Issues in the design of a large-scale test of English

Nesamalar Chitravelu, Khong Chooi Peng, Tan Soon Hock

The issues we came up with are the perennial ones and for those among the audience who have been working on tests, the feeling of déjà vu, the so-what's new, is inevitable. Why we are reporting our private experience of these very public perennials is in the hope that our unique combination of circumstances may have something of value to others.

Most standardized tests of language proficiency share the following characteristics:

a. they have been written for rather specific purposes (usually to predict adequacy of language command for future educational success where the medium of instruction is English)

b. they have been written for candidates who are usually not the nationals of the country which is trying to establish their English language proficiency through the given test and which is therefore only solely interested in the assessment aspects of the test; and

c. the tests are usually designed by people who are not educators of the test population. At best they would have taught them English as a second language.

By contrast, our test, the English Language Proficiency paper, is being designed to fulfil a multiplicity of purposes within the framework of the educational objectives of the country whose own citizens would form the test population. The test has been developed by people who are themselves not native speakers

Melbourne Papers in Language Testing 1992 Volume 1.2 pp. 45-71. The Language Testing Research Centre, The University of Melbourne.

and, perhaps most importantly, are primarily educators and not purely assessors. The distinction we draw here is that assessors give information which is crucial to the future education of the candidates; educators have the option of using the test itself as an agent to influence what and how the domain being tested (English language proficiency) is learnt. In that sense, the test is not merely an instrument of evaluation but also a pedagogic agent.

We hope that this unique combination of features may have helped us to stumble on to `insights' which may be of benefit to future test design, if not in determining what should be done, at least in highlighting the dangers of allowing amateurs to make important design decisions. We hope, though, that, like Dryden's Shadwell, we too may have `sometimes deviated into sense'. We hoped that being unlumbered with reputations to uphold, we may have more intellectual elbow room to think differently and even unwisely!

First, the paper will outline the aims of the ELP and the sociocultural and pedagogic framework within which it would function. Next, it will proceed to discuss some of the issues that we took into consideration in evolving our test design principles. The paper will then go on to discuss the structure the test finally took. It will conclude with a discussion of some of the adaptations that had to be made to the design of the test as a response to the findings of a pilot run.

Purpose of the test

The brief given us contained these specifications:

> `To prepare a national level criterion-referenced test to give information regarding the English Language Proficiency of those who, on completion of their Sixth Form enter the Malaysian job market or get into institutions of higher learning both in Malaysia and overseas'.

In interpreting this brief and determining what purposes the test should have, we took into consideration three factors:

a. the socio-economic ambience within which the test would operate;
b. our own understanding of the nature and acquisition of language proficiency;
c. our grassroots understanding of the teaching-learning situation in Malaysia.

The theoretical underpinnings of the test are made explicit in the discussion of the choices we made at each juncture in test design, and therefore need not concern us at this point in the description of the test. A brief outline of the Malaysian situation is, however, probably essential.

The socio-economic needs

With the development of a modern economy with its corollaries of rapid industrialization, international trade, and turn-key systems involving foreign participation in local projects, the needs for English have been burgeoning. But there has not been a corresponding increase in the number of Malaysians equipped with the language wherewithal to meet these multifarious needs. There is a tremendous gap, particularly in meeting the needs of the private sector and the middle and upper echelons of government service. In surveys conducted in 1980 (UMSEP Research Reports 1980/81) and 1985 (Chitravelu, 1985), it was found that English was still a vital constituent of success in careers, particularly in the professions, in business, diplomacy and in all those government jobs which involve contact with the private sector and the outside world as well as in areas where technical expertise can primarily, if not only, be assessed through the medium of English.

A further point that perhaps needs mentioning here is that there are many people who need information about the English Language proficiency of their potential students/employees but there is no means by which such information can be obtained. This is because our national exams, probably like national exams

everywhere else in the world, serve a different purpose. Their primary objective is to provide a means of comparing performances. They are thus, of necessity, norm-referenced and therefore capable only of giving information about relative performance. The users of the test - the institutions of higher learning and the employers - need a reliable guide of absolute ability.

