshops Laboratories Library Tutorial rooms Field sites	College Hospital Teaching areas Library Workshop	Hospital surgery wards Operating theatre Lecture rooms Seminar rooms Library Common Room
Daily use 10-12 hours per day throughout year	Daily, all day Up to 10 hours p day	Weekdays 6 hours, less at weekends, During training course	5 days per week 9 hours + per day Regularly whilst in UK
Learner-instructor Outsider-insider Beneficiary-benefactor Non-native-native Insider-insider Adult-adult Professional-professional *Junior-senior(+vv) Advisee-adviser *Man/woman-man/woman *Equal-equal Friend-friend Guest-host	Learner-instructor Outsider-insider Non-native-native Adult-adult Professional-professional Junior senior Man/woman-man/woman Student-student	Learner-instructor Non-native-native Insider-insider Adult-adult Professional-professional Equal-equal Man/woman-man/woman Customer-server Member of pub-official Guest-host	Learner-instructor (+vv) Therapist-patient Adviser-advisee (+vv) Consultant-client Leader-follower Adult-adult Professional-professional Professional-non-professional Senior-junior (+vv) Equal-equal

	Instrumentality	P1. Business	P2. Agriculture
Spec. 4	Medium	Listening Speaking Reading Writing	as P1
	Mode	Monologue Dialogue (spoken and written to be heard or read; sometimes to be spoken as if not written)	as P1
	Channel	Face-to-face Print Tape Film	Face-to-face Print
Spec. 5	Dialect	All sections: Understand British Standard English dialect. Produce acceptable regional version of Standard English accent.	

Spec. 6 — Target Level (in the 4 media for each section)

Dimensions:		L	Sp	R	Wr	L	Sp	R	Wr
(max=7)	Size	6	3	7	3	2	1	7	3
	Complexity	7	4	6	5	2	1	6	3
	Range	5	4	5	5	2	1	4	2
	Delicacy	5	5	6	6	1	1	5	3
	Speed	6	4	5	6	3	2	5	3
	Flexibility	5	5	3	3	1	1	2	1
Tolerance Conditions		L	Sp	R	Wr	L	Sp	R	Wr
(max=5)	Error	3	4	3	3	4	5	1	2
	Style	4	4	5	4	5	5	4	4
	Reference	3	4	2	2	5	5	3	3
	Repetition	3	4	2	3	5	5	5	3
	Hesitation	3	4	4	3	4	5	3	3

P3. Social	P4. Engineering	P5. Technician	P6. Medicine
as P1	as P1	as P1	as P1
as P1	as P1	as P1	as P1
Face-to-face Telephone Print Public address Radio TV Disc Tape recorder Film	Face-to-face Print Film Pictorial Mathematical	Face-to-face Telephone Radio Print Tape recorder	Face-to-face Telephone Print
Dialect	All sections: Understand British Standard English dialect. Produce acceptable regional version of Standard English accent.		

L	Sp	R	Wr	L	Sp	R	Wr	L	Sp	R	Wr	L	Sp	R	Wr
4	3	4	1	6	3	7	3	6	4	5	3	6	5	6	4
4	3	4	1	6	5	6	5	6	3	5	3	6	4	6	4
7	3	5	1	5	4	6	4	6	5	6	3	6	4	6	4
4	4	4	1	6	4	6	5	6	5	6	3	6	5	6	5
6	4	4	1	6	3	4	4	6	3	5	2	5	4	5	4
6	4	4	1	5	3	4	3	3	2	1	1	6	5	6	4

L	Sp	R	Wr	L	Sp	R	Wr	L	Sp	R	Wr	L	Sp	R	Wr
3	4	3	5	1	3	3	2	4	4	3	4	3	4	3	4
4	4	4	5	2	3	3	3	5	5	5	5	3	3	3	3
2	2	5	3	5	4	5	5	5	5	5	5	3	3	4	4
2	3	5	4	3	4	3	5	5	5	5	5	4	3	4	3
2	3	4	4	4	5	4	4	3	4	3	3	3	3	4	4

Spec. 7 Events/Activities

P1. Business	P2. Agriculture	P3. Social
1 Lectures	**1 Reference Study**	**1 Official discussions**
Listen for overall	Intensive for all infm	Reading forms
Comprehension	Specific assignments	Complete documents
Make notes	Evaluative reading	Discuss with officials
Ask for clarification	Main infm rdg	
2 Seminars/Tutorials	**2 Current Literature**	**2 Social in Britain**
Discuss given topics	Routine check	Personal information
Listen for comprehension	Keep abreast	Invitations
Make notes	For main information	Mealtime conversation
Ask for clarification		Complaints
		Polite conversation
3 Reference Study	**3 English lessons**	**3 Places of Interest**
Intensive reading	Test study	Reading text for infm
Reading for main infm	Teacher exposition	Entrance/tickets
Assignment rdg	Group work	Guidebooks
Assessment rdg		Listen to commentary
		Ask for information
4 Writing Reports	**4 Other**	**4 Shopping**
Sort out information	(Note: English is not	Attract attention
Factual writing	much used in this	Discuss goods
Evaluative writing	Spanish context,	Give choice
	outside the study area)	Arr payment
		Complaints
		Sale documents
5 Keeping up-to-date		**5 Health**
Routine checking		Appt-person/phone
Reading for intensive		Discuss symptoms
Reading for infm search		Complete forms
		Medical directions
6 Indust/Comm Visits		**6 Restaurants/cafes**
Discuss topics		Attract attention
Discuss after visit		Place order(s)
Listening for infm		Deal with bill
Take notes		Complaints
Ask for clarification		
		7 Travel
		Timetables, schedules
		State destination
		Pay fares
		Maps, explanations
		Road signs/symbols

P4. Engineering	P5. Technician	P6. Medicine
1 Lectures	**1 Lectures**	**1 Diagnosis**
Work sheets	Listen to explanations	Questioning,
Notes/diagrams	Listen to instructions	rephrasing
Displays/models	Coord with colleagues	Compr garbled infm
Seek description	Take notes	Notes for records
Understand lectures	Record test results	Ask for clarification
	Questions & comments	
	Read instr for test	
	Read instr re specimen	
2 Tutorials	**2 Reference Study**	**2 Instruct Staff**
Sheets, notes, displays		Groups or individuals
Seek clarification		Question to check
Evaluate schemes	Rdg for main information	Write notes (med codes)
Problem solving	Intensive reading	Requests re instructions
Mathematical probs	Take notes	
Assignment apprec		
3 Experiments	**3 Give Recommendations**	**3 Write**
Prove hypothesis	Prepare notes	Personal letters
Solve problems	Speak to notes	Case descriptions
Write up experiments	Talk about diagrams	Note form
Report on projects	Answer queries	Full reports
Explore principles		
4 Reference Study	**4 Self-Access**	**4 Students Seminars (conduct)**
Intensive experiments	Tape-slide uses	Explain themes
Intensive re applics	Reading for main infm	Question, correct
Refer to tables, data	Intensive reading	Present peer seminars
Subject periodicals		Notes, handouts
		Blackboard, OHP
5 Field Work		**5 Attend Less/Seminars**
General site visit		Comprehend overall
Periodical work visits		Selective retention
Survey instruments		Notes for reconstruct
Experimental surveys		Ask for clarification
Discuss problems		Present own topic
Write up experiments		Informal discussions
		6 Reference Study
		Intensive reading for all
		Reading for main point
		Reading for spec. assignment
		Assess position
		Routine check
		Exophoric reading

Spec. 8 Attitudinal Tone Index

(This list gives the superordinate terms and the 'P' profiles which indicate their significance eg 4, 5, 6, indicates that P4, P5 and P6 record this tone)

Superordinate polarity	'P' occurrences
Happy - unhappy	6
Contented - discontented	5 5
*Pleasant(ing) - unpleasant(ing)	1 1 4 4 5 5 5 6
Cheerful - dejected	6 6
Frivolous - serious	5 5 5 6
Rejoicing - lamenting	6
Entertaining - tedious	4 5 5^h
Exciting - unexciting	5
Humorous - humourless	5 5 6 6
Sensitive - insensitive	4 4 6 6
Hoping - hopeless	4 5 6 6
Courageous - fearing	6
*Cautious - incautious	1 1 2 4 4 4 4 4 4 5 6 6
*Caring - indifferent	1 1 2 4 4 4 4 5 6
Wondering - unastonished	6 6
Modest - proud	5 5 5
*Formal - informal	1 1 1 2 4 4 4 4 5 5 5^h 6 6 6
Friendly - unfriendly	5 6 6
Courteous - discourteous	1 1 4 5^h
Sociable - unsociable	6
Unresentful - resentful	6
Pleased - displeased	6 6
Patient - impatient	1 6
*Grateful - ungrateful	1 4 4 5 6
*Honest - dishonest	1 1 2 4 6 6
*Disinterested - biased	1 1 1 2 5 6
*Respectful - disrespectful	1 4 4 4 4 5^h 6 6
Admiring - contemptuous	5
Praising - detracting	1 5 6
*Approving - disapproving	1 1 1 2 4 5 6 6
Regretting - unregretting	5^h 6
Temperate - intemperate	6 6
Excitable - unexcitable	6 6
Willing - unwilling	1 1 4 4 6 6 6
Resolute - irresolute	4 6 6 6
*Inducive - dissuasive	1 1 1 2 5 6 6
Active - inactive	1 1 4 6 6 6
Concordant - discordant	1 1 1 2 6
Authoritative - unauthoritative	1 1 1 2 6 6
Compelling - uncompelling	1 1 1
*Certain - uncertain	1 1 1 2 4 5 5 6 6 6
*Intelligent - unintelligent	1 1 1^h 2 5 5^h 6 6
*Assenting - dissenting	1 1 1 2 4 5 5 6

Notes (1) P3. (Social English) has been omitted from this list
(2) The symbol ^{h} denotes a hyponym
(3) Tones used by 4 or more of the 5 profiles are indicated with an asterisk.

