Complimentary Copy

WORKING PAPERS

IN LINGUISTICS

AND

LANGUAGE TEACHING

SPECIAL ISSUE ON

LANGUAGE TESTING

Language Centre
University of Hong Kong

NO. 4 June 1981

CONTENTS

ARTICLES

Acknowledgements

The Editors are greatly indebted to the assistance rendered by Ms. Amy Lam (printing and coordination) and Miss Connie Kwok (typing), as well as to the University Printing Unit.

POLICY

Working Papers in Linguistics and Language Teaching will publish work in a number of areas, including: general linguistics, language teaching methodology, information about and evaluation of teaching materials, language curriculum development, language testing, educational technology, language and language teaching surveys, language planning, bilingual education. Articles on Chinese and Chinese language teaching may be published in Chinese.

Working Papers is aimed primarily at Hong Kong, and as such, is intended to be informal in character, but it is hoped that it may also be of interest to specialists in other parts of the world, especially in Asia. There are two issues per year.

SUBSCRIPTION

The 1981 issues will cost a nominal HK$5 each or HK$10 for the year (US$2.50 overseas, including postage for two issues). If you are interested in receiving them, please fill in the accompanying slip and return it as directed, or write directly to the Editors.

CONTRIBUTIONS

Contributions in the form of articles, reports, in-depth reviews or simply comments on earlier articles are welcome. They should conform to the style sheet on p. iii and should be addressed to:

The Editors,
Working Papers in Linguistics and Language Teaching,
c/o The Language Centre,
University of Hong Kong,
Pokfulam Road,
Hong Kong.

# STYLE SHEET
for
Working Papers in Linguistics and Language Teaching

1. Typescripts should be double spaced and use only one side of each sheet of paper.

2. Section headings will be italicised at the time of printing. Please indicate all such subheadings clearly by single underlining.

3. Diagrams should be on a separate sheet.

4. Footnotes should not be used. All notes should be in a separate section at the end of the article. Each note should be indicated by a superscript numeral in the text.

5. References should be on a separate sheet (or sheets) at the end of the paper, following the notes. It should be headed 'References'.

6. Journal articles should be referenced in the following way:

   Oller, J.W. and Streigg, V. 1975. Dictation: A test of grammar-based expectancies *English Language Teaching Journal* 30(1):25-36.

7. Books and pamphlets should be referenced in the following way:

   Bullock, A. et al. 1975. *A Language for Life.* HMSO, London.

8. Articles in books should be referenced in the following way:

   Kvan, E. 1969. Problems of bilingual milieu in Hong Kong: Strain of the two-language system. In T.C. Jarvie and J. Agassi (eds.) *Hong Kong: A Society in Transition,* pp. 327-343.

Contents from Working Papers in Language and Language Teaching - Nos. 1, 2 and 3

No.1 - May, 1979

No. 2 - December, 1979

No. 3 - July, 1980

In the last few years, there has been a growing awareness in Hong Kong of new developments in the field of language testing. This interest was triggered off by changes brought about in language teaching at both primary and secondary levels. To enable the new syllabus to function in the desired spirit, tests have to be designed whose objectives are in line with the new curriculum. Already reports have been published on different attempts made at translating concepts such as communicative competence and functional proficiency into testable forms.

One of the main interests of the Language Centre of the University of Hong Kong is in language testing with particular reference to the measurement of language-use for academic purposes. Research is still in progress to establish profiles of intended language behaviour, and different attempts have been made to incorporate these profiles into tests.

Another aspect of the Language Centre's activity that is related to language testing is the training of test researchers. Among the many units offered in the M.A. course in Language Studies, the Language Testing and Evaluation unit helps to stimulate and sustain interest in the area. A number of research papers have been published as a result of staff and student collaborative ventures.

The papers presented in this issue, though they do not necessarily represent any standard policy or view, are all on topics which are of interest to its staff members. It is anticipated that they will serve to stimulate discussion on language testing, be it practical or theoretical in nature. All views and comments received will be included in the next issue of Working Papers.

COMMUNICATIVE TESTING AS AN OPTIMISTIC ACTIVITY

Graham Low
Language Centre
University of Hong Kong

Introduction

The popularity of use-of-language testing seems to have been growing dramatically, in recent years, particularly where second language learners are involved; and the term 'communicative testing' appears to have become a battle cry for all that is 'progressive'. By espousing communicative testing, the impression is given that one can throw off at a stroke the shackles of test irrelevance and artificiality, and any need for statistical examination or treatment of test results. The present article is an attempt to suggest that communicative testing is not - so far at any rate - an instant panacea, and that if an appropriate metaphor is required, that of 'uncharted minefield' might be more suitable.

The difficulty of defining the field

It is sometimes assumed that the term 'communicative test' is by itself sufficient explanation of what is intended. But since practically every test may be considered an attempt, however crude, to simulate a communicative encounter between tester and testee, and since language may conceivably be used for purposes other than communication, as Canale and Swain (1980) rightly point out, the term 'use-of-language test' is perhaps preferable to 'communicative test', and will therefore be used for the rest of this paper. The term 'pragmatic test', as used by Oller (1979) is not used here for several reasons. Firstly, Oller's own definition is hard to understand and does not correspond to the usual idea of 'pragmatics'. Secondly, there is much disagreement among linguists about what the 'usual' idea and scope of pragmatics is anyway (eg. Hudson 1975 on whether questions relate to pragmatics or semantics). Thirdly, pragmatics, by definition, is generally held to exclude matters that are primarily syntactic or semantic, and although a test which asked testees to discover and rectify syntactic errors could be justified as a use-of-language test (since it reflects what many writers do both during and after composing a text), some people might object to it being called a pragmatic test.

In order to examine how, when and why language is used, it is obviously important to establish the physical and physiological conditions and limits within which people function most efficiently and around which they, not unnaturally, tend to organise social activities and language encounters. We may, for short, describe these as 'preferred human (linguistic) operating conditions'. It is presumably to such preferred human operating conditions that Oller (1979:38) is referring when he requires that language in a test must be processed under 'normal contextual constraints for that language' before the test can be called pragmatic. A simple illustration of the above might be the observation that most people are happier and function best, and organise themselves accordingly, when only one person is talking at a time and when this one person is doing so without undue background noise or signal distorti It is, of course, immediately apparent that research data on this topic cannot be applied uncritically to language-use tests. Consider the much-quoted case of airline pilots, who are required to perform flawlessly in situations involving gross signal distortion and, in many cases, high ambient noise (Sumby 1960). What is important is that the tester is able to compare the specific needs of the population being tested with what we have called 'preferred human (language) operating conditions'. Moreover, for his[1] test results to be at all interesting, he must be able to state any divergences in a clear and explicit way.

To sum up so far, it seems important to be able to test syntax, for example, where this forms part of a language-related activity that the testee is required to perform, and to see this as a language-use test. More generally, it is not enough for a text, or even a test question, to have meaning; it should have one or more purposes, and the tester must be able to justify these purposes for each intended testee. It will be suggested later that the argument can in fact be extended beyond this point, and that a strong case can be made for making not just individual test questions but a whole test or series of questions purposeful.

Problems of competence and performance

There seems to be no general agreement about precisely what is meant by the terms competence and performance when applied to language and language-related activities (v. the different approaches taken by Canale and Swain 1980, Valian 1979, and Wiemann and Backlund 1980). Thus there is no one model of language use on which one can, with complete confidence, base a test. Indeed, if one adopts a narrow view of both competence and performance, restricting competence to abstract knowledge about a language and performance to physiological phenomena such as coughs, undesired false starts and certain slips of the tongue, it is quite possible to end up with a range of activities concerned with evaluating and using one's knowledge of a language which cannot be called either competence or performance. One recent way out of this sort of problem is to pretend it does not exist and claim that tests of language-use do not need to be based on a-priori views of what language ability entails. (For a discussion of this point, see Lee 1981). I would argue that this is an ostrichlike position if taken to extremes, and that, even if one does not affiliate to a particular model of competence-performance, a clear theoretical stance is nevertheless required, even if this is simply to state that a broad view of language competence is being taken, and that the assumption is being made that a test of performance in a context which is relevant for the testee will give a result which reflects underlying competence to a large degree.

Notwithstanding the problem of ever defining competence and performance in a satisfactory way, there is the additional problem of deciding which of the two one ought to be in fact testing. On the one hand, one might argue that coughs and occasional false starts have nothing to do with someone's language proficiency, but on the other hand, in many cases (e.g. testing future salesman, interviewers, actors and numerous others) it might be extremely important to discover how they cope with, or 'repair', coughs or false starts; or indeed, how they cope with them when they come across them in people they have to communicate with. What is more, this latter argument could easily be extended to cover a much broader range of phenomena than just coughs and false starts.

Since competence is so difficult to define satisfactorily and since performance factors, however interpreted, can be shown to be of importance to a tester in a number of cases, it seems reasonable to conclude that it is performance testing which is likely to lead to a generally more satisfactory test, and that a tester who claims to be testing competence rather than performance is required to support the claim with empirical and theoretical evidence. For example, the statement that language competence includes a person's knowledge about a language cannot be used to justify using a question like 'How many tones has Cantonese?' as direct evidence of a testees competence in Cantonese, unless it can further be shown that it is precisely this type or form of knowledge that is in fact drawn upon by a language-user when words are actually uttered or comprehended. As a justification of this sort would, given the present state of our understanding of speech production and perception, be extremely hard to write, one is entitled, I think, to be a little sceptical of tests which do claim to be testing the linguistic aspects of communicative competence.

Producing a use-of-language test

The main stages in the production and exploitation of a use-of-language test, whether for testing proficiency or for research purposes, can be stated fairly clearly:

1. Discover what you want (or are required) to test.

2. Construct a good test based on this information.

3. Establish that you have actually produced a good test (and modify it accordingly).

4. When you have confidence in the resulting scores, process them to give you the information you need.

Such clarity, however, obscures the fact that we have very little idea of how to perform *any* of the four when we are dealing with a use-of-language test. To illustrate this in depressing detail, each of the categories will be considered separately, except for categories 3 and 4, which will be treated together for the sake of brevity.