An instrument to do this does not currently exist in Malaysia. Foreign universities therefore rely on standardized tests - TOEFL and ELTS mainly. The local consumers have no alternative but to use results on the available norm-referenced tests or, if they are universities, to use exemption and placement tests of their own.

The need for a fixed standard is imperative not only for the synchronic needs outlined above but also for a diachronic description of Malaysian English. There is much talk today about falling standards, but there is no way of establishing what we have fallen from, or where we have actually got to, in the continuum of proficiency since the tests Malaysians sit for reflect the norms of the day.

Educational needs

The story of Malaysian education follows the stereotype of most newly-emergent nations. Almost all courses in tertiary institutions are taught through the medium of the national language. However, the primary tool of independent research, both at undergraduate and postgraduate levels, is still English since English still remains the language of the textbooks and the journals that keep the scholar updated.

Despite the existence of six universities in Malaysia and the escalating costs of foreign education, the need to go overseas for an education still remains a big one. In 1983, the profile of higher education looked like this:

Malaysian students in overseas and local institutions from 1970 to 1983

	1970	1975	1980	1983
Overseas institutions	n.a.	31,500	40,000	58,000
Local institutions	13,324	31,529	38,125	55,072

The numbers have somewhat diminished in more recent years, but the number of Malaysians going overseas and the corresponding need for English are by no means small.

What to test and how to test: some issues

Our decisions on what to test were intuitive. Alderson (1988) explains that the first aim of test specification is:

> `...to provide a statement of what the test is supposed to test: in short, what its construct is, what view it takes of language proficiency generally, or with reference to particular components of that proficiency, what theory of reading, or communicative competence, or grammar or whatever, is being proposed or adopted by the test designers'.

We did not articulate explicit answers to the kinds of questions Alderson raises. Our system of working consisted of arriving at an initial consensus on certain polarities which seem to be perenially present in the development of any test, and in the process of arriving at and rationalising a consensus on each of these polarities and the questions they each spawned. We worked out an operational definition of our stances on the questions raised by Alderson. Once we worked out the philosophical and theoretical stances that the test would express, we omitted the stage of detailed specification based on needs analysis. The descriptions of skills such as the one by Munby (1978) seemed of little practical use in this context. On the one hand, they were so detailed as to be daunting. On the other hand, the more vital question of the dynamic combinations and

permutations of these skills in real communication remain largely unexplored.
What follows is a selective summary of our decisions on some of the broader issues. These have been specifically chosen for reporting because we felt that some of our decisions and our reasons for making them may be of interest to others attempting to design their own tests.

Achievement vs proficiency

We share the view of Read (personal communication) that the demarcations between achievement and proficiency are not as well-defined as was once thought. We believe that every test is on a continuum, one end of which is total explicitness of syllabus (the achievement end) and the other of total implicitness (the proficiency end). We chose the proficiency end of the continuum for two main reasons.

a. The first is theoretical. We feel that the more explicit we want a syllabus to be the more certainly and accurately we must know the nature of what it is we are testing. Since the issue of what constitutes proficiency in a language is still very much a moot point (Oller,1983; Alderson, 1988; Davies, 1988) we felt that a proficiency test that relies on our intuitions was a better bet;

b. The second reason relates to the educational framework within which the test is to operate. In Malaysia, we have three public examinations, one at the end of primary school, another at the end of lower secondary and one at the end of Form V. Each of these examinations tests degree of success in mastering the content of the syllabus to which it is pegged. These tests are graded, each syllabus carrying on from where the earlier one had left off. Now, we felt that if we produced another achievement test, the automatic assumption would be that this was next in the series. The corollaries of such an assumption would be that this syllabus must begin where the Form V syllabus left off and that the targets it

sets, like the targets of its predecessors must be achievable by the average Malaysian Sixth Former. Our brief was to design a test that would reflect the proficiency targets to be met by people wishing to operate at the upper and middle echelons of Malaysian society and given the levels the `average Malaysian' achieves at the end of Form V, such a target did not seem a viable objective to follow in the series of targets set by the previous three national examinations. We still feel, however, that a criterion-referenced test is viable in the Malaysian context since there are Malaysians (especially from the urban areas and from middle class homes) who can be groomed to achieve the necessary level of proficiency.