Inventory of Language Skills

We now record which Profiles require the Language Skills of the Munby list, to which refer for expansion of the abbreviated titles below. Skills required by 4 or more profiles (out of 6) are marked with an asterisk.

Skill Category	Abbreviated title		
1	Discriminating sounds in isolated words.		nil
2	Articulating sounds in isolated words.		nil
3	Discriminating sounds in connected speech.		
	3.1	Strong/weak forms	4
4	Articulating sounds in connected speech.		
	4.1	Strong/Weak forms	4 5 6
	4.2	Neutralisation	5
	4.3	Reduction vowels	5
	4.4	Sound modification	5
	4.5	Word boundaries	5 6
	4.6	Allophonic variation	5 6
5	Discriminating stress within words.		
	5.1	Accentual patterns	5
	5.2	Meaningful patterns	5
	5.3	Compounds	5
6	Articulating stress within words.		
	6.1	Accentual patterns	5 6
	6.2	Meaningful patterns	5 6
	6.3	Compounds	5 6
7	Recognising stress variations in connected speech.		
	7.2	Meaningful prominence	3 4 6

8	Manifesting stress variations in connected speech.		
	8.1	Rhythmic considerations	6
	8.2	Meaningful prominence	3 4
9	Recognising stress in connected speech.		
	9.1	Information units	1 6
	9.2	For emphasis	1 3 6
	9.3	For contrast	1 3 6
10	Manipulating stress in connected speech.		
	10.1	Information units	5
	10.2	Emphasis	3 5 6
	10.3	Contrast	3 5 6
11	Understanding intonation patterns (neutral)		
	11.1-10	Fall-rise-multi tones	3
12	Producing intonation patterns (neutral)		
	12.1	Falling moodless	3 5
	12.2	Falling interrogative	3 5 6
	12.3	Falling imperative	5 6
	12.4	Rising interrogative	3 5 6
	12.5	Rising non-final	3 5
	12.6-8	Rise/fall	5
	12.9	Question tags	3 5 6
13	Intonation, interpreting attitudinal meaning.		
	13.1	Rising moodless	3 4
	13.2-7	Various tones	3
14	Intonation, expressing attitudinal meaning.		
	14.1	Rising moodless	3
	14.2	Rising interrogative	3 4 6
	14.3	Front shift	3 6
	14.4	Rising imperative	6
	14.5	Falling interrogative	3 6
	14.6	Front shift	3 6
	14.7	Others	3

15	Interpreting attitudinal meaning.		
	15.1	Pitch height	1 3
	15.2	Pitch range	1 3 4
	15.3	Pause	1 3
	15.4	Tempo	1 3
16	Expressing attitudinal meaning.		
	16.1-4	as for last drill	4 6
17	Reorganising the script.		
	17.1	Graphemes	3 5 6
	*17.2	Spelling	3 4 5 6
	17.3	Punctuation	3 5 6
18	Manipulating the script.		
	18.1	Graphemes	3 5 6
	*18.2	Spelling	3 4 5 6
	18.3	Punctuation	3 6
19	Deducing meaning of unfamiliar lexical items.		
	19.1.1	Stress, roots	1 2 4
	19.1.2	Affixation	1 2
	19.1.3	Derivation	1 4
	19.1.4	Compounding	1 4
	19.2	Contextual clues	1 2 3
*20	Understanding explicitly stated information.		
			1 2 3 4 6
*21	Expressing information explicitly.		
			1 3 4 5 6
22	Understanding information not explicit.		
	*22.1	Inferences	1 2 3 6
	22.2	Figurative language	3 6

23	Expressing information implicitly.	
23.1	Inference	6
23.2	Figurative lang	6
24	Understanding conceptual meaning.	
*24.1	Quantity	1 2 3 4 5 6
*24.2	Definiteness	1 2 4 6
*24.3	Comparison	1 2 3 4 6
*24.4	Time	1 2 4 5 6
*24.5	Location	1 2 4 6
*24.6	Means	1 2 4 5 6
*24.7	Cause, etc	1 2 4 6
25	Expressing conceptual meaning.	
*25.1	Quantity	1 4 5 6
*25.2	Definiteness	1 4 5 6
*25.3	Comparison	1 4 5 6
*25.4	Time	1 3 4 5 6
*25.5	Location	1 3 4 5 6
*25.6	Means	1 4 5 6
*25.7	Cause, etc	1 3 4 5 6
26	Understanding communicative value (re context)	
*26.1	With indicators	1 2 3 6
*26.2	Without indicators	1 2 3 6
27	Expressing communicative value	
*27.1	With indicators	1 3 5 6
27.2	Without indicators	1 5 6
28	Understanding relations within sentence	
28.1	Structure elements	3 5
*28.2.1	Premodification	1 2 3 5
*28.2.2	Postmodification	1 2 3 5
*28.2.3	Disjuncts	1 2 3 5
28.3	Negation	3 5 6
28.4	Modal auxiliaries	2 3 5
28.5	Connectors	2 3 5
28.6-7	Embedding + theme	2 3 5

29	Expressing relations within sentence.	
29.1	Structure elements	3 5 6
*29.2.1	Premodifications	1 3 5 6
*29.2.2	Postmodifications	1 3 5 6
*29.2.3	Disjuncts	1 3 5 6
29.3	Negation	3 5 6
29.4	Modal auxiliaries	3 5
29.5	Connectors	5 6
29.6	Complex embedding	1 6
29.7	Focus + theme	6
30	Understanding lexical cohesion devices.	
30.1	Repetition	3 6
30.2	Synonomy	2 3 6
30.3	Hyponomy	2 6
30.4	Antithesis	2 6
30.5	Apposition	3 6
30.6	Set/collocation	1 6
30.7	General Words	2 3 6
31	Using lexical cohesion devices.	
31.1	Repetition	3 6
31.2	Synonomy	1 6
31.3	Hyponomy	1 6
31.4	Antithesis	6
31.5	Apposition	6
31.6	Set/collocation	1 3 6
31.7	General words	2 3 6
32	Understanding grammatical cohesion devices.	
*32.1	Reference (c+a)	1 2 3 4
32.2	Comparison	2
32.3	Substitution	1 2
32.4	Ellipsis	1 2 3
32.5	Time/place relaters	2 3
32.6	Logical connectors	1 2 3

33	Using grammatical cohesion devices.		
	33.1	Reference	1 3 6
	33.2	Comparison	6
	33.3	Substitution	1 6
	33.4	Ellipsis	1 6
	33.5	Time/place relaters	1 3 6
	33.6	Logical connectors	1 3 4
34	Interpreting text by going outside		
	34.1	Exophoric reference	1 3
	34.2	'Between lines'	1 3
	34.3	Own experience	1 2
35	Recognising indicators		
	*35.1	Introduce idea	2 3 5 6
	35.2	Develop idea	2 3 6
	35.3	Transition	1 3 6
	35.4	Concluding	3 6
	35.5	Emphasis	2 5 6
	35.6	Clarification	3 6
	*35.7	Anticipation	1 2 3 6
36	Using indicators.		
	36.1	Introduce idea	3
	36.2	Develop idea	1
	36.3	Transition	1
	36.4	Concluding	1
	36.5	Emphasis	3
	36.6	Clarification	6
	36.7	Anticipation	1 3
37	Identifying main/important point.		
	37.1	Vocal underlining	1 3
	37.2	End-focus	-
	37.3	Verbal clues	1 3
	37.4	Topic sentence	1 2 6