Establishing what is to be tested

There are two separate parts to this first step, just as there are when designing teaching rather than testing material. Firstly, there is a profile of what the testee (or relevant native speaker) is supposed to do when interviewing, writing essays, arguing with works supervisors or whatever. Secondly, there is a list or statement of the things that will actually be tested. The first requires a framework to describe how language is used; the second requires, above all, a clear set of selection criteria, or objectives. (An exam syllabus for a public exam would constitute a further level somewhere in between these two). Although much criticism could be aimed at the way test objectives are often set up, the present paper will simply focus on problems of profiling.

There must now be literally thousands of studies, many of them excellent, which examine aspects of written and spoken discourse; but from the language tester's point of view there are effectively none which allow the speaker to have an infinite number of simultaneous motives or intentions, and which have a realistic number of levels of analysis - which in practice seems to mean a non-finite system. I would like, as a brief attempt to illustrate both of these contentions, to consider the possibilities inherent in a short adjacency pair dialogue which is assumed to take place in a restaurant:

Customer: Can you recommend something?

Waiter: The *homard* is good, Sir, very good indeed.

One might wonder what could motivate the waiter to reply like this. He could of course simply be responding to the conversational obligation to reply, in an apparently satisfactory way, to a question; indeed there might actually be no lobster (*homard*), and he could in fact be attempting to mislead the customer, or perhaps embarrass the management or even the cook. Equally, he might be trying to increase his commission by recommending the most expensive dish on the menu. On the other hand, he might also be trying to embarrass an obviously poor or vulgar customer. He could at the same time be attempting to impress the customer with his own breeding/foreign language ability/ lack of menial status, and indeed he might well be attempting to manipulate the power relation between the two of them. All of these motives could structure or constrain the conversation; the point is that one can always imagine a situation where there is yet one more motive operating. A similar point may be made about the levels on which the waiter operates, though here the question of scope, or generality, is relevant as well. The waiter might well see his answer as forming part of a general wasting of time till the restaurant closes. Less generally, he might see it as part of dining room (as against kitchen) encounters, customer (as apart from staff) encounters, as part of general food ordering routines, and as part of a pre-ordering subroutine in particular. He could also see it as part of a general routine whereby the customer is invited to choose, or hurry up, or order the most expensive food. It could, however, just as easily form part of a wider routine involved with shotgunning customers into thinking they should not order cheap dishes generally (and this could in fact contain 'ordering' as part of it, rather than the other way round). Again, one could go on and add more. Indeed, it seems impossible to find a single perfect solution. It is always possible to invent a situation where the hierarchy of levels is altered, or just one more level can be added.

The problems do not, of course, end with considerations of motivation and rank. One important aspect for the language tester is the ability of the profiling system to reflect the often complex ways in which the parts of subparts of a conversation or essay-writing activity, etc. are structured. Carroll (1980) strongly advocates the use of Munby's (1978) profiling system in the design of language-use tests, but this allows only 3 levels and simply lists items on each level. Such a system, it has been argued (Low, forthcoming) is therefore unable to describe how a 'typical' university student writes an essay, using certain types of background reading in different ways at different points in the process, and performing different types of revision and editing, depending on what has gone before. The general conclusion is that a framework which simply lists rather than describes structure and complex relationships within and between levels is of limited value to the language tester. The point to the whole argument is not to criticise work in discourse analysis, but simply to note that, although the constant production and reassessment of what are known to be imperfect theories is a sign of a robust and thriving academic subject which is continually questioning itself, it nevertheless constitutes a serious problem for the language tester.

The topic of deciding what to test involves more, however, than the selection of an appropriate theoretical framework. To illustrate this, two possible pitfalls will be discussed. Firstly, one must be extremely careful not to profile the wrong population. This amounts in practice to being careful not to apply research from an apparently similar group unthinkingly to one's own test. Although this warning may sound ludicrous and totally unnecessary, it is not unusual to find, for example, the results of research on the study needs of postgraduate university students being applied wholesale to the teaching and testing of undergraduates. The fact that some postgraduates happen to read a great deal, take part in cut-and-thrust seminars, or need to produce their own definitions is not a valid basis for assuming that all or even most undergraduates behave similarly.

A second pitfall is to generalise from a non-representative sample. For example, to stay with the topic of university students, it is extremely hard to devise a single test which measures the ability of a broad range of students to work satisfactorily in a lecture situation. The reason is simply that lecture situations vary widely, from a lecturer asking for no notes to be taken, to a situation where all language is rendered redundant by astute use of diagrams and equations, or one where the lecturer reads verbatim from a script (a not uncommon situation where the language of instruction is not the lecturer's first language). The implication here is that a test which involves a 15-minute audio tape of an unscripted lecture (or lecturette) asking the testee to write down the main points of what is said, may well be a very poor an inappropriate test for many students.

Constructing a good test

Although there are numerous discussions detailing the qualities of a 'good' test (eg. Dearden 1979; Guklford and Fruchter 1978: 407-457), there is minimal discussion about what a good use-of-language test might look like, how a tester might go about constructing one without losing control over the variables involved; or about the ways in which he might estimate, with a reasonable degree of precision, how good the resulting test was. This is perhaps surprising, given the amount of use-of-language testing that goes on.

Let us first consider the question of test directness[2]. Such language performance be tested directly or indirectly? This apparently simple question presupposes firstly that the directness of a test can actually be measured; yet there are almost no studies which attempt to do this (v. Lee and Low 1981, and Low, forthcoming, for some work on estimating the directness of essay-writing tests). Any attempt to answer the question of direct versus indirect, however, requires (preferably several) studies examining the relative ability of tests of varying degrees of directness to predict aspects of desired linguistic behaviour. A study by Lee (1981) hints that in certain cases a direct test might possibly act as a better predictor than an indirect one, but it is not conclusive. If this hypothesis could actually be empirically supported, then a very useful and (to the author at any rate) intuitively satisfying step towards a theory of language-use testing could be taken. Not only could a usable parameter of 'directness' be established, but the general theory could be given direction, since the 'highly direct' end of the scale could be taken as more central to the theory than the 'highly indirect' end. We could then say that in an ideal situation the most accurate test for a given situation would be likely to be the most direct one. Since in practice the tester would often be working under constraints (such as lack of time or money), the theory would envisage him being prepared to put up with a certain (hopefully specifiable) amount of indirectness. The idea of 'producing the most direct test possible in the circumstances' could thus be given some theoretical justification, and one could state the characteristics of the resulting test with much greater explicitness than usual. The study by Lee and Low (1981) suggests that degree of directness can in fact be reflected in test scores; but much more investigation is needed.

The problem of how to actually construct a use-of-language test is equally problematic. In Low (forthcoming) it is argued that where the initial language-use profile shows that speakers or writers are exploiting a developing context, then the test, assuming it is to be a direct test, should do so too. This implies that, in many cases, a storyline or line of development could profitably act as a thread through a series of subtests, linking them together (v. Harding 1977 for an early example of this, or Low, forthcoming, for an attempt to put a line of development through an essay-writing test). Although there has been little research done as yet on the use of storyline techniques and their effects, it is already clear that a considerable extra burden is being added to the shoulders of the language tester, since the idea of positively encouraging the testee to make use of earlier sections of the test in order to resolve a particular point, conflicts with the general desire that all measurements (or marks) should, as far as possible, be independent of each other.

Since it appears to be crucial that language-use tests involve language in some sort of context, a related question is: How should context be manipulated and presented?. The above argument in favour of the use of storyline techniques clearly assumes that meaningful measurements can often only be made if complex relationships (ie. complex contexts) are presented as part of the test. In the absence of studies which examine precisely how much context is generally needed before people can confidently select appropriate, or intentionally inappropriate, language, we might simply point out that the work by the Conversation Analysis school of sociologists (eg. H. Sacks 1972, E.A. Schegloff 1971, G. Jefferson 1974, and H. Garfinkel 1972) lends support to the contention being made here that in most cases a considerable amount is involved (v. for example the argument on the sophistication of the monitoring techniques speakers acquire and use, at the end of Sacks, Schegloff and Jefferson, 1974). If this is accepted, then the onus is on the test designer to justify the use of minimal contexts, of the sort suggested in Morrow (1977).

Following this line of reasoning, we could perhaps set up another hypothesis, which might be seen as a subpart of the directness hypothesis (above), to the effect that the ideal test would present the testee with full contextual data and that any selection or restriction that the tester cares to introduce for the purpose of testing should be justified explicitly. It should be understood that this proposal implies that general theory can and should be used in test design, but it does not, however, imply that a rich context is always the most desirable. As an illustration of this, we might examine a test in ESL that was produced for Chinese policemen in Hong Kong. After being asked their name and a few questions about the job, designed to relax them, testees were asked a question such as 'How do you get to Nathan Road?' Now, this would not appear to reflect what happens in real life, where tourists come straight up and ask their question without giving the policeman time to get used to the accent. Application of the idea of maximum test directness, or context authenticity, would here suggest that the provision of an extended context (for that is what in effect is being provided) might not in this case be provided in real life. One possible answer might be to simply restructure the test and place the direction question(s) first. This is a very simple example, but it does not show that (a) a minimal context can sometimes be justified, on validity as well as administrational ease grounds, (b) the application of theoretical ideas influences test structure as well as content, and (c) no meaningful discussion about test design is possible unless the designer explicitly states what he has done and why. In passing, it might be worth noting that the same argument can be applied to the use of 'padding' or redundancy, particularly in listening tests. The argument that listening tests should contain great redundancy and therefore should not involve the use of full transcripts has been made very forcibly by Frankel (1978). Yet, it has already been pointed out that many students have to spend hours listening to lectures which are read from just such transcripts, and one might object quite strongly to an ESL test for pilots which included numerous anecdotes from Air Traffic Control, designed to brighten up life in the cockpit.