Competence vs performance

`Competence' is generally defined as knowledge about a language and its rules stored in an individual's mind whereas `performance' is the integration or orchestration of this knowledge to convey personal meaning.

Following the developments in communicative language teaching, there is increasing concern that proficiency testing should move away from the testing of knowledge of language and its rules, to focus on assessment of communicative language skills. In advancing this view, several arguments have been put forward. Since the ultimate aim of all language learning and teaching is to `get things done with words', it is argued that task-based performance tests have greater construct and face validity because they reflect this dynamic, purposeful use of language. Furthermore, it is contended, learners' purposes for learning the language are acknowledged as varied and having legitimate claims on decisions regarding what is to be taught and learnt. This points to the need for test content to reflect the kinds of situations examinees will find themselves in in real-life. Assessment of proficiency, therefore, should not be confined to linguistic accuracy but should also incorporate other criteria necessary for the assessment of effective communication of ideas in specific language-use situations. To this end, `direct' tests of English proficiency for communicative purposes have been

developed. These include those administered by Royal Society of Arts Examination Board (The Communicative Use of English as a Foreign Language), and the Association of Recognised English Language Schools Examination Trust (ARELS Oral Examinations) which assess general communicative skills. Similar tests in English for academic purposes include those developed by The English Language Testing Service (The ELTS Test), and The Associated Examination Board (Test in English for Educational Purposes).

Nevertheless, despite what seems to be a general accord on the value of assessing communicative performance, some discordance remains over what it is that a test really ought to consist of. There is, on one hand, the above view that assessment of performance is best done through `direct' tests. On the other hand, there are those who hold that competence-oriented tests that correlate highly with actual performance are both possible and even desirable. Such a view is clearly shared by tests such as TOEFL and TOEIC.

While recognising the value of the principles underlying the above tests, we agree with another view - that it may not be worthwhile or even possible to distinguish between testing for `competence' per se versus testing for `performance'. Rea (1985:21) argues that:

> `all linguistic behaviour, whether it involves phoneme recognition, assigning a meaning to a single lexical item or, at the other end of the spectrum, interpreting stretches of discourse, constitutes instances of performance'.

She does, however, make the distinction between performance that is `non-communicative' (meaning- independent; concerned with grammatical accuracy or `well-formedness'), and performance that is `communicative' (meaning-dependent; the `successful performance of which reflects an integration between grammatical, sociolinguistic, strategic and other competencies' (Rea, 1985:24). This view, that performance underlies all

language behaviour, is identical to that of Widdowson (1978:3) who defines `usage' and `use' as both aspects of performance.

Other considerations concerned with the large-scale nature of the test to be designed have to also be taken into account. `Direct' performance testing will entail not only large sums of money but will require a complex system of administration in order to ensure its successful implementation on a national scale. Within the context of ELP, apart from the considerations of expense and administrative feasibility, the lack of trained personnel and the large scattered population also precluded sole dependence on `direct' measures.

Furthermore, there is the undeniable fact that performance-based tests yield fairly limited samples of language from the examinees, raising the question about the validity of inferring proficiency from these samples. In addition, the view among the designers of ELP is that `non-communicative' tests could be categorised along a continuum from `direct' to `indirect' and that it was possible to select from along the continuum those that are not totally focussed on usage but are meaning-dependent which, together with the `communicative' tests, will provide a comprehensive profile of the examinee's proficiency. The decision was made, therefore, to include both `communicative' and `non-communicative' tests in the ELP. What these consist of is described in the section on `Integrative vs discrete skills' below.

Some skills vs all four macro skills

The issue whether ELP should test some or all four skills was decided after taking into account the following factors.

Tests such as TOEFL and TOEIC have demonstrated that their validity in predicting overall language proficiency based on performance on one or two macro skills is high. Their findings would justify not having to include all four skills. For a large-scale population (potentially approximately 50,000), considerations of test proficiency and economy are important. Furthermore, the testing of the listening and, in particular, the

speaking skill entail administration complexities and greater expense.