38	Indicating main/important point.		
	38.1	Vocal underlining	3
	38.2	End-focus	-
	38.3	Verbal clues	1 3 6
	38.4	Topic sentence	6
39	Distinguishing main idea by differentiation.		
	39.1	Primary/secondary	2 4 5
	*39.2	Whole/parts	1 2 4 5
	39.3	Process/stages	2 4 5
	39.4	Category/exponent	2 5
	39.5	Statement/example	2 5
	39.6	Fact/opinion	1 2 5
	39.7	Proposition/argument	1 2 5
40	Extracting salient points to summarise.		
	40.1	Whole text	1 2 5
	40.2	Idea	1 2 5
	40.3	Underlying point	1 5
41	Extracting relevant points re.		
	*41.1	Coordination	1 2 5 6
	41.2	Rearrangement	1 6
	*41.3	Tabulation	1 2 4 6
42	Expanding salient points into.		
	42.1	Whole text summary	1
	42.2	Topic summary	1
43	Reducing text through rejection of.		
	43.1	Systemic items	6
	43.2	Repetition etc.	6
	43.4	Example compressions	6
	43.5	Abbreviations	1 2 6
	*43.6	Symbols	1 2 4 6

44	Basic reference skills.		
	*44.1	Layout	1 2 3 4 5 6
	*44.2	Tables, indices	2 3 4 6
	44.3	Cross-reference	4 6
	44.4	Catalogues	1 6
	44.5	Phonetic transcriptions	6
45	Skimming to obtain.		
	45.1	Gist	1 2 6
	45.2	Impression	1 6
46	Scanning to locate.		
	46.1	Simple search (single)	3 6
	46.2	Complex (single)	2 6
	46.3	Simple (more than 1)	6
	46.4	Complex (more than 1)	1 2 6
	46.5	Whole topic	1 2 6
47	Initiating a discourse.		
	*47.1	Initiate	1 3 5 6
	47.2	Introduce new	6
	47.3	Introduce topic	6
48	Maintaining a discourse.		
	*48.1	Respond	1 3 5 6
	48.2	Continue	1 5
	48.3	Adopt	1 3 5
	48.4	Interrupt	1 3
	48.5	Mark time	1
49	Terminating a discourse.		
	49.1	Boundaries	-
	49.2	Excuse	1 3
	49.3	Conclude	3

50	Planning and organising discourse (rhetorically)		
	50.1	Definition	1 4
	*50.2	Classification	1 4 5 6
	*50.3	Properties	1 4 5 6
	*50.4	Process	1 4 5 6
	*50.5	Change of state	1 4 5 6
51	Transcoding information from diagrams.		
	*51.1	Conversion into sp/wr.	1 3 4 5 6
	*51.2	Comparison in sp/wr.	1 2 5 6
52	Transcoding information from sp/wr.		
	*52.1	Completing a diagram	1 4 5 6
	*52.2	Constructing diagrams	1 4 5 6
53	Recording information.		
	Nil		
54	Relaying information.		
	54.1	Directly	3 5
	54.2	Indirectly	3 4

Spec. 10 List of Micro-Functions

Include all micro-fuctions from each of the Scales 1-6 for educational/training purposes, and micro-functions from Scale 7 for social survival purposes. Functions to amplify content of Language Skill Number 26.

1 Scale of Certainty

Affirmation, certainty, probability, possibility, nil certainty and negation. Conviction, conjecture, doubt and disbelief.

2 Scale of Commitment

Intention and obligation.

3 Scale of Judgement

Valuation, verdiction, approval and disapproval.

4 Scale of Suasion

Inducement, compulsion, prediction and tolerance.

5 Argument

Information, agreement, disagreement and concession.

6 Rational Enquiry

Proposition, substantiation, supposition, implication, interpretation and classification.

7 Formulaic Communication

Greeting, farewell, acknowledgement, thanks, apology, good wishes, condolence, attention signals.

Appendix B

TWENTY IMPORTANT STUDENT CATEGORIES

Rank order	Programme	% of Participants	% Cumulative
1	Agriculture (incl. Fisheries, Timber, Vets)	17	
2	Engineering (excl. Agricultural Engineering)	13	
3	Medical (including Dental & Paramedics)	10	40%
4	Economics and Development	8	
5	Administration (Public)	7	
6	Education (+ Education Administration)	5	60%
7	English Teaching	5	
8	Mining & Geology	4	
9	Accountancy, Banking and Insurance	4	
10	Sciences	4	
11	Physical Planning	4	
12	Sociology	3	81%
13	Business Admin, Management & Marketing	3	
14	Media	3	
15	Industrials	2	
16	Statistics, Demography	2	
17	Transport	2	
18	Aviation	2	
19	Laws	1	
20	Marine Engineering, Ports, Harbours	1	100%

Appendix C

Acknowledgements to staff assisting in preparation of specifications

Thanks are given to the following staff members who prepared participant specifications:

P.1.	Business Studies	Roger Hawkey
P.2.	Agricultural Science	John Munby
P. 3.	Social Survival	Shelagh Rixon
P. 4.	Civil Engineering	Melvin Hughes
P. 5.	Laboratory Technician	David Herbert
P. 6.	Medicine	Elizabeth Smyth

The major contribution to the operation has been John Munby's thesis, 'Specifying communicative competence; a sociolinguistic model for syllabus design,' shortly to be published by CUP[1]

Controller and Deputy Controller, English Language Division have also given advice on the requirements of the English Language Testing Service.

Directors ETIC and ELTI are thanked for allowing us to use staff for the specifications.

[1] Munby, John. *Communicative syllabus design.* CUP, 1978.

Appendix D

A statement of abilities required of first year entrants (Engineering Science) into Northern Universities (Joint Matriculation Board)

1 Knowledge and understanding of:

Terms, conventions and units commonly used in engineering science

Particular principles (or laws) and generalisations of engineering science, and their effects and interrelationships

Specialist apparatus and techniques used for the demonstration of the principles referred to above, and the limitations of such apparatus and techniques

The use of different types of apparatus and techniques in the solution of engineering problems

2 Abilities

Understand and interpret scientific and other information presented verbally, mathematically, graphically and by drawing

Appreciate the amount of information required to solve a particular problem

Understand how the main facts, generalisations and theories of engineering science can provide explanations of familiar phenomena

Recognise the scope, specification and requirements of a problem

Understand the operation and use of scientific apparatus and equipment

Recognise the analogue of a problem in related fields of engineering science and practice

3 Ability: Communication

Explain principles, phenomena, problems and applications adequately in simple English

Formulate relationships in verbal, mathematical, graphical or diagrammatic terms

Translate information from one form to another

Present the results of practical work in the form of reports which are complete, readily understandable and objective

4 Ability: Analysis

Break down a problem into its separate parts

Recognise unstated assumptions

Acquire, select and apply known information, laws and principles to routine problems and to unfamiliar problems, or those presented in a novel manner

5 Ability: Synthesis and Design

Design the manner in which an optimum solution may be obtained and to propose, where necessary, alternative solutions

Make a formal specification of a design or scheme

Make a plan for the execution or manufacture of the design or scheme

Use observations to make generalisations or formulate hypotheses

Suggest new questions and predictions which arise from these hypotheses

Suggest methods of testing these questions and predictions

6 Ability: Evaluation and Judgement

Check that hypotheses are consistent with given information, to recognise the significance of unstated assumptions, and to discriminate between hypotheses

Assess the validity and accuracy of data, observations, statements and conclusions

Assess the design of apparatus or equipment in terms of the results obtained and the effect upon the environment and suggest means of improvement

Judge the relative importance of all the factors that comprise an engineering situation

Appreciate the significance of social, economic, or design considerations in an engineering situation.

REACTION TO THE CARROLL PAPER (1)

Caroline M Clapham, University of Lancaster

The Carroll report states that the Davies Test (EPTB) is now unsatisfactory because:

1 It was not designed to cope with the number of students and diversity of courses that there are today

2 Many students fail to finish their courses because their English is not good enough

3 The emphasis in language teaching and testing has changed from an atomistic approach to a broader sociolinguistic one

4 The advent of ESP has led to fewer teachers and testers working towards the needs of all language users.

I am not sure whether 1 matters much for a test of the Davies kind and I know too little about the test's concurrent and predictive validity to comment on 2. However, a combination of 3 and 4 has led to such a drop in ithe test's face validity that it is losing the confidence of its users and will probably have to be changed. (Whatever Palmer and Bachman may think, face validity is of the utmost importance when a test is administered by non-statistically minded bodies.)

I should have liked to have been able to discuss the differences in content between the Davies Test and the proposed replacement, ELTS, but in his *Specifications* Carroll does not go so far as to describe the items in any detail. I can only, therefore, comment on some of the issues that lie behind them.

ELTS, as reported in these specifications, is planned to be radically different from the Davies Test, and I am rather alarmed by the number of changes envisaged. The proposals follow three swings of the pendulum of language teaching and testing theory: the new test is to test communicative competence, it is to be divided into different modules to test ESP, *and* it is to be criterion rather than norm referenced. There are very good arguments for all of these, but, since none of the three is yet well tried and tested, I wonder if it is wise to go for all of them at the same moment.

Even if we accept the arguments for the first two, what about the move to criterion referencing? At the present state of our knowledge, is this practicable, and is it in any case necessary?