It is important that the language tester realises that 'authenticity' includes the testee's mental processes as well as the content of what is spoken or written. This raises the difficult general question of whether one should test both process and product, or since people work in different ways, product alone. As an example of this we could describe an essay as a 'product', and a method of composing one as a 'process'. One can take different views on this topic, but it is surely vital that any and every use-of-language test be accompanied by a statement by the designer setting out the position that he has ultimately taken when designing the test. Assuming that one might wish to test process (and there seems to be little serious discussion - apart from polemic - on this topic in the literature) then the further question arises of how precisely does one ask questions, say, of a reading passage, in an authentic way, such that the 'natural' processes of reading for a particular purpose are not seriously altered or destroyed in the process? Reading passages pose an interesting problem. If you take a text and follow it with a series of questions, particularly in such a way that each part of the text has an equal probability of being tested, you must face several unpalatable facts:

(a) the testee must read the test before discovering why he has read it,

(b) the points you ask are unlikely to resemble the way a reader generally extracts information,

(c) you simulate partically no known reading task, firstly since not all information generally carries the same weight to a reader (indeed writers spend large amounts of revision time trying to downgrade or highlight chunks of data or language effectively, and secondly since humans do not generally carry out intensive reading for a very long period of time or on very long texts (and thus tend not to be very good at doing it), and

(d) unless a clear purpose is given to reading the passage, the testee is highly likely to interpret your questions in a idiosyncratic way.

If your response to this is to adopt the so-called 'search reading' format and to put the questions at the front, you are stuck with the observation that since most people appear unable to carry more than one or possibly two 'new' questions in their heads while trying to read a text at the same time, either they do not perform efficiently or else they simply restructure the task so that the questions come at the end. Another alternative, chopping up a text into short sections and adding questions between sections, has the added disadvantage that it seems to be almost impossible for the reader to hold the writer's line of argument in his brain when going from section to section. Such a test is also, in the author's own experimence, far more exhausting than the simple activity of reading and comprehending the text when it is not exploded, though this may relate to proficient readers more than to beginners(3).

Although the example could be extended by examining various other possibilities, I would like to suggest that a general conclusion can already be drawn: that is, that test questions which look like test questions, rather than like parts of a language-related activity which occur at such and such a point because that is where they normally occur within the activity (or because a certain problem has arisen), are often not perceived as 'authentic' by the testee, who promptly fails to produce precisely the behaviour one is trying to test. I interpret this to imply that not only do we need to re-examine, combing imagination with empirical data, the way subtests are related and linked together (whence the line-of-development concept), but there needs to be some serious rethinking about the whole concept of what a test question is and should look like.

This section can usefully end with some brief comments on a topic that brings together at a general level many of the points so far discussed. One of the major problems of simulating reality in language-use tests is that it conflicts with several of the principles developed for measuring things in general. For example, a direct use-of-language test necessarily confronts the testee with a mass of inter-relating variables, which conflicts with the belief that individual variables should be isolated and tested separately, if the tester is to retain an acceptable degree of control over the test. On the other hand, the effect of altering a situation (even to the minimal degree that it is simply a question of the testee knowing that it is a test) can seriously alter the testee's resposes and thus call into question the accuracy of the test. Indeed, the examiner, where it is a question of an interview-type oral test, is no more exempt from such 'behaviour-modification' than is the testee, as one tends to respond quite differently to halting speech or to continual repetition, or perhaps to a question couched in rude terms, when it occurs as part of a test; quite differently, for example, from the way one might respond to a similar occurrence at a bus stop. In this particular situation, a strong case can be made for altering the examiner's role (assuming he must take part in the interview), to one where he personally engineers problems and communication breakdowns. This means that these must be treated as important parts of the test (which further implies that the examiner can explicitly describe what type of breakdown has occurred and precisely how; and can also describe how well the testee has extricated himself) rather than as undesirable, if disturbingly common, events.

What is really being argued, through all the individual points, is that no serious test development can take place without the test designer being aware of what exactly he has constructed and why. This amounts, in an ideal state, to total designer accountability. Only when a designer can describe what he has constructed can any modification be carried out with any degree of confidence. This brings us to the next section, which concerns the evaluation of the test and the interpretation of test scores.

Validating a use-of-language test and interpreting test scores

This section examines problems in two areas, the first focussing on the test itself, the second focussing more on the testee.

In a sense, unless statistically established, all validity is related to perceived, or face validity. The perceiver can be the tester, a relevant independent expert in whatever the testee is to use language for, or the tester himself. Much can be established by designing a questionnaire to accompany the test. The problem here is partly the question of deciding what exactly it is that one wishes to find out. Within certain limits, well-designed questionnaires can provide invaluable data about biases that testees and experts feel are operating, numerical indices of perceived directness, a 'shock value' index for prognostic tests, and even an indication of perceived gross mismatches between actual test scores and the scores that testees feel they ought to have achieved. Despite the difficulty and the extra work involved, this sort of information is vital. A short illustration will suffice. A considerable amount of effort has been expanded in recent years on the design of English-language proficiency tests for intending university students where the medium of instruction will be English. Assuming that the vast majority of testees will be science students, the tester is faced with two basic options. He can either go for a series of subject-specific tests (with the disadvantage that some subjects will have been done at school, so that testees taking these papers will have a greater familiarity with the subject and the way it is expressed) or, alternatively, he can attempt to construct a single test for all testees, possibly using a neutral topic. Now if the tester opts for a series of subject-specific tests, then it is crucial to establish comparability between the tests and this would not seem to be at all possible without surveying both testees and relevant experts in the various fields at the very least. Similarly, if the tester prefers the concept of a single overall test, a survey of testees and relevant experts would seem to be the only way to obtain data like the following:

(a) What is the range of subject areas (ie. different 'sciences') to which the test can validly be applied?

(b) Do the apparently perfect distractors provided by the neutral topic in fact complicate the test unnecessarily, or even have the effect of destroying reading or listening processes/strategies which the testees normally employ quite satisfactorily when deciphering unfamiliar language in their own or familiar subject areas?

(c) Do some students feel more penalised than others; and if so, why?

(d) Are emotional reactions to the test, particularly adverse ones, significantly affecting test results and/or testees' confidence in the test?

This example makes it clear that not only are questionnaire and interview follow-ups an important part of use-of-language test development and monitoring, but that they ought, if possible, to be used with every single testee, not just with some small pretest sample.

Reliability

The other major area concerned with test evaluation is the question of reliability. The problem here is that the term 'reliability' is quite unclear as to its scope. Simply to say that a reliable test is one in which one can have confidence is no help at all. Nor is it any use simply lifting concepts, and hence formulae, from the evaluation of metal rulers or mechanical metering devices. One rather disturbing illustration of this is the concept of test-retest reliability. Now, it is clearly desirable that if a metal ruler is used twice by the same person to measure the same thing, the two measurements should be virtually identical. The demand for replicability is conceptually sound, for the simple reason that remeasurement is perfectly possible, since the act of using the ruler a second or third time is unlikely to affect the operating ability of the user. The trouble with language-related activities, though, is that people respond differently to new information (eg. seen as a challenge) than to old information when it is overtly repeated (eg. seen as irritating, boring, not a challenge). Indeed, when a sequence of words is repeated, it has different functions from those it had when it was uttered for the first time. At a more practical level, people's use of language appears to vary both quantitively and qualitatively depending on their current state of mind, and without access to this information it would be difficult to explain differences between test-retest scores with any confidence. Thus, where natural language-use is the object rather than just the instrument of testing, potentially serious theoretical as well as practical objections can be raised as to the suitability of using test-retest reliability (or, by implication, some indirect attempt to measure the same thing) as a test development tool. Similar objections can be made to other dimensions sometimes subsumed under the term 'reliability', but the above should be sufficient to show that a complete rethink is necessary in an attempt to produce a set of concepts which are :

(a) appropriate to language use

(b) clearly differentiated rather than being thrown into a single rag-bag category, and

(c) feasible both to operationalise in mathematical terms and to use in practice.

While on the topic of reliability, we should also draw altertive to two problems that are commonly met by language testers. The first is the unfortunate dilemma that (although several measurements ought to be made of any given point before one may have much confidence in the results) the more one concentrates on language use, particularly where one is aiming at a high degree of directness, the harder it becomes to design tests where the fact of repetition does not destroy the authenticity. It is candidly acknowledged by the author that the further demand that test questions should themselves be purposeful and resemble parts of natural activities makes test design infinitely harder.

The second problem is that, since use-of-language tests necessarily involve complex situations and interrelating sets of variables, subtests tend partially to overlap, with the result that composite or total marks do not necessarily represent what the tester imagined they would. Although multivariate techniques like factor analysis, which allow common factors to be reasonably well distinguished from (individual test) specific variances and which allow the tester several extra possibilities when weighting subtotals, are of extreme value to the tester (enthusiastic accounts can be found in Guilford 1948, and Rummel 1970), the results are of no use unless the tester has previously formulated the pattern he expected or intended to emerge.

Thus, although Morrow (1977) may be justified in rejecting the overuse of bivariate statistics, any potential language-use tester who goes on to infer that he is thereby exonerated from in any detail the characteristics of his test or from working out the expected pattern the results should take, is seriously deluding himself.