However, the decision was made to include all four macro skills. This decision hinges on several main considerations:

1. Depending on his educational background and other factors, an examinee's proficiency in each of the four macro skills may be at a different level. In the Malaysian context, for example, it is possible for an examinee, particularly from the rural areas, to be more proficient in reading than speaking.

2. Different prospective employers, training centres and institutions of higher learning make different demands on language proficiency. For example, Malaysian universities require a high reading proficiency whereas certain professions place greater emphasis on speaking. In order to meet the needs of these `consumers', profile reporting on each macro-skill is necessary.

3. It is universally demonstrated that teachers will teach toward public exams. An exclusion of any macro skill would inevitably lead to the neglect of that skill. A recent development in the Malaysian school curriculum is to include the teaching of all four skills. It was felt that this trend should be upheld.

ESP vs general proficiency

Although in the language teaching scene the trend seems to be towards greater and greater specialisation, and a corollary development would seem to be the development of tests which test language skills using subject-specific content, our team opted for general proficiency.

The primary reason for our decision to construct a proficiency test was a pragmatic one. Since the anticipated candidates for the tests, and ultimately, the users of the test had diverse needs, the

number of versions of the test we would need to construct would render the whole enterprise untenable. Firstly, it would not be cost-effective; and secondly, because it would reopen the contentious issue of the comparability of the multiple versions of the test.

Our experience in teaching and designing ESP courses at the university also persuaded us that expediency need not be the sole rationale for choosing general proficiency over ESP. We found in our teaching that in many instances the students could not even begin to access the ESP materials. This was not because they did not know the language of their specialism, but because they did not have a threshold of English on which to peg their new learning. Some of the time, when students did have the language, they did not have some other constituent demand of the communicative situation viz the assumed previous knowledge, the intelligence etc. These elusive `other factors' are as much components of the communicative competence that we seek to measure in our tests and yet we do not test them. So why the need to test specialised knowledge? As Davies (1988) concludes of ELTS:

> `ESP is a valid construct but ...the variability is more varied than the simple subject-specific structure permits, involving language skills and no doubt other variables not tested in ELTS'.

Post facto, years after our decision, we have found research literature to vindicate our choice, and to show that there does not seem to be any compelling proof of the superiority of ESP tests over general tests. Davies (1988), for example, asserts that the validation study that his team did on the ELTS test showed that the general component G1 contributes 0.83 to the overall band score, while some of the specialised components contribute only as much as 0.5.

One test or several tests

The issue of ESP or general proficiency also, in part, collapses into the issue of one test or several tests.

1. One option is to have one test, comprising both a general section which tests general language proficiency and a modular section which examines subject-specific skills. This is the option that T.E.E.P. and ELTS take. Both have a general test complemented by a modular section. T.E.E.P. has a general section and two modular sections, one for Arts and Social Sciences and the other for the sciences. ELTS has a general section and six modular versions.

2. Another alternative would be to have a different test for different expected proficiency targets. This is the option John Read's ELI test and the CUEFL test took. Our own decision, like TOEFL's, was to have a single test to express several expected criterion levels. The rationale for this will be explained in the section on the establishment of criterion level and the reporting of scores.

Integrative or discrete skills

The question of how performance (whether communicative or non-communicative) ought to be tested, remains the subject of much debate. The predominant view is that `communicative' aspects of performance can best or should be tested using `global' or `integrated' means, whereas `non-communicative' or `competence-oriented' aspects are efficiently tested through `atomistic' or `discrete' means.

Discrete point tests, based on the premise that language, being a linguistic phenomenon is most efficiently tested through its linguistic constitutents, includes measures of phonology, points of grammar, and vocabulary. Tests designed along `discrete point' lines include standardised tests such as the MLA Cooperative Tests, the Graduate Record Test and TOEFL. Oller (1976:156) in strongly advocating a non-discrete approach to testing, lists cloze, dictation, translation, essay, and the oral interview as examples of integrative measures.

Davies (1978:215) summarises the main points commonly raised to argue for an integrative rather than a discrete approach. Firstly, language is not made up of unrelated bits; it forms a whole. The bits must therefore be integrated and tested in combination with one another. Secondly, there is always a communicative purpose in language learning, and it is this communicative ability which must be tested. Thirdly, discrete point tests are too general to be useful and should be replaced by specific tests (reflecting special purpose in language teaching).