Criterion Referenced Tests

To take the question of practicability first: for a criterion referenced test to work it must have a comprehensive list of language objectives to which it can be tied, and it must also be capable of being pretested to see whether each item tests the criterion in such a way that those who know it pass, and those who do not, fail. Carroll tackles the first of these very thoroughly – one of the main aims of his *Specifications* is to present a list of language objectives – but what about the pretesting? How, for example, should the proposed 250 multiple choice items be analysed? Traditional item analysis is, of course, norm referenced, with items being assessed according to a comparison of the reactions they elicit from high and low ranking students. In criterion referenced testing, though, the ranking of students is, by definition, irrelevant. Testers are not interested in whether more students at the top than at the bottom get an item right. They want to know whether those who know the subject, and those who do not, pass or fail accordingly. It may well be that since items also have to be checked for level and ambiguity, some sort of initial norm referenced analysis will have to be used, but what should happen after that?

Carroll treats this problem very lightly. He implies that since the aim will be to match students with their language requirements rather than with their fellows, representative samples will not be needed for the test preparation. The implication seems to be that norm-referenced tests need to be tried out on representative samples but that criterion-referenced ones do not. Carroll specifically says that once the communicative demands are defined, it is the test's job to decide how a particular candidate measures up to them, not to see how 'performance characteristics are distributed throughout a population of applicants . . .' He seems to be confusing the preparation of the test with its final administration. If a validated test is criterion-referenced, each candidate's performance will of course be compared with the language specification and not with that of other examinees, but before that stage is reached, the test **must** in some way be tried out on representative samples for level, reliability and validity. (Confusingly, Carroll does say in direct opposition to what he says elsewhere, that, 'our performance standards will derive from ongoing courses and their students'.)

Since there are these problems of construction, does this proficiency test need to be criterion referenced? I agree that a student's level of English should be compared with the level he needs for his course rather than with

that of the other candidates (it would make no sense for a proficiency test to pass people according to percentiles, as it is rumoured some O and A Level boards do) but with a proficiency test, is such a fine diagnostic tool needed? Would not a norm-referenced test with set, validated target levels for each subtest, serve the purpose as well? As Carroll proposes, the marks in each language area could be set differently for the different disciplines and course demands, and the final score sheet could provide all the information described. I do not want to stray further into the marking system just yet, but I do want to question the necessity of embarking on the ill comprehended area of criterion referenced testing for proficiency, when there are at hand hardy statistical methods for norm referenced tests.[1]

The profiles

If a test is to be criterion referenced (and indeed preferably when it is not), there needs to be an adequate specification of the candidate's language requirements, and this specification is, of course, the nub of the Carroll report.

I shall not comment on the coverage, applicability and feasibility of these specification criteria in language learning and use, since they are based on John Munby's description of communicative needs (Munby, 1978). What I shall do is look at the manner in which they are used here. However, before I do that I must say something about the six profiles described in the report. I should have liked to have seen how much they helped to straighten the tester's muddled way through communicative competence and ESP, but unfortunately I cannot do this, as their use here is vitiated by the fact that five of them seem to have been invented by their compilers. Carroll gives reasons for this and I can see that 'comprehensive observational studies of the participants' would have been very time consuming. However, without such studies, surely the profiles are almost useless. Although Carroll points out that the field work depends too much on the subjective judgement of the compilers, he still draws conclusions from it. For example, he says that the profiles will be used to identify common areas, and areas of specific need on which diversified tests can be based. Indeed, most of his findings throughout the report, for example target levels of courses, variation in demand between different disciplines, and extraction of factors, are based on this

[1] Since preparing his *Specifications* Carroll seems to have tempered his views on criterion referenced testing. In Carroll 1980, page 10, he says 'Emphasis on the pre-specification of communicative tasks lends itself to criterion referenced techniques, but it is far too early to consider dispensing with the elaborate and well worked-out procedures of norm-based statistics.'

'data', and cannot therefore be trusted. This is a pity, because it is an excellent way of setting out the demands of different courses in a tangible and comparable way. In his explanation for collecting the data in this manner, Carroll rather startlingly says, 'The ultimate validation of our methods would be the effectiveness of the test based on their results.' To spend time constructing tests according to possibly false data would seem a waste of time; and if the tests were invalid, how would one know whether the data or poor test construction was at fault?

Even if the profiles were not just the result of educated guesses they would have been of little use because they are 'personalised' in order to 'counteract the natural but dangerous tendency to overgeneralise about communicative needs'. Unfortunately, this personalisation, far from counteracting it, actually encourages overgeneralisation. Nothing can prevent the reader, or indeed the writer, from treating the profiles as typical. That Carroll himself is misled is shown when he says that the first necessity is to study the needs of 'a typical student'.

The specifications

For the purposes of testing, the specifications fall into four categories:

Cat. 1 : Spec. 0	Student's background
Cat. 2 : Spec. 1, 2, 3, 4, 5, 7	Setting
Cat. 3 : Spec. 8, 9, 10	Manipulation of Language
Cat. 4 : Spec. 6	Target levels

Cat. 1 Carroll's argument is that the candidate is to be matched with his language requirements regardless of his background. In this case, age, mother tongue and previous English knowledge are strictly irrelevant. (This, of course, ignores the use of contrastive analysis in ESP testing.)

Cat. 2 and 3 These two categories form the basis for the criterion referenced list of objectives mentioned earlier, and had the profiles been based on solid research, would have provided the raw material from which a test such as ELTS could be constructed. There is obviously little point in looking at the substance of these profiles here, but I had hoped that we might be able to see how such material would be transformed into test items. Unfortunately, though, the report does not takes us to this stage.

If they were well researched, categories 2 and 3 would also provide invaluable evidence for or against the shift to ESP. Research is urgently needed into whether an ESP proficiency test is actually necessary, as the preparation of

parallel ESP modules makes the construction of a valid and equitable test time consuming and difficult.

Cat. 4 With specification 6, target levels, we come on to marking, and a host of unsolved problems, of which the two most important are:

a) How does one set reliable and valid target levels?
b) How does one marry these up with reliable test scores?

Munby's target level specification consists of a two dimensional matrix giving size, complexity, range, delicacy, speed and flexibility by verbal medium, with levels ranging from 1 to 7 (see *Specifications* Appendix). This is set beside a tolerance matrix giving error, style, reference, repetition and hesitancy by verbal medium, with levels ranging from 1 to 5. When he introduces them, Carroll uses Munby's scales, but in his succeeding discussion converts the 7-point scale to percentages, for, he says, comparative purposes. This seems to me to be unnecessary since two 7-point scales can easily be compared, and it is also dangerous as it makes the scale look deceptively equal interval. Indeed Carroll seems to treat it as such, for he has worked out means on the strength of it.

Here again the inadequacy of the data means we can make no deductions about comparative levels of course demands, but we can look at how the system might work. Presumably the plan is that the testing staff would fill in the course target levels after consultation with instructors and heads of departments. It is an attractive scheme, but I doubt whether there would ever be enough time for it to be implemented, especially since it would need frequent updating as course demands changed. I doubt too, whether many heads of departments would want to be so involved. In practice, time constraints would probably prevent the matrices being used, and test compilers would be happy if they were able to get amalgamated listening, reading, speaking and writing levels.

Of course, whether the levels are simple or complicated the same problem remains: how can they be made valid and reliable? The report admits that the profile levels it gives may not be reliable, but it does not say how they could be made so.

It is also not clear from the report how Carroll intends to use the tolerance levels since they are not included in the marking scheme graph. Although the idea of tolerance levels is very appealing, I wonder how much they would improve the precision of the results. Since the target and tolerance levels are based on different scales it is difficult to compare the two, but if research bore out Carroll's statement that tolerance ratings are negatively correlated

with level of demand, and if this correlation was a high one, then tolerance might well be omitted. Certainly the marking system would be much easier if tolerance could be left out.

Setting the target levels is hard enough, but matching these with test results is even harder. If the whole test was subjectively scored according to the same 7-point scale, it might be possible for both setters and markers to determine their levels in the same way, though even here, decisions would have to be made about such questions as how much flexibility to allow. (I am not sure, for example, where Carroll's four point error comes from, nor what his 'significantly above the level' means.) Once there is a multiple choice element in the test, the difficulty is compounded; there would **have** to be many trials of the test, and the results would **have** to be correlated with students' actual abilities and with the target levels. This would take time, and would lead to all the usual validation problems, but it would be absolutely essential if the test was to be fair both to the prospective departments, and to the students whose careers were at stake.

Test Design

The mention of multiple choice questions brings me to the proposed test format, and it is surely only once we have a detailed idea of what this will be that we can know whether the *Specifications* are indeed the breakthrough in test design that the author claims. It is only once we see how they can be applied that we can know whether the ensuing battery will have face, content and construct validity. Alas, the report stops here. It does give a bare outline of the proposed test, listing the number of items and areas to be tested, but it does not describe the items in any detail. All it says, tantalisingly, is that 'in our preparatory work, we have had no difficulty in devising test types to measure the varied communicative features revealed in the specifications . .'

Finale

The Carroll report makes far-reaching suggestions for changes in proficiency testing, and by providing a concrete plan for people to criticise, should advance our knowledge of how to test communicative competence. However, a very great deal of research will have to be carried out before a reputable international test can be based on it.