The term 'criterion-referenced' test is often used to denote a test where a scale of anticipated testee performance is set up and individual test scores are read off the scale, without necessarily being compared with each other. For the sake of clarity, scales of anticipated performance will be called 'performance criteria' to distinguish them from task profiles, context hypotheses and the like, which will be called 'test design criteria'. Now, the establishment of even remotely adequate performance criteria against which to assess testee performance is perhaps the hardest task in use-of-language testing. Any scale that is set up must be sufficiently 'natural' (itself hard to define here) for markers to be able to judge in terms of it; and expressed in such a way that the relationsip between any two scales used as part of the same test can be explicitly stated. An even more demanding view than this would like to see the scales justified by showing that these are the concepts (or clusters of concepts) that are actually present in people's minds when they engaged in communication situations (Hinofotis, forthcoming; Hinofotis and Bailey 1981). The encouraging point about the latter studies is that the authors realise the importance of explicit description (though this does not of course mean that one must necessarily agree with their descriptions) and of empirical support. The problem with designing scales of performance tends to be that, despite of often praiseworthy desire to use modern terminology, the result is often as vague and/or ambiguous as its predecessor. I would, for example, argue that 'exploring evidence of principles dealt with in lectures and tutorials' (Carroll 1980:3-5), or attempting to differentiate Intermediate Level from Foundation Level candidates by demanding 'reasonable accuracy in pronunciation' rather than 'unambiguous sound distinction' (ESB 1981) is little real improvement over statements like 'must show awareness of adverb usage'. As with so much in use-of-language testing, a considerable amount of rethinking, and a large number of well-designed empirical studies are needed before any really satisfactory answers to what often appear to be impossible questions can begin to emerge.

Conclusion

This paper has tried in a diffuse sort of way to make the following points:

(a) Language-use testing is difficult, as very little is yet known about it. What is known, however, suggests that it is likely to become harder not easier to design good tests in future.

(b) It would be desirable if a coherent, empirically-supported theory of language-use), could be developed, since ideas and techniques used in other types of test cannot simply be applied wholesale to use-of-language testing.

(c) Despite the fact that no such theory yet exists, the tester ought nevertheless to be completely accountable for any tests he produces, and complete accountability implies the ability to describe in detail the content and structure of the test, plus the pattern which the scores are intended to take.

Notes

1. The words 'he' and 'his' should be read to mean he/she and his/her throughout.

2. Although it is easy to give a general definition of a direct test (eg. 'A test where all significant aspects of a task and the conditions under which it is performed are present', Low, forthcoming. For other discussions, see Clark 1975, or, on the similar topic of 'work sample tests', Guilford 1948), it is extremely hard to set up sets of specific parameters in order to measure directness.

3. The importance of distinguishing between proficient and non-proficient readers when discussing texts or writing systems is underlined by Perera 1979.

REFERENCES

Canale, M. and Swain, M (1980) Approaches to Communicative Competence, RELC Occasional Papers, 14. Singapore: SEAMEO Regional Language Centre.

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

Clark, J.L.D. (1975) 'Theoretical and technical considerations in oral proficiency test', R.L. Jones and B. Spolsky,(eds.) Testing Language proficiency. Arlington: Center for Applied Linguistics, 10-28.

Dearden, R.F. (1979) 'The Assessment of learning,' Brıtish Journal of Educational studies, 27, 2:111-24.

English Speaking Board (1981) Oral Assessments in Spoken English as an Acquired Language. Southport.

Frankel, M.A. (1978) 'The case for Unscripted Listening Comprehension Materials', Paper in SEAMEO RELC 13th Regional Seminar, 18th April, 1978.

Guilford, J.P. (1948) 'Factor analysis in a test development program', Psychology Review, 55:79-94.

Guilford, J.P. and Fruchter, B. (1978) Fundamental Statistics in Physchology and Education. 6th ed. McGraw-Hill.

Harding, A. (1977) 'Specimen tests', New Objectives in Modern Language Teaching; Defined Syllabuses and Tests in French and German . Oxfordshire Modern Languages Advisory Committee.

Hinofotis, F.B. (forthcoming) 'The Structure of oral communication in an educational environment: a comparison of factor analytic rotational procedures', J.W. Oller (ed.) Issues and Prospects in Language Testing Research.

Hinofotis, E.B. and Bailey, K.M. (1981) 'American undergraduates' reactions to the communication skills of foreign teaching assistants', J.C. Fisher, M.A. Clark and J. Schachter (eds.) On TESOL 1980: Building Bridges: Research and Practice in Teaching English as a Second Language, Washington, D.C.: TESOL.

Hinofotis, F.B., Bailey, K.M. and Stern, S.L. (1981) 'Assessing the oral proficiency of prospective foreign teaching assistants: instrument development', A. Palmer and P.J.M. Groot(eds.) The Validation of Oral Proficiency Tests: Selected Papers from the Colloquium on the Validation of Oral Proficiency Tests. Washington, D.C. TESOL.

Hudson, R.A. (1975) 'The Meaning of question', Language, 51,1: 1-31.

Lee, Y.P. (1981) 'Measurement and evaluation of communicative competence without necessary reference to a-priori theoretical models: the case for direct language tests', J.A.S. Read(ed.), Papers on Language Testing, RELC Occasional Papers 18. Singapore: SEAMEO Regional Language Centre, 86-98.

Lee, Y.P. and Low, G.D. (1981) 'Classifying tests of language use,' Paper presented at 6th AILA World Congress, Lund, Sweden.

Low, G.D. (forthcoming) 'The Direct testing of academic writing in a second language'.

Morrow, K.E. (1977) Techniques of Evaluation for a Notional Syllabus. Reading: Centre for Applied Language Studies, University of Reading, (Study commissioned by the Royal Society of Arts).

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

Oller, J.W. (1979) Language Tests at School. London: Longman.

Perera, K. (1979) Cruttendon Language in Infancy and Childhood: a Linguistic Introduction to Language Acquisition. Manchester: MUP.

Rummel, R.J. (1970) Applied Factor Analysis. Nothwestern U.P.

Sacks, H., Schegloff, E.A. and Jefferson, G. (1974) 'A Simplest systematics for the organisation of turn-taking for conversation', Language, 50:696-735.

Sumby, W.H. (1960) 'Control Tower Language', Language and Speech, 3: 61-70

Valian, V. (1979) 'The Wherefores and therefores of the competence-performance distinction', W.E. Cooper and E.C.T. Walker eds., Sentence Processing. Hillsdale, N.J.: L. Erlbaum Associates, 1-26.

Wiemann, J.M. and Backlund, P. (1980) 'Current theory and research in communicative competence', Review of Educational Research, 50,1: 185-199.

SOME NOTES ON INTERNAL CONSISTENCY RELIABILITY ESTIMATION FOR TESTS OF LANGUAGE USE

Y.P. Lee
Language Centre
University of Hong Kong

1. Introduction

Recent attempts to develop tests of language use (or 'communicative' tests) have been primarily for the purpose of making language tests more valid than are the average discrete-point structural language tests. However, the reliability of these tests of language use has hardly been investigated. This paper is an attempt to focus on some of the problems encountered in estimating reliability for tests of language use, and to offer suggestions on how existing statistical techniques can be of use.

2. Discrete-Point Tests and Tests of Language Use

2.1 Before starting out, it is necessary to discuss the terms 'discrete-point test' and 'test of language use'. Terms like these are, more often than not, either too vague, or too much used to be precise. The attempt to specify the meaning of the two terms here is simply to establish a common ground for the discussion in the present paper.

2.2 By 'discrete-point test' is meant here the type of language test which is composed of a number of generally unrelated items derived from elements identified in linguistic theory as constituting a particular language. The main characteristics of such a test are: (1) each item stands by itself and can be regarded as a subtest (v. Guilford 1954, Ch.14); (2) the test score is the sum total of correct responses; (3) items can be freely added on to or taken away from the test without affecting other items. Examples of discrete-point language tests can be found in a number of public examinations in English, e.g. TOEFL, and H.K.S.C.E. English Paper II.

2.3 By 'test of language use' is meant the type which measures the candidate's ability to handle specific language tasks, e.g. taking down telephone messages for secretaries, following a message under strenuous noise conditions (for example, air traffic controllers). Examples of this type can be found in the J.M.B. Test of English (Overseas) and the English Proficiency Test for Engineering Students devised by the Language Centre, University of Hong Kong (Lee 1979).

It is quite difficult to describe general characteristics of tests of language use, because they can be very different one from another, and they are still at a trial period of development. However, there seems to be one general feature in tests of language use which is of relevance to the present discussion. It can be generally stated that tests of language use require testees to handle extended pieces of language both receptively and productively. Consequently, testee answers are in the form of either an extended message or a number of interrelated questions which can even be objectively marked. For example, testees may be required to read or listen to the description of a process (say in engineering) and are required to label an empty flow-chart representing the process. It is clear that, in such a test, the items (cells in the flow-chart) are interrelated. They form a structured whole and cannot be easily manipulated as in the case of discrete-point tests. Moreover, the number of items that can be included in any one such test is also quite limited. To come back to the test just outlined, the type of process chosen can have a number of stages in the region of 20 or so. This is because anything more complicated than that (e.g. a computing procedure involving some 60 or more steps) would drastically reduce the chances of it being a viable language test. It would over-tax human comprehension power and memory.

3. Problems in Estimating Reliability of Tests of Language Use

3.1 Because of the limited space allocated to this paper, the discussion here will be restricted to an examination of tests whose score is a composite score derived from the sum total of objectively marked items. Furthermore, only internal consistency reliability will be discussed, and neither parallel form nor test-retest reliability will be covered in the present paper.

3.2 Two major problems exist in the estimation of internal consistency reliability which are relevant here. The first is establishment of test homogeneity. To the extent that a test measures the same characteristic (or characteristics) in all its parts (usually items), it is called a homogeneous test (v. Guilford and Fruchter 1978, 417-18). The more homogeneous a test is, the higher its internal consistency. In statistical terms, a homogeneous test has items that have more or less the same item intercorrelations. The second problem has to do with the relation between internal consistency reliability and test length. A long test, other things being equal, has a higher reliability coefficient than a short one.

3.3 In discrete-point language tests, homogeneity is more or less ascertained by basing test items on language units identified through theoretical linguistics (generally of the taxonomic structuralist type). Moreover, items are the units in such tests and can, therefore, be added into or subtracted from it.

In tests of language use, however, neither test homogeneity nor test length can be as easily controlled as in discrete-point tests. First of all, tests of language use are based on language as communication, of which we have, as yet, no clear and precise knowledge. More often than not, such tests are samples of concrete language behaviour and can be labelled as work-sample derived tests (v. Guilford 1948). Secondly, as pointed out in 2.3, the items in such tests are interrelated in a whole. The units of manipulation, therefore, are whole tests or subtests and not individual items as in the case of discrete-point tests.