Davies' own position, however, is that:

> `the most satisfactory view of language testing, and the most useful kinds of language tests are a combination of these two views, the analytical and the integrative ... Test reliability is increased by adding to the stock of discrete items in a test: the smaller the bits and the more of these there are, the higher the potential reliability. Validity, however, is increased by making the test truer to life, in this case more like language in use'.
> (1978:149)

He further advocates that:

> `it makes sense to see integrative and discrete point tests as forming a continuum'.

The approach we adopted in ELP is similar to that of Davies' above in that we included items which fall at different junctures of the continuum.

A criterion-referenced test or norm-referenced test

Many of the reasons as to why we chose a criterion-referenced test have already been discussed in the course of the discussion on other issues. All that needs to be done at this juncture, therefore, is to summarise the main issues already raised, and to discuss those still to be considered. These, roughly, are the points already raised:

1. Our brief already specified a criterion referenced test and we had no option but to accept this as de facto.
2. There is an urgent need for a test with unchanging standards both within Malaysian society and for purposes of international recognition.

Two more issues, each rather important from our point of view, still remain to be discussed. One concerns the definition of the term criterion-referenced and the other concerns the way in which our understanding of criterion can be operationalized. The issues are in some ways intertwined.

Generally, standardised tests attempt to predict performance on future tasks. The existing standardised tests that we know of operationalize the notion of criterion by establishing score levels which the test developers and administrators claim are the mandatory minimal thresholds necessary for effective performance of these tasks. This level of confident prediction, we felt, presupposed proper needs analyses and empirical validation procedures. Again, the pragmatics of our situation did not provide the necessary infrastructure. Our test needed to be produced in a hurry and since our target population spanned a very wide range of needs (local university vs foreign university, varying subject needs, jobs with different demands, etc.), it precluded, in any practical sense worth considering, all possibilities of empirical validation or needs analyses. How then, can we consider our test `criterion-referenced'? We are not aware whether our definition would be acceptable to the gurus of testing, although post facto we have discovered that others too have ventured similar definitions (Popham, 1978). We decided to make the standards of achievement postulated by the test immovable, but within the immovable yardstick, not to fix any point as that which needs to be reached by people in order to do X, Y or Z effectively. If our contention that the common factor in proficiency (whatever that is) is the most significant in accounting for variation, then the test should allow people with differing needs to each find their niche in the continuum that the test provides. This is a decision we stumbled on but having arrived, we rather think that, at least for our needs, this may be

the most economical and elegant solution. Since Malaysian English itself is in a constant state of flux and its future not entirely predictable (there are many attempts to arrest the fall in standards) a fixed standard that can be flexibly used will take care of the various needs. For empirical evidence each (not just each kind of) test user can fix the standard of performance his employee/student needs to possess. Besides, we, like Davies (1988), believe that proficiency is context-determined and we also additionally believe that criteria are fixed not strictly on what is absolutely necessary for the performance of a task but also by what, given the constraints of context, it is reasonable to expect. An employer in England would probably expect a higher standard of English from his secretary than an employer in Malaysia. Likewise, an employer in Malaysia would have expected a higher standard of English from his secretary in the 1950s and 60s than today. Even today, a secretary at the front desk of the Ministry of Foreign Affairs is likely to need more English than her counterpart in the Social Welfare Department. A test such as ours would allow for considerations such as these, as the cut-offs are market-determined.

Reporting of results

The description of our stand on criterion-referencing above may seem to suggest that we absolved ourselves completely from all responsibility to indicate standards. But this is not true. We do not, ourselves, claim the right to indicate which level of proficiency is the required minimum for performance of any one task. But what we do, is to use a style of reporting that is likely to be of maximum help to potential users. We provide information on overall performance as well as on performance on each of the skills. This is to take account of the fact that different users have need for information on different skills. In establishing rungs or bands on the continuum of proficiency, we have given precedence to comparability with existing tests. Thus, our test, like the ELTS, consists of 9 bands and each of these bands is described in roughly the same terms as existing standardized tests. Work is now being done to establish concurrent validity of our test both with TOEFL and with ELTS.