BIBLIOGRAPHY

CARROLL, B J
Testing Communicative Performance. Pergamon Institute of English. 1980.

MUNBY, JOHN
Communicative Syllabus Design. Cambridge University Press. 1978.

REACTION TO THE CARROLL PAPER (2)

Clive Criper, University of Edinburgh

The stated aim of the English Language Testing Service (ELTS) as set out in these specifications is quite clear. It is:

1 to test whether any student is already able to cope with the language needs of his academic course;

2 to assess the nature and duration of any language tuition that a student might need to bring himself up to the level at which he could cope with his academic course.

What is claimed to be new in ELTS is a matching of course requirements with the test instrument. ELTS is thus designed to be a model of a criterion-referenced test where the criterion is based on a close analysis of the real communicative needs of a student attending a particular course.

I think there can be no disagreement with these basic aims. They are aims which we can applaud without reservation and indeed feel virtuous that we have taken the path of righteousness.

Reality, unfortunately, cannot be kept entirely out of sight and out of mind as one reads the apparent basis for the ELTS test – at any rate as specified by Brendan Carroll in his paper. Let me take in turn some areas in which reality and ELTS ideology appear to be in conflict.

Communicative Needs of the Users

The whole argument against the use of general proficiency type tests for use as placing tests for Higher Education students rests on our ability to identify different student's needs. This is clearly the crux of the argument of the paper by Brendan Carroll and a large proportion of the paper appears to be spent on 'proving' this fact. The 'proof' offered, if it is meant as a proof rather than a statement of belief, is highly spurious.

ELTS Argument

The basic starting point of Brendan Carroll's work was Munby's needs analysis. Without commenting on Munby's thesis as such, it will suffice to say that Carroll follows the outline of Munby's specification parameters. These are a set of typologies which are meant to cover all the important linguistic

and social areas which might affect the language to be used in any particular situation. Each typology then divides up the world into a number of discrete categories against which a student's needs for English are matched.

In the present instance this kind of matching has been carried out for six 'students', the majority of whom, it appears, are imaginary. The needs have been analysed on an intuitive basis by staff who have some knowledge of the subject area.

When it comes to specifying the proficiency target levels in each of the four skills a numerical figure is given on a subjective basis. Notwithstanding the disclaimer, **these figures are then used as if they are genuine experimental figures on a true equal interval scale.** Scores are added and averaged, and are treated as being on an equal interval scale from which conclusions can be drawn about the length of study necessary to reach a certain level.

In another area – that of the 'Essential Language Skill Categories', a further quantitative comparison is made between the six subjects and the same spurious 'proof' of connection between various of the subjects is made.

There are other areas of the specification parameters, eg microfunctions, where much of the theoretical basis of the specification might be challenged and, inevitably, many areas where one could argue at length about the rating of needs provided by the analysts. Such arguments would only be of interest, however, in showing that the analysis made is essentially a theoretical one and not an experimental one.

Course Requirements

There is an unstated assumption in the whole paper that individuals picked out for illustration of the scheme are going to institutions which are sufficiently similar for generalisations to be made about the communicative needs of their students. The ideology of the ELTS scheme requires a close matching between student and institution.

I am extremely doubtful whether the language needs of students going to do postgraduate work in Agriculture, for example, have more in common than between some students doing, say, Medicine and Engineering. If one tries to specify the content of lectures in Engineering, it becomes apparent that the individual variation in lecturers, techniques and approaches outweighs anything that the content may have in common.

In addition, as Carroll rightly points out, Universities and other institutions in the UK have considerable autonomy. Within most institutions there is also

considerable variation in Faculty policies and, even more importantly, in departmental policies. It is also true that individual Supervisors within the same department have very different views of the minimum level of English that they require from overseas students. This latter fact of life has, in the past, been one of the major reasons why Universities have found it an impossible task to specify clear-cut language requirements for their post-graduate students.

The implication of this is two-fold. Firstly it will never be possible to specify **in detail** the requirements in the various skills for a particular group of subjects across all Universities. Testing Centres, such as the British Council overseas offices, will not be able to match institutions' profiles very closely. It follows that, secondly, a fine assessment of needs, in test terms, will be wasted.

Practical Considerations

There are three main areas to be considered – testing centres, particularly overseas, the UK 'customer', be it University or Technical College or hospital and the test producer.

Test Producer – Reference has already been made to the difficulty of producing reliable generalisable 'profiles of needs' except where there are gross differences. Leaving aside any argument about the ease or difficulty in designing test items to cover the 'specification parameters', a major problem comes up in the plan to use subject tests, eg reading comprehension using specific subject area texts. While such a procedure appeals to common sense and thus has great face validity there are at least two types of difficulty.

Firstly, the subject specialist, whether testee or teacher, tends to require more and more specialist texts. To the specialist there is no such thing as an 'agricultural' text covering all related branches of the subject, any more than there is a 'medical' text. The idea of a 'special purpose' text for a wide range of sub-disciplines is contradictory and paradoxically may potentially be more subject to criticism on the grounds of non-validity than a more general text.

Secondly, it may be more difficult to control the texts for background knowledge of the testees. Background or factual knowledge is an enormous advantage in answering comprehension questions. While it may be argued that there is a certain basic minimum knowledge that can be expected of any student in a particular subject, in practice no such minimum knowledge exists, both because of the educational and cultural background of different students and because of the existence of a multitude of sub-disciplinary backgrounds that students may have. A language test as such cannot afford to be

seen to be classifying students according to their subject knowledge rather than their language ability, otherwise receiving institutions may come to reject its use.

Testing Centres - Carroll makes reference, quite rightly, to the importance of cost and time that would be involved in the ELTS overseas and states that there is a need for a quick screening test. In any overall assessment of ELTS I think that the time/money cost has to be weighed very carefully against the extra information which a test based on an assessment of projected communicative needs requires. This is particularly so if the testing centres will not, in practice, have the information about the real requirements of the receiving institutions. Considerable judgement will also be required to make recommendations on the basis of the test and the way that the Davies test has sometimes been administered and interpreted leaves one with considerable doubts about using a far more sophisticated instrument.

UK Customers - My experience suggests that in Universities at least the level of sophistication in interpreting and using English test scores is very low indeed. At Edinburgh, two test scores are widely used, Davies (EPTB) and the English Language Battery (ELBA), and only a limited number of people understand what the somewhat odd figure of 'Davies 40' means, and the similar odd figures of 'ELBA 50 and 70'. Only the specialists have an idea of the relationship between the two. Considerable difficulties will inevitably arise in interpreting either scores or band scores for different skills and I fear that many institutions, or at any rate individuals within them, will operate on some rule-of-thumb averaging operation. If that happens, then the whole purpose of the ELTS 'profile' design will be vitiated.

Summary

The need 'to test whether a student is already able to cope with the language needs of his academic course', is crystal clear and happily the British Council has taken up the challenge. Carroll's 1978 presentation of the specifications for ELTS, aimed at testing a student's potential ability to operate in a study environment raises issues in testing as interesting and as problematic as those in the teaching of ESP. What will be needed will be a programme of development and validation over several years which will deal with the real world of testing and needs rather than the hypothetical constructs of Carroll out of Munby.

BACKGROUND TO THE SPECIFICATIONS FOR AN ENGLISH LANGUAGE TESTING SERVICE AND SUBSEQUENT DEVELOPMENTS

Ian Seaton, ELTSLU, The British Council, London

Consideration of Carroll's paper divorced from a knowledge of the context in which it was produced and the developments following its publication is problematic, since the questions 'What led to these specifications?' and 'What has been done or what is likely to be done about them?' recur in the reader's mind. The paper reproduced above represents but one phase, although a vital one, in the complex process of establishing the English Language Testing Service. No further information on the subsequent development of the service had been made public when the reactions to Carroll's paper were written. Some information on the background of ELTS and more recent developments is therefore given below to provide a context for the reactions and discussion.

In the latter half of the 1970's the British Council was faced with the need to introduce a new or modified English proficiency testing system geared to the changes in ELT developments, notably in ESP, and to the changes in the needs of sponsored students seeking to come to Britain. However, it was faced with two closely linked constraints – one professional, the other financial. The first was that even in January 1978 there was no body of research into the testing of ESP which could be drawn upon. English proficiency tests were being conducted for special groups at that time, but not on anything approaching the scale that the Council would be required to test. The major ESP test system established in Britain by then was the PLAB test administered by the General Medical Council, and some industrial companies had commissioned publishers or other groups to construct ESP tests for internal use in their own training programmes. But results of any research that may have been carried out on those tests had not been published. This contrasted sharply with the volume of research by Lado, J B Carroll and others that was available to the constructors of the TOEFL, EPTB, ELBA and other English proficiency tests more than 15 years previously. Secondly, the Council was entering a period of increasing financial stringency which precluded the possibility of commissioning elaborate in-depth research.