So, both test homogeneity and test length pose serious problems for reliability estimation in tests of language use. It must be admitted that little thought has been given to the solution of these problems. Some may even dismiss them as more or less unimportant for tests of language use. The present writer thinks, however, that, unless we are willing to pay the price for not having reliable tests of language use, we will have to come face to face with the problems.

The remainder of this paper is an attempt to establish test homogeneity and to tackle the problems of test length for tests of language use.

4. Towards a Solution

4.1 It seems profitable to suggest a solution by way of an example. The Language Centre of the University of Hong Kong administers each year a battery of English proficiency tests to all incoming first year Engineering students. The battery consists of 3 reading, 2 listening and 3 writing subtests which are all language-use tests and are derived from actual English language tasks required in first year Engineering academic work. A detailed account of the battery, known as the English Proficiency Test for Engineering Students, is found in Lee (1981).

All the subtests with the exclusion of the writing subtests are composed of objectively scored items and are of the type described in 2.3 above. In the discussion that follows, the writing subtests will not be considered.

4.2 The internal consistency reliability estimates (r_{tt} - reliability coefficient, and $r_{t\alpha}$ - reliability index), for the reading and listening subtests taken separately are as follows:-

	r_{tt}	$r_{t\alpha}$
Reading 1	0.70	0.84
Reading 2	0.58	0.76
Reading 3	0.80	0.89
Listening 1	0.64	0.80
Listening 2	0.84	0.92

Table 1: Reliability estimates of the reading and the listening subtests in the English Proficiency Test for Engineering Students.

Nunnally (1978) suggests that the lower limit for r_{tt} should be 0.70, if we are dealing with exploratory studies; but should be 0.80 or more, if we are dealing with more practical decision making, like student placement. From this point of view, the set of reliability coefficients (r_{tt}) in *Table 1* does not appear to be particularly promising. Test length is clearly a problem. The number of items in each subtest is:

Reading 1	13
Reading 2	6
Reading 3	10
Listening 1	8
Listening 2	14

It is particularly relevant to observe that the two lowest figures of r_{tt} (0.58 in Reading 2, and 0.64 in Listening 1) are both associated with the two shortest subtests.

The problem of homogeneity is even more serious. There is simply no a-priori basis to argue that the subtests are homogeneous; and whatever conclusion can be drawn as regards test homogeneity has to be established a-posteriori through statistical techniques.

4.3 Factor analysis was employed to provide a solution for both the homogeneity and the test length problem. Basically, factor analytic technique is a method of identifying underlying dimensions measured by a number of tests. In our case, it can be used to find out whether the 5 subtests are measuring 5 different characteristics, or whether there are fewer than 5 distinct dimensions being measured. In the latter case, factor analysis can also show how the subtests are clustered together. (The interested reacher can refer to Guilford 1954, Ch. 16 for a detailed description of factor analysis.)

All the subtests, including Writing, were factor analysed, using Principal Component Analysis without Iterations. Four factors were extracted for rotation, their eigenvalues being 2.42, 1.19, 0.93, 0.89. The percentages of variance accounted for by each of the four factors were: 30.3, 14.9, 11.7, 11.1. The Varimax rotated factor matrix is as below in *Table 2*:

	Factor 1	Factor 2	Factor 3	Factor 4
Reading 1	0.14	-0.12	0.78	0.25
Reading 2	0.02	0.10	0.07	0.90
Reading 3	0.04	0.38	0.70	-0.16
Listening 1	0.13	0.82	0.03	-0.02
Listening 2	0.17	0.59	0.09	0.36
Writing 1	0.81	0.08	-0.04	0.17
Writing 2	0.82	0.11	0.10	-0.01
Writing 3	0.76	0.14	0.15	-0.04

Table 2: Varimax rotated factor matrix of the Reading, the Listening, and the Writing subtests (cut-off point for factor loading 0.50).

The factor pattern is very clear. Factor 1 should be a writing factor, Factor 2 a listening factor, Factor 3 a reading factor. Factor 4 is a bit more complex. The two subtests loading high here are Reading 2 (0.90) and Listening 2 (0.36). The latter is of a rather uncertain status, because it is below the cut-off point (0.50) adopted here. However, the overall loading pattern in Factor 4 would give Listening 2 considerable prominence. If this interpretation is accepted, Factor 4 can be described as an information transfer factor, because both Reading 2 and Listening 2 involve transferring a linguistic discourse into graphic representation.

As far as the five (reading and listening) subtests are concerned, it can be seen that the factor analysis solution has identified three distinct underlying dimensions which, by definition, are homogeneous. Moreover, each dimension is the combination of two subtests, which, as a result, can be collapsed together to make one longer test. It should be pointed out that, in combining the subtests, due consideration was given to the factor loadings which indicate the relative strength of association of the subtests with the factors. These loadings, were then used as weights on the scores of the subtests, before they were combined.

The results of the factor analysis have given us 3 sets of homogeneous scores with 22 items (Listening 1: 8; and Listening 2: 14) in Factor 2, 23 (Reading 1: 13; and Reading 3: 10) in Factor 3, and 20 (Reading 2: 6; and Listening 2: 14) in Factor 4. It should, perhaps, be necessary to point out that Listening 2 appears in two of the above combinations. A test thus loading significantly in more than one factor is known as a factorially complex test. Such tests are relevant to more than one underlying dimension in the factor matrix. However, the weighting according to factor loading would ensure that due consideration is given to possible differential prominence such complex tests have in the relevant factors.

4.4 Internal consistency reliability was estimated for the three sets of factor scores with the subtests properly weighted. The method employed was from Guilford (1954, 393) - Reliability of Composite Scores:

$$r_{ss} = 1 - \frac{\Sigma w_j^2 s_j^2 - \Sigma w_j^2 s_j^2 r_{jj}}{\Sigma w_j^2 s_j^2 + 2\Sigma w_j w_k s_j s_k r_{jk}}$$

where r_{ss} = reliability coefficient (r_{tt}) of a sum of components

w_j = weight assigned to any component J

w_k = weight assigned to any component K

s_j, s_k = SD's of components J and K, respectively

r_{jj} = reliability coefficient for any component J

r_{jk} = intercorrelation of components J and K

The reliability coefficients (r_{tt}) and the reliability indices ($r_{t\alpha}$) are as below in *Table 3:*

	r_{tt}	$r_{t\alpha}$
Factor 2	0.82	0.91
Factor 3	0.80	0.89
Factor 4	0.80	0.89

Table 3: Internal consistency reliability estimates of factor scores.

These sets of estimates are noticeably improved as compared with those in *Table 1.*

5. Conclusion

It appears that the employment of the factor analytic technique can help to resolve statistically the two major problems related to internal consistency reliability estimation for tests of language use. It offers positive suggestions on how the problem of test length can be overcome. In actual practice, justification of whether two or more tests subsumed under one single factor can be collapsed or not has still to come from test content.

Moreover, the procedure just described may also be of wider application, in cases where a-posteriori methods are required in relıability estimation. One possible case in point would be the cloze procedure. The general practice has been to derive a cloze score from the unweighted sum of the item (blank) scores. Such a scoring method assumes that the cloze procedure is more or less homogeneous (or factorially simple). However, little experimental work has been done to validate such as assumption. It is possible that the method proposed in this paper may here find a useful application.

References

Guilford, J.P. (1948) 'Factor analysis in a test-development programme', Psychology Review, 55, 79-94.

———————— (1954) Psychometric Methods, 2nd ed., McGraw-Hill, New York.

Guilford, J.P. and Fruchter, B. (1978) Fundamental statistics in psychology and education, 6th ed., McGraw-Hill, New York.

Lee, Y.P. (1979) A Comparison between 'Global Integrative' Language Test and 'Task-Based' Communicative Skill Language Test as Predictor of Language Proficiency. Unpublished MA dissertation, University of Hong Kong.

————— (1981) 'Evaluation and Measurement of Communicative Competence without Necessary Reference to A-Priori Theoretical Models - The Case for Direct Language Tests', Papers on Language Testing: RELC Occasional Papers No. 18, Singapore, SEAMEO Regional Language Centre, 86-98.

Nunnally, J.C. (1978) Psychometric Theory, 2 ed., McGraw-Hill, New York.

THE USE OF CLOZE PROCEDURE IN THE FORM III ENGLISH LANGUAGE SECONDARY SCHOOL SCALING TEST

R.K. Johnson
School of Education
University of Hong Kong

Introduction

The Hong Kong Department of Education recently established a Secondary School Scaling Test to determine which Form III students will be allocated places at Form IV level. For the English Language paper, an explanatory statement was distributed to all secondary schools along with sample papers to illustrate the types of items that would be used in the various sections.

The general approach adopted for the English Language Scaling Test was summed up as follows:

'The emphasis is placed upon English as a "tool for use" rather than as a formal system, upon the communicative functions which the forms serve rather than on the forms themselves.'

A clear distinction is made between the types of item to be used in the Scaling Test and those used in 'conventional language tests' which 'focus directly upon formal features of the English language'. I assume that the writer is referring to test: made up primarily of discrete point structural items.

Given this clearly stated philosophy, it was surprising to find that cloze procedure was used in the section for testing reading comprehension.

In a recent paper (Johnson 1981), I discussed a number of aspects of cloze testing, amongst them my reasons for believing that normal reading and the completion of cloze passages have little in common. The purpose of the experiment reported in this paper is to determine to what extent the multiple-choice cloze test used in the Reading Comprehension section of Sample Scaling Test A differs from an equivalent test consisting of discrete point structural items; and thus, by implication, whether cloze is an appropriate procedure to use in the Scaling Test, given the assumptions upon which the test is based.