Operationalizing the decisons in test content

The listening skill

The listening component comprises three separate sections:

a. 10 MCQ items each in sentence form test the understanding of stress and intonation to get at the literal meaning of an utterance, to infer implicit meaning, and to understand the illocutionary force or intent behind an utterance.
b. 20 short exchanges between two participants constitute the listening input. Each exchange is asked orally by a third speaker. The inclusion of this type of item is based on the fact that there are other sub-skills which can best be tested to discrete form within an interactive context. The examinee is tested on his ability to distinguish meaning, to recognise appropriacy, to make inferences, to understand degree, purpose, sequence, etc.
c. A fairly long conversation between two speakers on one or more topics provides a general listening purpose test. In order to understand the questions some of which are MCQ, the examinee is required to listen for specific information, to obtain the gist of what is said, to distinguish fact from opinion, to extract salient points, to summarise argument, and to judge the speaker's attitude.

The oral skill

This is done in two parts:

1. The examinee is given 25 MCQ questions requiring him to identify appropriate forms and functions within the context of an exchange between two speakers. The successful completion of each item requires an integration between linguistic and communicative competencies.
2. The examinee is tested using a tape-recorder. The examinee hears and responds to taped stimuli. His

responses are also recorded on tape. The whole test takes about 20 minutes.

The questions which the examinee is required to respond to are in 4 stages. (These stages are not overtly labelled as such on tape).

Stage 1 : The examinee is asked to give some basic information (e.g. index number). His performance at this stage is not graded;

Stage 2 : This begins the actual test. The examinee is asked questions about himself (e.g. hobbies);

Stage 3 : At this stage the examinee is asked to look at some visual material (e.g. pictures) and to speak about the material for 1 - 2 minutes.

Stage 4 : The examinee is presented with a specific problem (e.g. problems of drug abuse) and is asked to give his opinion based on the input. He is given about 2 minutes to do so.

The reading skill

The Reading component comprises four main sections:

1. 15 MCQ items based on one long text which tests the global comprehension, reading strategies and locational skills that we felt were some of the marks of an efficient reader. The time allocated for the task ensures that only effective use of strategies will ensure successful completion of the tasks.

2. True/False/Can't Say items which test what we called `discourse features'. They were based on several short texts. The texts for this section of the test were selected from a wide variety of sources (newspapers, research reports) and highlighted for consideration issues primarily related to the pragmatic meaning of the text and the rhetorical functions (exemplification, corroboration, etc.) that different portions of text (e.g. reports of findings, provision of statistics) perform.

Section 1 and 2 mentioned above attempt not only to provide a fair sample of the skills an efficient reader needs to possess, they also attempt to increase the visibility of what the test constructors

deem as important skills for students to learn. Test-wiseness, it is hoped, would in this case be not a beating of the system but a way of using the system as a catalyst to stimulate an educative process.

3. The third element in the Reading test is a cloze test. The text for this is a continuation of the long text mentioned in 1. above. The section is included for two reasons: one, as a test of the candidates' ability to literally comprehend a text like the one used to test higher order skills and two, as an antidote to the prescriptive nature of the two sections already mentioned.

4. The fourth test is a 15-item multiple choice test of vocabulary. This, of the four tests, is the closest to the usage end of the performance continuum and was included partly to increase the reliability of the test (see Davies 1988 quoted above) and partly to accede to the generally held view that vocabulary is an important constituent of Reading.

The writing skill

As with the other skills, writing is tested through items which fall along different points of the performance continuum.

Proficiency in writing deals with not only how well a candidate is able to express himself clearly, concisely and accurately, but also how well he is able to organise his thoughts. Accordingly, the writing section includes the following:

1. A test of structure
This section comprising 20 MCQ items was included for three main reasons. Grammar is accepted by many scholars to be the common denominator underlying all language use. It is also a well-tried and reliable predictor of general proficiency. The format adopted for the testing of structure here is one that has been used for several years in the TOEFL test. A knowledge of the problems faced by Malay-medium candidates (and in general, second/foreign learners of English) in their mastery of English structure informs the choice of items as there is yet no reliable

knowledge of the relative functional importance of the various structures in the language.