Nevertheless a decision was made in 1977 to commission six small teams of qualified teachers and consultants to devise the specifications that Carroll has reported. The teams chose to use the Communicative Needs Processor proposed by Munby (1978) to organise their survey and specifications.

Early in 1978 the recommendations of Carroll's report were accepted in principle and new teams drawn from the British Council English Language Division and the University of Cambridge Test Development and Research Unit edited the specifications further and produced items for a preliminary version of the test known as ELTS. This test observed the two phase (screening test and subject specific modules) system proposed by Carroll and was trialled in Britain later in the year. After analysis of the results, revisions were made and a second version pre-tested overseas in 1979. After further modifications a third version was produced and put into operation in a number of selected countries from early 1980. It can be seen that although the speed of introduction was carefully controlled, resources were fully committed and it was not possible to publish reports of the developments as they took place. However the *User Handbook* containing details on the nature of the test was published in late 1980, and the *Specialist Handbook* with technical details of the tests is scheduled for publication in late 1981. Details of the pretesting and analysis of the results will be abstracted from the handbook and published separately as a very brief report at the same time. Copies of these publications can be obtained from the British Council English Language Testing Liaison Unit or from the University of Cambridge Local Examinations Syndicate.

One of the latest and most important developments is that within the overall validation framework an extensive follow-up validation study of the test is being undertaken by the English Language Testing Service in cooperation with the Institute of Applied Language Studies, University of Edinburgh. This study should give information which will be valuable to the test consumers and which could well lead to modification of certain specifications or formats in the future.

REPORT OF THE DISCUSSION ON TESTING ENGLISH FOR SPECIFIC PURPOSES

J Charles Alderson, University of Lancaster

The purpose of the discussion was to consider the possibilities and problems of testing within an ESP framework, and not to focus on the English Language Testing Service recently established by the British Council and University of Cambridge Local Examinations Syndicate. However, to date almost no attention has been given within testing circles to the problems of ESP testing, so that one of the very few articles of relevance to the debate is the *Specifications for an English Language Testing Service,* written within the British Council by Brendan Carroll. In addition, the ELTS is one of very few cases so far in the United Kingdom of an attempt to carry out ESP testing. (One other case is the PLAB test of the General Medical Council.) Inevitably, therefore, much of the debate centred on the ELTS since it provides a practical example of the problems of ESP testing. For this debate, the *Specifications* document proved to be an excellent starting point, raising as it does so many issues, and attempting to introduce ideas into the field of testing from the 'outside' EFL/ESL world, as well as from applied linguistics. It should be remembered that this document was originally produced as a paper for discussion **before** the final specifications were worked out.

Proficiency versus Achievement

The discussion confined itself to the topic of proficiency testing for ESP. This was partly because the *Specifications* paper itself is concerned with proficiency testing, but more importantly because there is a sense in which the development of achievement tests of or for ESP simply does not present a problem. Any achievement test must crucially depend on its content. That is, to be valid, an achievement test must be based on the syllabus which has preceded it: otherwise it is by definition not an achievement test. Thus the validity problem of an achievement test is essentially a sampling problem. To the extent that it is possible to develop a syllabus for specific purposes, it is also possible to develop a specific purpose test, since it 'merely' has to reflect that syllabus. The problem of what an ESP syllabus looks like: what items, skills or content it contains, and how that content is determined (be it through prior needs analysis, negotiation with learners, fiat, or whatever), is simply not the concern of the constructors of achievement tests. Proficiency tests, on the other hand, are not based upon any particular syllabus, again by definition. One is, therefore, faced with the problem of deciding what must be tested.

The Need for Specific Tests

Once it was agreed that the discussion was properly concerned with proficiency tests, it was then necessary to clarify why proficiency tests should test ESP. The *Specifications* document suggests, in the Foreword, that it is necessary to 'specify the communication needs' of potential testees, because of the inadequacy of previous test instruments:

> 'there is always a number of students who have to abandon their studies and return home because of their language inadequacy and the progress of a much larger number is adversely affected in one way or another by language problems.'

Thus the *Specifications* document aims to explore 'ways of devising a more up-to-date system which will be able to cope with a problem of the size and diversity of which the earlier system had not been designed to meet'. Later it is made clear that the need is for tests 'which will cater more completely for the many different types of programme (of courses of study) we are testing for'. Thus, there is a need for a new test or series of tests because poor students are getting through, or rather the Davies test (EPTB) is failing to identify students who have problems, and it does not cater for the needs of a wide variety of students. Unfortunately we are not offered empirical evidence that the existing test has in fact failed to identify students with problems. Indeed it was suggested that it may be the case that 'poor' students are being accepted despite low EPTB scores, and that the problem is not so much the identification of weakness, but the lack of remedial action.

We are usefully given criteria by which a new instrument can be judged: it will identify such students, and it will meet the needs of that variety of students more adequately. However, it does not follow from the 'fact' that the existing instrument is deficient that what is needed is an ESP test, or a battery of specialist tests: one plausible solution might simply be a better general test, constructed along similar lines to existing instruments. The evidence suggests that different academic departments do indeed place different language demands upon overseas students. It is certainly plausible that an undergraduate course in Engineering will have different linguistic requirements from a postgraduate course in linguistics. It is not clear, however, that this implies the development of separate tests for Engineers and Linguists. Even if the activities they have to contend with are entirely dissimilar – for example, a taught course contrasted with a masters degree by dissertation alone – it does not follow that different tests of language ability are required. It could be that all that is needed is that different levels of proficiency are required for different subject disciplines. Thus, in order to succeed in Engineering, a student 'needs' an EPTB score of, say, 36, whereas

to succeed in Linguistics, a student 'needs' 42 on EPTB. Indeed, this is typically the way in which entry requirements have been varied for different disciplines, in the UK and in the US. It may be the case that separate tests are required, but we do not have satisfactory evidence yet that this is so.

One major argument advanced for specific tests is that of face validity: a test for Engineering students should look like a test for Engineering students and not like a test for Social Scientists, or worse, Generalists. There is a very real problem with face validity arguments of this kind which is related to the question: Suited for whom? Will all engineers – electronic, electrical, chemical, civil, mechanical – agree on the face validity of an Engineering test?

Perhaps the most powerful argument for specific tests is that of the diagnostic value of a profile of a student which can be matched against the communicative needs of his particular course of study. Regardless of the presence or absence of predictive validity of such a profile – predictive, that is, of final academic grades, or whatever – there is, or may be, value in profiles of students' abilities, relatable to institutional criteria, for both administrative purposes (that is, admission decisions) and for pedagogic purposes, since hopefully such information would allow remedial action to be taken on a language course, for example.

One further advantage of such a profile is that it might encourage people – institutions – to be explicit about what they want and expect students to be able to do (with language), if the students are to succeed. This, however, presupposes that it is actually possible for subject departments – or indeed, even applied linguists – actually to specify what the language-related requirements are. This may not be the case: it may be impossible both to determine what the linguistic demands being made on any individual actually will be, and, furthermore, it may be very difficult to specify in advance what difficulties a particular student will have in meeting those linguistic or language-related demands. Some students, it was argued, will learn to cope much more easily than others. Thus a proficiency test, which simply labels a student at one point in time, gives no information about learning potential, and for that very reason may be inadequate. Two students may achieve the same proficiency score, but have very different potential: one student may have greater aptitude or adaptability than the other, perhaps having learnt the foreign language for only six months, whilst the other has studied it for fifteen years: in such a case one might expect the student with the shorter learning history to have greater potential for coping in a foreign language environment. Thus, what may be needed is not only a proficiency test, but in addition an aptitude test, or an adaptability test, or details of individual learning histories.

The problem with predictive validity of any test is that so many variables enter into a student's ultimate performance, in addition to whatever the particular test is measuring, that one is unlikely to get higher validities for aptitude tests than for proficiency. This would be an argument against replacing proficiency tests with aptitude tests. The issue was raised of whether it is in any case the task of a language test, be it general or specific, to predict performance in, say, Physics. Might it not be perhaps less presumptuous and more valid, simply to require that a language test should predict how much a student will improve in language, and to what level? Thus what one needs to know is not to what extent EPTB or any other proficiency test correlates with academic performance, but to what extent it correlates with itself, or another 'relevant' measure of language ability, at course end, when final academic success is being judged. The diagnostic argument is that we need to be able to predict the difficulties students will have **because of** language: the crucial question is: Is this knowable? Be this as it may, existing tests are frequently used as if they were predictive of final academic success, or as if they predicted eventual language proficiency levels. EPTB scores, for example, are often interpreted as indicating a required number of weeks of English tuition before commencing academic study: a student achieving a score of 32 on EPTB may be expected to take a twelve week English course, whereas a student with a score of 36 might be required only to undergo six weeks. This is a misuse of the test score, because the test was not validated in such a way, and is in any case unintelligent because it ignores language learning history.

How Specific is Specific?

For the sake of the argument, it was assumed that specific tests are needed, that evidence is, or will one day be, available which indicates incontrovertibly and uncontroversially that a general test is simply not doing its job. (This assumption, it was noted, implies that we know what the job of a proficiency test is: that we can answer the question: Proficiency for What?)