Subjects

The subjects consisted of 264 Hong Kong Secondary school students who speak English as their second language. Of these, 64 were Form IV students, 35 Form III students and 165 Form II students.

Test Materials

For the reasons stated above, the study made use of the multiple-choice cloze test in the Reading Comprehension section of sample Form III Scaling Tests, which were distributed to schools in 1980. The sixteen items of this test were isolated from the discourse, modified where necessary so that the meaning was clear out of context, and ordered in such a way that no possible clues either of cohesion or coherence (Halliday and Hasan 1976) remained to assist the subject in selecting the best slot filler. Thus two tests were prepared. *Test 1* was the discrete point test. *Test II* was the cloze test. It was assumed that any differences in performance by subjects on these two tests could be ascribed to this difference in format.

Method

The subject were divided into two matched groups, Group A and Group B, either by assigning one class at a particular form level to Group A and another class to Group B, or by dividing each class into two and allocating half the subjects to Group A and half to Group B. There was an even distribution of Form II, III & IV students in each group. All subjects took both tests: Group A in the order *Test I* first and *Test II* second, Group B in the order *Test II* first & *Test I* second. This was done in order to eliminate the order in which the tests were taken as a variable. It also permitted a comparison of 'practice effects' i.e. the extent to which performance on the second test taken showed an improvement over performance on the first.

The tests were administered by the class teacher in timetabled English language periods no more than two days apart. No subjects did the tests sequentially. It is assumed that the subjects would notice the similarities between the two tests, but this was not brought to their attention or mentioned by the teacher (if at all) until after both tests had been completed.

Results

A. *Test I* vs *Test II*: Discrete Point vs Cloze Test Formats

Table I (pg.) shows the breakdown of results for each item showing the percentage of subjects who chose A, B, C and D respectively. Category E is for nil response on that item. The items are listed in *Test II* in the order which corresponds to the sequential order of items in *Test I* (e.g. Item 6 in *Test II* corresponds to Item 1 in *Test I*). In reporting the results, the item is referred to as in *Test I*, the corresponding number in *Test II* can be identified by reference to *Table I*.

TABLE I

(Group A & Group B Combined)

TEST I							TEST II							
Item	A	B	C	D	E	Total	Item	A	B	C	D	E	Total	Test (II-I)
1	181	25	46	8	2	262	6	195	17	35	17	-	265	4.8%
%	69.1	9.5	17.6	3.1	0.8		%	73.9	6.4	13.3	6.4			
2	22	173	32	35	-	262	11	12	173	40	35	4	265	-0.5%
%	8.4	66.0	12.2	13.4			%	4.5	65.5	15.2	13.3	1.5		
3	18	154	37	52	1	262	7	26	152	29	55	2	264	-1.2%
%	6.9	58.8	14.1	19.8	0.4		%	9.8	57.6	11.0	20.8	0.8		
4	48	50	124	37	3	262	2	40	66	132	25	1	264	2.7%
%	18.3	19.1	47.3	14.1	1.1		%	15.2	25.0	50.0	9.5	0.4		
5	49	15	161	36	1	262	3	41	31	161	30	1	264	-0.5%
%	18.7	5.7	61.5	13.7	0.4		%	15.5	11.7	61.0	11.4	0.4		
6	69	42	51	98	2	262	12	36	34	49	144	1	264	-17.1%
%	26.3	16.0	19.5	37.4	0.8		%	13.6	12.9	18.6	54.5	0.4		
7	46	64	17	134	1	262	5	30	58	24	148	4	264	5.0%
%	17.6	24.4	6.5	51.1	0.4		%	11.4	22	9.1	56.1	1.5		
8	30	139	46	46	1	262	13	19	135	61	47	2	264	-2.0%
%	11.5	53.1	17.6	17.6	0.4		%	7.2	51.1	23.1	17.8	0.8		
9	136	48	57	18	3	262	15	139	53	49	19	4	264	0.8%
%	51.9	18.3	21.8	6.9	1.1		%	52.7	20.1	18.6	7.2	1.5		
10	48	99	25	90	-	262	1	41	92	28	101	2	264	-3.0%
%	18.3	37.8	9.5	34.4			%	15.5	34.8	10.6	38.3	0.8		
11	16	79	96	69	2	262	16	22	55	95	89	3	264	-0.6%
%	6.1	30.2	36.6	26.3	0.8		%	8.3	20.8	36.0	33.7	1.1		

TABLE 1 (Con't)

TEST I							TEST II							
Item	A	B	C	D	E	Total	Item	A	B	C	D	E	Total	Test (II-I)
12	28	121	25	87	1	262	4	30	129	28	75	2	264	2.7%
%	10.7	46.2	9.5	33.2	0.4		%	11.4	48.9	10.6	28.4	0.8		
13	30	84	98	47	3	262	8	34	68	125	35	2	264	9.9%
%	11.5	32.1	37.4	17.9	1.1		%	12.9	25.8	47.3	13.3	0.8		
14	46	30	126	59	1	262	14	36	38	125	63	2	264	-0.8%
%	17.6	11.5	48.1	22.5	0.4		%	13.6	14.4	47.3	23.9	0.8		
15	36	89	53	83	1	262	9	34	86	44	97	3	264	5.0%
%	13.7	34.0	20.2	31.7	0.4		%	12.9	32.6	16.7	36.7	1.1		
16	56	69	35	101	1	262	10	55	78	24	103	4	264	0.5%
%	21.4	26.3	13.4	38.5	0.4		%	20.8	29.5	9.1	39.0	1.5		
Total of correct responses					2,024 48.3%	4,192						2,145 50.8%	4,224	2.5%
Total w.outj 6 & 13					1,828 49.84%	3,668						1,876 50.76%	3,696	0.9%

It will be noted that, with the exception of items 6 and 13, there is a remarkable degree of consistency in the responses.

In gross percentage terms, which are not necessarily significant, only these two items show a greater than 5% difference in the numbers of correct answers. Overall, including items 6 and 13, all subjects scored 48.3% correct answers on *Test I* (the discrete point test) and 50.8% correct answers on *Test II* (the cloze test). With items 6 and 13 removed from the analysis the percentage of correct answers on *Test I* and *Test II* respectively are 49.8% and 50.7%, a difference of only 0.9%.

These figures, and in particular the generally consistent pattern which emerges overall, strongly suggest that the two tests are not different in any important respect, and (taking into account Table 2 below) that they are in fact testing very much the same thing. As a further check, those test papers on which names had been written (197 out of the 264) were extracted so that degree of consistency of performance on the two tests could be subjected to a x^2 analysis of significance. *Table 2* shows for each item, the number of subjects who selected the correct answer in *Test I* & *Test II* (A), those who made an incorrect choice in both tests (B)*1, those who chose incorrectly in *Test I* but correctly on *Test II* (C) and those who chose correctly on *Test I* but incorrectly on *Test II* (D).

A x^2 test of significance of difference was made on the figures in C. & D. (to find out whether there was a significant difference between the response patterns of the two groups of subjects who performed inconsistently on the two tests.) The performance of subjects in category A was consistent, and for category B is assumed to have been consistent. The behaviour of subjects in categories C and D is assumed to have been random except where the x^2 test shows a significant difference between the two groups. (The x^2 value at the 0.05 level of significance is 3.84).

TABLE 2

Item		A	B	C	D	Total	x^2
Test I	Test II	+ 1 +	- 1 -	- 1 +	+ 1 -		
1	6	131	38	17	11	197	0.893
2	11	99	49	26	23	197	0.082
3	7	84	41	37	35	197	0.014
4	2	72	75	30	20	197	1.620
5	3	96	57	20	24	197	0.205
6	12	54	72	49	22	197	9.521
7	5	78	71	25	23	197	0.021
8	13	77	71	20	29	197	1.306
9	15	84	66	26	21	197	0.340
10	1	41	117	20	19	197	0.-
11	16	43	90	32	32	197	0.016

*1 Note, these subjects were consistent in that they chose incorrectly; they did not necessarily select the same distractor in each test, though this was generally the case.

TABLE 2 (Con't)

Item Test I Test II	A + 1 +	B - 1 -	C - 1 +	D + 1 -	Total	x^2
12 - 4	82	65	31	19	197	2.420
13 - 8	70	59	48	20	197	10.721
14 - 14	70	79	23	25	197	0.021
15 - 9	46	110	24	17	197	0.878
16 - 10	47	113	18	19	197	0.-
Totals (all items)	1174	1173	446	359	3152	9.188*
w/out items 6 & 13	1050	1042	349	317	2758	1.443

Table 2 shows that over all items there is a significant difference (x^2 = 9.188) between groups C & D in that of those subjects who were inconsistent, a significant number followed the pattern of wrong choice in the discrete point test and correct choice in the cloze test. However, this degree of significance is accounted for by two items only - item 6 and item 13 - and for no other item was there any significant difference between the performances of the two categories of response pattern. With these two items removed from the analysis, the remaining fourteen items are shown to be not significantly different (x^2 = 1.443). Thus the overall picture suggested by the response pattern shown in *Table I* is confirmed, and for this test, reorganising the cloze test into discrete point items has no important effect. Performance was affected by the change (adversely) on only two items, and though the change was shown to be statistically significant, it should be borne in mind that the x^2 calculation did not take into account the large number of consistent response and the actual number of responses which show a pattern in favour of the cloze format is a very small proportion of the total number (2.76%). This suggests that the level of statistical significance between the groups which performed inconsistently (C + D) has no actual importance

in the context of the total overall response pattern, that for all practical purposes the results obtained under the two testing conditions are identical, and the cloze test used in the sample paper of the Form 3 scaling test is no different so far as these subjects are concerned from the discrete point test derived from it.

B. Practice Effect: Group A vs Group B

Group A took *Test I* (discrete point) first and *Test II* (cloze) second.

Group B took *Test II* (cloze) first and *Test I* (discrete point) second.

Table 3 shows the analysis and the sixteen items in terms of the following categories:

P:- Practice effect: a clear gain in the percentage of correct responses from *Test 1* to *Test 2* regardless of test type.