2. Paragraph Organisation
This section consists of five gapped paragraphs, each of which develops its central idea in different ways. The gap is located in different strategic positions in the five paragraphs and tests the candidate's ability, firstly, to recognise what rhetorical function the sentence in the gap should fulfil and then to find the sentence that best realises this function. The ultimate aim of this section, therefore, is to gauge, in a fairly discrete way, the candidate's mastery of specific writing techniques, especially those involving the perception and production of coherent paragraphs. The information obtained from this section, it is hoped, would complement the information from the essays in the performance section.

3. Essay
This section consists of two writing tasks: a free composition and an essay based on information provided in non-linear texts. The writing ability of the candidate is reflected in his ability to orchestrate all the skills necessary to produce the written form.

Structure of the test
Pursuant to the theoretical considerations for test design, the table in the Appendix outlines the test structures:

Structure of Original Test
The test was divided into five parts, labelled as Papers 1 - 5. Each paper had varying numbers of sections accompanied by their individual sets of items. The entire test took 5 hours 20 minutes.

It should be noted that the contents of each paper was limited to one macro skill. Its varied composition was instead affected by factors such as

a. time - it was felt that each paper should not be too long as to be mentally taxing.

b. manageability and feasibility of administration -
this took into consideration the way in which the items are to be answered. As far as possible those that needed computer sheets were grouped together, while those that needed answer sheets had to be collated separately. Obviously, the taped output section needed a separate administration altogether.

As a consequence of the first pilotting of the Test, a number of issues were raised. These will be discussed via a description of changes that were made to the test, as reflected in Table 2 in the Appendix.

Structure of Revised Test

1. The test was shortened to four papers, totalling 4 hours 15 minutes.
2. Sections within papers were dropped to effect a shorter test. The decision to make such adjustments were based on the following considerations:

A. Statistical

i. unmarked stress (Paper 1, Section A): It was discovered that the two tasks of unmarked stress and short exchanges which test listening comprehension are both highly correlated at 0.79. This may imply redundancy in testing, and it was felt that since short exchanges enabled a wider coverage of subskills than unmarked stress, the latter was dropped.

ii. Listening to a mini lecture (Paper 1 Section C) and a long conversation (Section D): Correlation of the two with the other subskills of listening is high, though they were sufficiently different from each other (0.57). It was felt that only one for any version of the test would be retained.

iii. In the writing section, the intercorrelation between subskills was found to be substantially low, thereby justifying an inclusion of all of them in the test. This would ensure a good balance between the competence and performance tasks.

B. Balance in tasks

i. the writing subskills reflected this balance.

ii. the test of speaking involved only two sections - discrete feature oral and the taped output. A rather high correlation of 0.74 might tempt a decision to delete one of the two without much loss to the efficiency of the instrument as a test of speaking. However, from our experience at the University of Malaya we have recognised the necessity to maintain a battery of indirect discrete point items to counterbalance the problems of unreliability inherent in performance tests like the taped output.

C. Difficulty Level vs Format of test

The correlation between reading subskills is not very high. There seems to be sufficient difference between them to warrant the presence of all of them on a test of reading. However, since this skill showed poor performance in relation to the rest of the skills, one possibility in reducing difficulty level was a slight reduction in test length. We therefore looked for a subsection that might appear to be similar in intent to another. The textual features section seemed a possibility. Even its format is a variant of the cloze procedure used in the MCQ section in reading an extensive text. Hence, it was decided that the textual features section would be dropped.

D. Content

There was a change made to the content of the performance tasks in writing. Originally, the writing tasks were related to the reading text. It was felt that the undue influence of reading should be eliminated. The two tasks would now be based, one, on non-linear information, and the other, a composition on another topic.

Conclusion

Whether the assumptions and theories on which our test is based are valid or whether our decisions on issues like market-determination of cut off are sensible only time can tell. In the meantime, however, work on statistical validation is on-going.

References

AEB. 1984. *Test in English for Educational Purpose.* The Associated Examining Board.

Alderson, J.C. 1988. New Procedures for validating proficiency tests of ESP? Theory and Practice. *Language Testing*, Vol. 5, No. 2, pp. 220-232.