The problem that was addressed was: how specific should a specific test be? Is it, in other words, possible to draw up a real specification for a language test? Carroll claims that the development of tests within ELTS represents 'a process of diversification of test instruments to meet the diversity of test situations'. The question that inevitably arose was: when are 'test situations' no longer diverse, but similar, or similar enough? The ultimate specification of a test situation must be that of one individual at one point in time: above that level, a claim of specificity must be invalid for some individual at some point in time. Yet it is in principle impossible to devise an instrument for one individual at one point in time, which is in any sense reliable and valid, since to determine the extent of such reliability and validity, one has to be able to

compare performances on the same instrument. Thus, *a priori,* a specific test is impossible. However, it was felt that there may be practical reasons for constructing a 'more or less' specific test – a test for engineers or for chemical engineers, or for chemical engineers going to study at Lancaster. Nevertheless, it was pointed out, there are practical problems in matching specific students to tests. Which ELTS modular test, for example, out of the six presently available (Physical, Life, Social and Medical Sciences, Technology and General Academic) should be taken by a student of Urban and Regional Studies, whose course will include Law and Economics courses as well as courses in Technology? Should such a student take a General Academic test, (ie less specific), or should a test be developed for Urban and Regional Studies, (ie more specific)? What about the (frequent) cases of students who have a background in Physical Sciences, who are coming to the UK to do a (to them) novel course in Technology? Do they take the Physical Science test or the Technology test? It is not clear that any principled decision is possible, and if the tests are not comparable, then students suffer the luck of the draw. How is a student to decide which test to take? On what basis can he choose? How can a British Council English Language Officer advise him? The point is that the level of specificity chosen for the test is inevitably arbitary. One can attempt to analyse communicative needs – looking at study situations, for example – and then find what different study situations have in common. One thereby extracts the specific from specific situations, abstracting generalities in order to cover more situations. To what extent can such an endeavour be characterised as constructing an ESP test? Might it not be the case, as suggested in the discussion about communicative tests, that if one abstracts far enough, one might end up with linguistic proficiency or Grammar, as being common to all language-related situations?

Another problem frequently encountered with specific tests is that of previous knowledge of the subject matter; at what level of specificity or generality can one be relatively sure that one is not testing subject-matter knowledge rather than linguistic or communicative abilities? Can one (should one) be sure that prior (subject) knowledge will not give one candidate an advantage over another candidate? A related problem is that of lack of knowledge: a specific test might well assume or presuppose subject knowledge that the testees do not have; native-speaker A-level students might have such knowledge, and the text may be premissed upon that, but differences in educational and/or cultural backgrounds may mean that overseas students may not have the knowledge. Two questions arose: does it matter, since overseas students will in any case have to read texts premissed upon pre-existent knowledge, usually within an English-speaking community? And how can one possibly avoid involving prior knowledge, since comprehension and presumably production must depend upon the prior existence of some set of knowledge?

One might, if one were simply to characterise the tasks that students have to engage in their target situations, be able to specify a set of Study Skills which are common to study situations, and one might then attempt to measure such study skills in the proficiency tests. Such is, in effect, what the ELTS test is becoming. The question then is whether one actually needs a Munby-type specification at the level of microskills, of the sort advocated by Carroll, in order to arrive at a relevant set of study skills for inclusion in a language or study skills test. In fact, it is, as suggested already, probably impossible for an ELTS-type test to base itself on specifications of the type advocated in the document: the specifications are simply too specific, and certainly do not allow the generation of texts or text types. Criper has already mentioned the danger of specific texts: that they satisfy no-one because they are not specific enough. The fact is that the Communicative Needs Processor does not help one to select texts or items for a test.

The point of the specification of microskills and the like is that such specifications should be reflected in the final test, after editing and pre-testing, in the proportions (with the weighting) that the specifications indicate as necessary. Traditional item analysis procedures and criteria for item rejection must therefore be applied with caution if the final test is to reflect the original specification.

One of the problems of the specification of microskills is that not all can be tested on any one text. This is revealing of the nature of such microskills: if they are not applicable to a particular text, to what extent are they generalisable skills? To what extent are they not rather **product**-oriented skills than **process**-oriented – in other words they are in effect glossable as 'the skill of processing X text feature' rather than '**how** X text feature is processed'. If X feature is not present in the text, the specific microskill as currently defined cannot be tested. If they were defined as process skills, then it might be possible to measure them, not on X feature, but on Y or Z features, which require a similar process. In fact, at present, the microskills are nothing more than the ability in general to process a certain linguistic feature.

Be that as it may, if one believes in the necessity for a specification of test content based on a needs analysis which will identify the types of skills that students do need, then it is crucial that test results should show how far individuals have met such specifications. One problem raised was: To what extent does the **lack** of such (pre-) specified skills lead to student problems or student failures? A further problem lies with the interpretation of the scores that result with a test specified in such a manner. Does any given score equal the same score gained by other testees? Are all 70%s to be interpreted in the same way? Surely, any less-than-perfect score will be composed of a different constellation of 'microskills'. Yet the claim that seems to be being made

regarding microskills is that they are all equally necessary: no suggestion has been made that as long as one achieves, say 80% – **any** 80% – of the skills, one will no longer be 'at risk' linguistically. Presumably until further information is gathered about the relative importance of each microskill one will either have to regard perfection – 100% scores – as the only adequate test results, or, alternatively, be prepared to report 'scores' for each microskill – a sort of profile within a profile. It is true that the same predicament presents itself on a grammar test: any less-than-perfect score will comprise different successfully completed items. However, there is no claim associated with grammar tests that one **needs** mastery of modals, tenses, concord or whatever. How the score is arrived at is usually (although possibly erroneously) regarded as unimportant. The claims for enabling skills, microskills, are at present much stronger than that – possibly too strong, it was felt.

The question of specificity raised two further issues: that of extrapolation and that of comparability. Taking the latter first: how can one compare performances on two different, ie specific tests? If a medical student takes test A and an engineering student takes test B, how is one to determine whether a score of 50% on one test is equivalent to a score of 50% on the other? How are the parallel tests to be balanced and calibrated? Presumably one can only compare test performances if they are criterion-referenced: that is, scores are comparable when they all meet the criterion, or when they all fail to meet the criterion, since criterion-referenced scores are essentially binary. Thus, specific tests may well be necessarily criterion-referenced: the problem is, how to develop criterion-referenced tests. How does one establish the internal validity of such a test, and in particular how is one to conduct item analysis? (One answer suggested was point-biserial correlations). The problem of extrapolation, familiar from the discussion of performance tests, also exists with specific tests: How is one to predict from one performance on a specific test to performances in 'real-life'? Although the problem seems to be solved by needs analysis, which purportedly helps one to identify the real-life tasks and texts which one can incorporate in one's test, the fact is that increased specificity of the type brought about by needs analysis, particularly of a Munby nature, decreases the likelihood of extrapolability: the more specific the test/task, the less general can one be in one's conclusions from that text/task.

To what extent do proficiency tests have to be specific? Having seen the problems of specifying the level of specificity required, the discussion returned briefly, to the original question ('Who needs specific tests'?) to consider the extent to which any student coming to the UK to study can survive with only a limited range of skills, of the type identified by a needs analysis. Students themselves are reported as perceiving their greatest problems as being not related to Study Skills nor specific to their academic discipline, but rather to survival in British society: students consistently

mention problems of adaptation to the UK, the problems of being immersed in a foreign environment. Frequent mention in the literature (see Davies, Moller & Adamson (1975), is made of the importance of social and welfare problems in the minds of overseas students: it is possible, however, that these 'welfare and social' problems might, to some extent at least, be linguistic or 'communicative' problems. Academic tutors may regard their students' subject knowledge (or lack of it), or their particular academic orientation, or their (non) adaptability as being their main problems. Yet these perceptions are often at odds with the students' own: are the tutors overlooking the real problems, or are the students unaware of their main difficulties, or reluctant to admit them?

What do Students Need?

The only way such questions can begin to be answered is through empirical study, both of the type carried out by Davies, Moller and Adamson (1975), and also by means of needs analysis. The aim of needs analysis is to answer the question: Proficiency for what? In this respect the *Specifications* document is valuable in pointing up the need to determine the communicative needs of students before establishing test content. Such needs analyses must be data-based: speculative research about the needs of 'typical' student is only of value if it susgests areas that empirical research might look at. One attempt to establish communicative needs is being made by Cyril Weir, with the Associated Examinations Board. Using an observation schedule adapted from Egglestone, Galton and Jones (1975) – the Science Teaching Observation Schedule – together with interviews with subject teachers, students, and follow-up questionnaires, Weir is attempting to do a needs analysis of interaction and events. His aim is to gain an overall impression of what happens in lectures, seminars and practical classes in Science, Engineering and Social Sciences courses with a view to finding activities and requirements which are common across disciplines, and which could therefore be incorporated in a test (which need not *a priori* be subject-specific). He aims to get a profile of types of activities, in order to see whether students can deal with the particular activities that the analysis shows they have to be able to deal with. It is hoped that a national questionnaire sent to staff and students in relevant departments will provide a broader basis for a description of what students have to do through language in respect of their courses and throw light on the language problems that staff and students have noticed. Although this is an interesting attempt to provide empirical justification for Study Skills tests or even subject-specific tests, there is a danger that mere observation will lead to a confusion of frequency of occurence with importance of an activity: 'the more frequent, the more important'. In fact, this is not necessarily the case: students may find relatively infrequent activities very difficult, and of crucial importance. The problem is to identify common areas of difficulty which are of importance. The hope, of course, is that if the test is built on a specification of what the student has to do, the receiving institution can judge whether a failure to do it is important for their particular course or for that particular student.