X Little or no difference between the percentage of correct responses on *Test 1* and *Test 2*, regardless of test type.

C A higher percentage of correct responses was recorded on the cloze item regardless of whether the cloze test was taken first or second.

D A higher percentage of correct responses was recorded on the discrete point item regardless of which test was taken first.

Pc A practice effect was recorded (gain in percentage of correct response from 1st to 2nd Test) but the trend was considerably stronger on the cloze item.

Pd A practice effect was recorded, but the trend was considerably stronger on the discrete point item.

TABLE 3

	P	X	C	D	Pc	Pd	Total
No. of Test Items	2	4	4	1	2	3	16

P. Two items (3 + 7) recorded a pure practice effect with a gain of 6.2% in Group A, and 9.0% in Group B for item 3 and 13.5% in Group A and 4.3% for Group B for item 7.

X Four items (2, 9, 11 and 16) showed little or no differences between the percentage of correct answers on the test taken first and the test taken second.

C Four items showed a clear bias in favour of the cloze format: i.e. Group A showed a gain and Group B a loss. The percentages are given for each item in *Table 4* showing the percentage correct on each test and the respective gains and losses from first test to second test.

TABLE 4

Item	Group A			Group B		
No.	Test 1	Test 2	% Gain	Test 1	Test 2	% Loss
1	71.9%	78.7%	+ 6.8%	68.7%	66.1%	- 2.6%
6	38.5%	56.6%	+18.1%	52.3%	36.2%	-16.1%
13	25.2%	42.6%	+17.4%	52.3%	50.4%	- 1.9%
15	32.6%	34.6%	+ 2.0%	39.1%	30.7%	- 8.4%

It is strongly suggested by these figures that, in item 6, the cloze format is providing clues which are not available to the subject under the discrete point format. Item 13 shows a similarly large gain for Group A, but the Group B students seem to have benefited from doing the cloze test first, and scored almost identically with the discrete point format. This was not the case with item 6. Presumably, candidates who were correct with the cloze format and then incorrect in the later test, failed to recognise the item on its second appearance.

D Only one item (10) favoured the discrete point format for Groups A & B (See *Table 5*)

TABLE 5

Item	Group A			Group B		
No.	Test 1	Test 2	% Loss	Test 1	Test 2	% Gain
10	43.7%	42.6%	-1.1%	26.6	31.5	+4.9%

Other items showed some practice effect, but clearly favoured either the cloze format or the discrete point format.

Two items showed practice effect but favoured the cloze format (see *Table 6*)

TABLE 6

Pc

Item No.	Group A			Group B		
	Test 1	Test 2	% Gain	Test 1	Test 2	% Gain
4	46.7%	54.4%	+ 7.7%	45.3%	48.0%	+ 2.7%
12	46.7%	52.2%	+ 5.5%	45.3%	45.7%	+ 0.4%

Three items showed practice effect but favoured the discrete point format (see *Table 7*)

TABLE 7

Pd

Item No.	Group A			Group B		
	Test 1	Test 2	% Gain	Test 1	Test 2	% Gain
5	62.2%	64.7%	+ 2.5%	57.0%	60.6%	+ 3.6%
8	54.1%	54.4%	+ 0.3%	47.7%	52.0%	+ 4.3%
14	50.4%	51.5%	+ 1.1%	43.0%	45.7%	+ 2.7%

The totalled results show an overall practice effect favouring the cloze format in Group A (5.0% gain for *Test II* over *Test I*). For Group B there is a minimal practice effect (0.2%) favouring *Test I* over *Test II* (i.e. the discrete point test).

The percentage of correct responses overall was rather consistent: (see *Table 8*).

TABLE 8

% of Correct Responses	Group A		Group B	
	Test 1	Test 2	Test 1	Test 2
	48.1%	53.1%	48.3%	48.5%

The figures suggest that there was some advantage overall in taking the cloze test after the discrete point test, whereas taking the cloze test before the discrete point test offered no advantage. It will be noted that the overall percentage of correct answers for Group A on its first test was 48.1%, for Group B on its first test 48.3%; and again the figures confirm that under normal test conditions the two forms of this test are equivalent. The fact that a difference emerges on a repeat test using the alternative format is arguably irrelevant since normal testing procedures do not permit such conditions.

C. Differences between Forms 2, 3 & 4

The numbers of subjects in each group are inconsistent, the sample for Form 3 being in particular undesirably small for such a comparison. Only the overall figures are given in *Table 9*.

TABLE 9

	No. of Students	% Correct		% difference
		Test I	Test II	
Form 2	165	40.3%	42.1%	1.8%
Form 3	35	41.7%	43.9%	2.2%
Form 4	64	72.8%	77.0%	4.2%

Table 9 shows very little difference between Form 2 and 3. Both show a slightly higher % correct responses for the cloze format. Form 4 subjects, as might be expected of a group which has survived the selection procedures at the end of Form 3, perform far better than the lower forms, (proportionally over 30% more correct responses on both tests) and also show an increased ability to use discourse clues to obtain correct answers under the cloze format, though this is still small overall (4.2%) and is derived from a few items only, primarily items 6 and 13. Overall responses on these items are shown in *Table 10* (below) for each Form.

TABLE 10

	Form 2			Form 3			Form 4		
	% Correct Test I	Test II	% Diff.	% Correct Test I	Test II	% Diff.	% Correct Test I	Test II	% Diff.
6	34.5%	46.7%	12.2%	32.4%	57.1%	24.7%	47.6%	73.4%	25.8%
13	44.2%	47.3%	3.1%	8.8%	11.4%	2.6%	34.9%	67.2%	32.3%
No. of Students	165			35			64		

There are instances where the discrete point format shows a gain over the cloze format, but these are less consistent across groups and of a lesser order of magnitude. As examples, Item 8 shows a 4.9% difference in favour of the discrete point format for Form 2 subjects, but shows 4.5% & 1.7% differences in favour of cloze for Forms 3 and 4 respectively. Item 10 shows a 13.2% difference in favour of discrete point for Form 3, and a 10.6% difference in the same direction for Form 4, but Form 2 results show a 1.8% difference in favour of the cloze format.

Discussion of the Results

The most important fact to establish with regard to this experimental study is that the results relate only and solely to the tests under discussion, and cannot be generalised to encompass the relationship between cloze tests and discrete point tests as a whole. In this particular study the results on two items seem to have been affected by the clues of coherence and cohesion which are associated by many people with the cloze format, and are the basis for calling cloze tests 'global' tests of language. These items (6 and 13) are discussed below.

For fourteen items, and for the two tests overall, there was a remarkable similarity and consistency in the response patterns. In other experiments using other passages of discourse as a basis for test construction, it is to be anticipated that in some cases more items would reflect the influence of discourse clues, and that other cases would reflect less, or indeed none. It follows that the randomised gapping procedures generally associated with the preparation of cloze tests are counter-productive, since it is impossible to determine to what extent, if any, a particular cloze test differs from a discrete point test based upon the same structural/lexical items. It may be desirable, in fact necessary, for those who intend to use cloze tests to undertake an investigation into the chosen passage of the kind reported in this study. The arguments against random procedures in the development of cloze tests, and in favour of principled decision-making which takes a large number of factors into account, are given in some detail elsewhere, and will not be repeated here (Johnson, 1981). I will only add that this study provides support for the general conclusion, arrived at by Alderson (1979), that cloze tests resemble discrete point tests more than any other kind of test, and in particular more than they resemble such tests of higher level language ability as reading comprehension.

The differences between the psycholinguistic processes involved in completing a cloze test and those involved in reading comprehension are again discussed in some detail in the paper already mentioned, and are not in any case directly relevant to this study. It is clear however that cloze cannot be equated with reading comprehension (unless discrete point items are also equated with reading comprehension) and if cloze is to be used in the Form III Scaling Test, it should be used in some other section and for some other purpose than as a test of reading comprehension.

A minor but interesting point of detail remains to be discussed: the nature of the discourse clues which facilitated the higher level of correct responses to items 6 and 13 under the cloze format.

The items are set out below, with the information available to the cloze reader, assuming other choices have been made correctly, and assuming that the reader is attempting to use meaning clues from beyond the immediate context in which the gap occurs.

Item 6/12

Discrete Point

I would not ______________ anything to happen to it.

A. do
B. fear
C. hope
D. want

Cloze

At first, I thought I would like to (*wear*) (*the watch you gave me*) to school tomorrow (*so that*) all my friends could see it. But after thinking about it, I (*decided*) not to, because I would not ____________ anything to happen to it.

Item 13/8

Discrete Point

The watch was exactly the right ____________.

A. measure
B. order
C. size
D. way

Cloze

Then I (*put*) (*the watch you gave me*) on and found that it was exactly the right _______________.

As *Table I* shows clearly (above), the major difference between the response pattern for item 6 in the discrete point test and item 12, the corresponding item in the cloze test, is the switch of responses from the distractor A to the correct choice D (see *Table 11*).

TABLE 11

	Test I	Test II
Choice	Item 6	Item 12
A	26.3	13.6
B	16.0	12.9
C	19.5	18.6
*D	37.4	54.5
E	0.8	0.4

* Correct choice

The use of *hope* (C) rather than *want* or *think* is a error which is common to second language speakers of English in all three continents in which I have taught. Its consistent effectiveness here as a distractor is not surprising. *Fear* (B) is an understandable error in item 6 if the learner associates with *fear* the structural patterns associated with verbs such as *expect* or *like*, or of course *want*. In item 13, however, *fear* is semantically unacceptable in the context provided: i.e. in a clause which provides a reason for deciding not to wear the watch. The fact that the semantic clues were so ineffective suggests that the subjects did not, or could not make use of these clues of coherence (There is evidence from the Form IV subjects that the better readers were more willing and better able to take account of such clues. 29% chose B in item 6, and only 8% in item 12.) The switch of responses from A in item 6 (*Test I*) to D in item 12 (*Test II*) is difficult to explain, since the same structural constraints operate against A in both items, while there could be some semantic justification for it in item 12. It certainly is not obvious to me that there are strong clues in the cloze passage which trigger this shift, and the more skilled readers (Form IV) were not attracted to choice A in the first place.