Chitravelu, N. 1985. *Status and Role of English in Malaysia: A Research Report.* (unpublished).

Davies, A. 1978. Language Testing: Survey Article. *Language Teaching and Linguistics Abstracts*, Vol. 2 3/4: part 1, pp. 145-159; part II, pp. 215-231.

Davies, A. 1988. Operationalizing uncertainity in language testing: an argument in favour of Content Validity. *Language Testing*, Vol. 5, No. 1, pp. 32-48.

Department of Immigrant and Ethnic Affairs. 1984. *Australian Second Language Proficiency Ratings,* Australian Government Publishing Service, Canberra.

Goon, C. 1980. Profile of Communication Needs of a Malay-Medium Arts Graduate, *University of Malaya Spoken English Project Research Report*, Language Centre, University of Malaya.

Khong, C.P. 1980. Profile of Communication Needs of an Administrative Officer in a Statutory Body. *University of Malaya Spoken English Project Research Report*, Language Centre, University of Malaya.

Oller, J.W. Jr. 1983. *Issues in Language Testing Research.* Newbury House, Roley, Massachusetts.

Popham, W.J. 1978. *Criterion-Referenced Measurement*, Prentice-Hall, Englewood Cliffs, New Jersey.

Rea, P.M. 1985. Language Testing and the Communicative Language Teaching Curriculum. *New Directions in Language Testing*, Lee et al (eds), Pergamon Press, pp. 15-32.

Widdowson, H.G. 1978. *Teaching Language as Communication*, Oxford University Press: London.

APPENDIX

Table 1 : Structure of Original Test

Paper	Section/ Task	Content	No. of Items	Time Allocated	Scoring	Score	Total Time
1	A	Unmarked Stress	15	15 mins	Objective	15	1hr 15m
	B	Marked Stress	10	10 mins	Objective	10	
	C	Short Exchanges	20	20 mins	Objective	20	
	D	Listening to a long conversation	10	20 mins	Objective	10	
	E	Listening to a mini-lecture	13	10 mins	Objective	15	
2	A	Discrete	25	15 mins	Objective	25	45m
	B	Features, Oral Vocabulary	30	15 mins	Objective	30	
	C	Paragraph Organisation	5	15 mins	Objective	5	
3	A	Structure	20	15 mins	Objective	20	1hr
	B	Discourse Features	40	30 mins	Objective	40	
	C	Textual Features	10	15 mins	Subjective	20	
4	A	Reading an extensive text	30	1 hr	Objective	30	2hr
	B	Essay writing	2	1 hr	Subjective	15	
5		Taped Output		- about 20 mins per student	Subjective	15	20m
		Total				270	5hr 20m

APPENDIX

Table 2 : Structure of Revised Test

Paper	Section	Content	Score	Scoring	Time in mins	Total Time
1	A	marked stress	10		10	
	B	short exchanges	20	objective	20	1hr
	C	listening to a long conversation	10		15	
	D	discrete features: oral	20		15	
2	A	structure	20	objective	15	
	B	essays	30	subjective	75	1hr 30m
3	A	vocabulary	15		15	
	B	discourse features	20	objective	20	1hr 30m
	C	reading an extensive text	20		40	
	D	paragraph organisation	5		15	
4		taped output	30	subjective	15m	
		Total	**200**			**4hr 15m**

APPENDIX

Table 3 : Summary of Adjustments to Review Test

Skills	Content	Original Test		Revised Test	
		No. of Items	Time in mins.	No. of Items	Time in mins.
Listening	marked stress	10	10	10	10
	unmarked stress	15	15	0	0
	short exchanges	20	20	20	20
	listening to a long conversation	10	20	10	15
	listening to a mini lecture	13	10	0	0
Speaking	discourse features, oral	25	15	20	15
	taped output	-	20	-	15
Reading	vocabulary	30	15	15	15
	discourse features	40	30	20	20
	reading an extensive text	30	60	20	40
	textual features	10	15	0	0
Writing	structure	20	15	20	15
	paragraph organisation	5	15	5	15
	essays	2	60	2	75
	Total		5hrs. 20mins.		4hrs. 20mins.