Another, practical problem of such research, apparent in Carroll's use of 'typical' data rather than real data, is the problem of sampling. Attempts to make general statements about what students need language for, inevitably come up against the sheer size of the research needed in order to be anything like exhaustive. Even a case study of one department in one university (see Murphy and Candlin, 1979), can take years of research without leading to generalisable results. Allied to such a practical problem is the problem posed by the need to specify, at least according to Munby/Carroll, the target levels and the tolerance conditions of language use. The whole area of tolerance conditions is very under-researched, and will require a vast amount of research effort before anything meaningful could begin to be said about the way in which native speakers judge foreigners' English: Are they more affected by occasional grammatical errors than by consistently poor handwriting or pronunciation? Does this vary from event to event (being different in seminars and tutorials, for example)? Including such parameters in our language tests is at best an ambitious goal, at worst impossible.

Profiles

The claim of *Specifications* and the justification for the existence of student 'profiles' is the 'fact' that a medical student needs x score on test A, y score on B and z score on test C, whereas an engineering student, for example, may require z score on test A, x score on test B and y score on text C. It was pointed out that we have no empirical evidence that such is the case (and we have suggested that it might equally plausibly be the case that a medical student simply needs a higher (lower) score overall on a particular test, than an engineering student). The point about profiles is that one needs different tests in order to produce them, and one needs to be able to show that such differentiated tests are necessary.

The aim of ELTS tests is to produce profiles, based upon the specifications arrived at by needs analysis. Information from an ELTS-type test might be of value diagnostically to teachers and syllabus designers. More traditional proficiency tests like EPTB and ELBA are not intended to yield diagnostic information, and although they are unfortunately frequently used as placement tests, they usually result in heterogeneous groups, in remedial or pre-sessional language courses for example. it could be that an ELTS-type test could be of use in identifying areas of students' weaknesses relative to their

communicative needs which would enable one to group together those students with a common problem and a common need to overcome it. This might suggest that whatever the problems involved in using an ELTS-type test for proficiency purposes, the diagnostic value of student profiles might be great. Proficiency tests cannot be used diagnostically, as they are designed simply to establish a criterion for a particular population in the most efficient way possible, whereas diagnostic tests, intended to yield richer information, could actually be used (though less efficiently) as proficiency tests. The question was raised as to why there are so few diagnostic tests in existence: is this merely a practical problem, or are diagnostic tests theoretically impossible? Or do they simply not exist because people – teachers – do not want them and would not use them? Even if it is possible to gather relevant diagnostic information, what could one do with the information? Students' problems may well depend on something not related to the point being tested, but on the content of the text, and a host of variables within the specific context of the problem. A diagnostic test will not in any case tell one which of the weaknesses identified are crucial weaknesses, since all it can do is establish whether a subject knows or does not know something about a particular area (although in a sense the importance of a weakness has been decided in advance by the very inclusion of items in that area, in the test). The view was put forward that diagnostic testing may be at best pretentious – making claims that it is unlikely to be able to live up to – or at worst a pseudo-procedure, because diagnosis is impossible: problems are not predictable. Of course, textbooks and syllabuses are just as pretentious to the extent that they attempt to eradicate or anticipate problems.

It was felt, however, that there may be a danger of requiring too much of our diagnostic profiles: we may not be able to achieve perfect diagnosis, but gross diagnoses may be of use. The information that a particular student cannot understand lectures but needs to, may well be of greater value than information to the effect that the same person has achieved a score of 32 on EPTB. Desirably, our diagnoses would yield information not only on the product – for example, for comprehension, 'this student has failed to understand this lecture' – but more valuably, would yield information on the process whereby the product was (not) reached. Product items cannot be used diagnostically, since they do not tell one anything about how the individual did or did not get the product, whereas process items might be of great value. Thus profiles might be of value if they are the right sort of profile: composed of 'items' that give information about process which is relevant to pedagogic intervention. This would seem to be less related to the question of the specificity of a language test, than to the question of test content in terms, for example, of enabling skills.

One real and important problem with profiles is that people – admissions officers and the like – do not seem to be able to cope with them. It was reported that there is a regrettable tendency to reduce ELTS-produced profiles to an 'average score', from which, of course, all diagnostic information has been removed. However desirable diagnostic profiles might be for some language teachers, it is unlikely that they will be usable by lay people. If tutors or admissions officers have difficulty understanding a division of scores into Listening and Reading, how likely are they to want, or to have the time, to interpret a profile? But even if such people have the time and the inclination to interpret profiles, to what extent will they be able to do so? There would have to be some sort of prior determination that x course in y department in z institution requires a certain sort of profile, and the fact is that we simply do not have that sort of information: neither the admissions officer nor the applied linguist is able to say what profile is required by any department in any institution. Thus, at best, a vast amount of research is necessary before such criteria could be established.

Research Needed

As was reiterated throughout this discussion, there is clearly a need for a great deal of research in the general area of specific purpose language proficiency testing before one can begin to make claims about the validity of particular approaches or tests. It would be unfortunate if ELTS-type tests were introduced without any sort of validation. Empirical evidence, rather than construct validity, is urgently required on these and similar tests, since already admissions decisions have been taken about students. It is to be hoped that follow-ups will be done of students who have been admitted with ELTS scores (although it is unlikely to be possible to follow-up students who have been rejected because of their ELTS scores). It would, for example, be perfectly possible to get ELTS-type profiles of students who emerge successfully from their course of study, and, over a period of time, to gather information which would lead to a profile of 'successful' students. The ethical problem of admitting or rejecting students without such information remains.

It was generally agreed that it is crucially important to find out what is happening on a test as influential as the ELTS test. There is a clear need to know how such 'ESP' tests relate to existing tests, for practical as well as academic reasons. There is a clear need to know what the new tests are predicting, and what they are capable of predicting. There is a need to know what sort of diagnostic information can validly be provided, and whether it can be used by both applied linguists and lay people. There is a need to specify much closer the outcomes to which the test is to relate: both the academic and the linguistic/communicative. There is a need to analyse the communicative needs of students in this country, and the extent to which the

problems of native speakers are similar to or different from those of non-native speakers. It is clear that the design and implementation of a new test instrument requires an enormous amount of research, development, effort and resources, which it is easy to underestimate. The same need for research would exist for any test, but particularly for a test that appears to be an ESP test, that claims to be innovative, to be an improvement over other tests and that deals with the future of people. We need to know whether the claims made in the *Specifications* document are substantiated by the evidence. Nevertheless, it was agreed that the *Specifications* document is important, despite its unsubstantiated claims because it highlights the central problem of ESP proficiency testing: matching the demands of test design with those of the people taking the test and with those of the sponsoring and receiving institutions.

SECTION 3

BASIC CONCERNS IN TEST VALIDATION[1]

Adrian S Palmer, English Language Institute, University of Utah, USA
and Lyle F Bachman, University of Illinois, USA

Introduction

Test validity is a complex issue, and to address its many facets in any degree of detail in the space available is a considerable challenge.[2] To make this possible at all, we have had to assume that the reader has some degree of familiarity with traditional views of validity. Consequently, we will review only briefly the basic types of validity. We then look in somewhat more detail into the nature of construct validity – the type of validity which we are currently investigating. Finally, we present some of the general results we have obtained in a recently completed construct validation study.

Types of Validity

Investigations of test validity are, in general, investigations into the extent to which a test measures what it is supposed to measure. This is however, a very general definition of validity, and it is useful to distinguish among several different types of validity. We will distinguish among four here.

Face validity

The first, and in our opinion the least important, type of validity is 'face validity'. Face validity is the appearance of validity – the extent to which a test looks like it measures what it is supposed to, but without any empirical evidence that it does. There is no statistical measure of face validity, and there is no generally accepted procedure for determining that a test does or does not demonstrate face validity.

---

[1] Prepared for presentation at the RELC Regional Seminar on the Evaluation and Measurement of Language Competence and Performance, Singapore, April 21-25, 1980.

[2] We would like to express our deepest appreciation to the participants in the 1979 and 1980 colloquia on the construct validation of oral tests, held at the TESOL national conventions. These individuals, too numerous to name here, have contributed to every phase of the research described in this paper – from the original expression of a need for such research to its design, implementation, and interpretaiton.