The evidence on this item, then, seems to suggest that semantic and/or discourse clues had little bearing upon the response patterns, and that under the cloze format, as in the discrete point, a limited context formed the basis for decision-making. Where the context provided by the 'cloze format' provided a strong presuppositional basis for rejecting one clue (B), these contextual clues were largely ignored.

The response patterns for items 13/8 are shown in *Table 12.*

TABLE 12

Choice	Test I	Test II
	Item 13	Item 8
A	11.5	12.9
B	32.1	25.8
*C	37.4	47.3
D	17.9	13.3
E	1.1	

* Correct response.

In this case the major overall shifts in pattern response are from B and D (in item 13) to C (in item 8). The likely explanation here is not that additional semantic clues were available for the cloze item, but that the discrete point item is a poor item. Both B & D are, in admittedly rather obscure circumstances, acceptable English. This explanation is considerably strengthened by the fact that the bulk of those who changed their responses in favour of the correct choice in the cloze passage were the skilled (Form IV) readers: 16 chose B in *Test I,* as against 6 in *Test II.* For choice D, it was 12 as against 7. Two subjects failed to make a choice in *Test I,* non failed to make a choice in *Test II.* The analysis of response patterns on this item (*Table 2*) suggests that it is these comparatively advanced readers who moved consistently to Choice C in *Test II* who are responsible for the level of significance in the response pattern shifts for this item.

For this item, no additional clues were provided by the cloze situation; rather the ambiguity in item 6 (*Test I*) was resolved, and the comparison between test formats is therefore invalid in this case since a valid item is contrasted with an invalid one. It is surprising therefore that Form II subjects showed an increase in correct answers (*Test II* over *Test I*) of only 3.1%, and Form III subjects of 2.6%, compared with the 32.3% increase in correct responses by Form IV students

Thus, although it might have been supposed that in these two items at least a clear distinction might emerge between what is tested under the discrete point format and what is tested under cloze procedure, this is not the case. In item 6, clear semantic clues provided by the cloze passage for rejecting B were largely ignored while the motivation for the shift from A to D is obscure, to me at least. In item 13, the apparent difference between discrete point and cloze tests proves to be essentially a difference between an ineffective item (because ambiguous) and an effective one.

The final conclusion must therefore be, that this cloze passage taken from a Sample Scaling Test and intended as a measure of reading comprehension, is not in any important sense a different measure from a set of equivalent discrete point items.

References

Alderson, J. Charles. (1979) 'The Cloze Procedure and Proficiency in English as a Foreign Language', TESOL Quarterly. Vol. 13, No. 2, June, pp. 219-227.

Halliday, M.A.K. and Ruqaiya Hasan. (1976) Cohesion in English. Longman, London.

Johnson, R.K. 1981. 'Some Questions about Cloze Tests.' In John A.S. Read (ed.) Directions in Language Testing. SEAMEO Regional Language Centre, Singapore.

Appendix I

Multiple-Choice Items.

For each blank, choose the best answer from the choices given below the item.

1. When I opened the box and saw what was inside it, I could hardly ______________ my eyes!

   A. believe
   B. rely
   C. trust
   D. understand

2. After thinking about it, I ______________ not to do it.

   A. concluded
   B. decided
   C. judged
   D. thought

3. I ______________ the watch on.

   A. measured
   B. put
   C. tested
   D. took

4. I was really ______________ to get such a beautiful watch.

   A. enjoyed
   B. gifted
   C. round
   D. lovely

5. When the parcel came, everyone gathered ______________ to watch me open it.

   A. before
   B. for
   C. round
   D. since

6. I would not ________________ anything to happen to it.

A. do
B. fear
C. hope
D. want

7. I tried to guess ______________ it might be.

A. how
B. that
C. this
D. what

8. Dad says I am a very lucky girl to have _____________ a kind and generous Auntie.

A. so
B. such
C. that
D. this

9. I already have ____________ for a gift.

A. an idea
B. idea
C. the idea
D. the ideas

10. I am writing to thank you very much for the birthday present which the postman _______________ this morning.

A. arrived
B. delivered
C. reached
D. sent

11. Your gift will _____________ surprise you!

A. beautiful
B. hardly
C. really
D. very

12. I got very ______________.

A. excite
B. excited
C. excitement
D. exciting

13. The watch was exactly the right ______________.

A. measure
B. order
C. size
D. way

14. We were very glad to see in your letter that you ______________ to visit us at Chinese New Year.

A. get
B. hoping
C. plan
D. will

15. I thought I would like to ____________ my watch to school tomorrow.

A. dress
B. carry
C. show
D. wear

16. I took it ______________ all my friends could see it.

A. for
B. to let
C. to make
D. so that

Cloze Passage

For each blank in this letter, choose the best answer from the choices given on the opposite page.

26th September, 1979.

Dear Auntie Mary,

I am writing to thank you very much for the birthday present which the postman ____(1)____ this morning. I was really ____(2)____ to get such a beautiful watch.

When the parcel came, everyone gathered ____(3)____ to watch me open it. I got very ____(4)____ as I tried to guess ____(5)____ it might be. When I opened the box and saw what was inside it, I could hardly ____(6)____ my eyes! Then I ____(7)____ it on and found that it was exactly the right ____(8)____ .

At first, I thought I would like to ____(9)____ it to school tomorrow ____(10)____ all my friends could see it. But after thinking about it, I ____(11)____ not to, because I would not ____(12)____ anything to happen to it.

Dad says I am a very lucky girl to have ____(13)____ a kind and generous Auntie - and I agree with him! We were very glad to see in your letter that you ____(14)____ to visit us at Chinese New Year. I already have ____(15)____ for a gift that will ____(16)____ surprise you!

Thank you once again for your marvellous present.

Love,

Amy.

THE EFFECTS OF THE SHIFTING OF INSTRUCTIONAL MEDIUM ON STUDENTS' PERFORMANCE IN SELECTED ANGLO-CHINESE SECONDARY SCHOOLS IN HONG KONG[1]

Peter T.K. Tam

School of Education
University of Hong Kong

ABSTRACT

In the Hong Kong primary schools, students can be identified by the language of the instructional materials they utilize. These are, the E.P. students, defined operationally in this study as those students who studied three science or social science subjects in the medium of English in the sixth grade in an Anglo-Chinese primary school; and the C.P. students who studied these subjects in the medium of Chinese . When both groups of students enter the same Anglo-Chinese secondary school where the instructional medium is completely or essentially in English, it is hypothesized that the E.P. students would gain an advantage over the C.P. students who face a sudden change of instructional medium. The major research question of this study is: What are the effects of shifting of instructional medium on both types of students in terms of academic performance? In this study, it is assumed that, when the E.P. and C.P. students are admitted to the same secondary school, both groups of students do not differ significantly in academic ability at the intake level. On the basis of results of a survey reported previously, six schools judged to have a balanced proportion of E.P. and C.P. students were selected to participate in this project. Within each school, both groups of students were compared with respect to family education, language confidence, language competence, internal examination results, language understanding, and language preference. Owing to the large number of comparisons to be made, all the null hypotheses were tested at the 0.001 level of significance in order to reduce the Type I error of the experiment. It was found out that in terms of family education, there is a greater proportion of E.P. to C.P. students progressing to or beyond the first degree level. However, the majority of both groups do not go beyond Form V level. With respect to the dependent variables, results strongly indicated that the E.P. students are relatively more confident in English language skills; they are more competent in English reading and listening tests. In terms of the internal examination results all of the comparisons were in favour of the E.P. students in English and Western History, and most of the comparisons were in favour of them in subjects such as E.P.A., Geograph, and Science. For subjects such as Mathematics, Chinese, Chinese History, Art, Physical Education, and Conduct, no convincing evidence is obtained that one group did significantly better than the other. The majority of

[1] This is the second part of the research project entitled "The effects of changing the instructional medium on students' learning" which was supported in whole by the University of Hong Kong.

both types of pupils preferred both their teaching and instructional materials to be presented in English with Chinese used in a supporting role. The majority of both groups of students believed, rightly or wrongly, that instruction in the medium of English can help them to improve their English. Although learning to be conversant in English is an important aim, the majority of both groups of students perceived English and Chinese to be equally important.

INTRODUCTION

In terms of the instructional medium used, it is possible to identify two types of students in primary schools in Hong Kong. Students of the first type study in Anglo-Chinese primary schools, where Chinese is the medium of instruction. In this study, students of the first and second type are referred to as E.P. and C.P. students. As the majority of local students in primary schools are of the C.P. type, they may experience great difficulty when they first enter an Anglo-Chinese Secondary school, where the instructional medium (particularly for textbook materials) shifts abruptly to English. When students fail to adjust to a new instructional medium, they tend to lose ground in performance. The detrimental effects resulting from a shift of instructional medium are likely to be considerable, but it seems that not much empirical data has been collected to test the proposition. The main objective of this study is to assess the effects of the shifting of instructional medium on both types of student in terms of school performance. Since the impact of language shifting on the E.P. students is theoretically less than that on the C.P. type, it is hypothesized that the E.P. students would gain an advantage over the other type when studying those subjects that require a fair amount of competence in English.

METHOD

Sample

In identifying the target population for this study, two problems have to be solved. The first is to identify those Anglo-Chinese secondary schools with a balanced proportion of the E.P. and C.P. students in Form I classes. As this kind of information is lacking, a survey was conducted to a large stratified random sample of secondary schools to identify the target schools. Part of the results of this survey was presented in another paper entitled *A Survey of the Language Mode Used in Teaching Junior Forms in Anglo-Chinese Secondary Schools in Hong Kong*. As a result of this survey, six schools which were judged to have a good proportion of both types of students were invited to participate in this study.

Table 1: The Sample Size

School		Anglo-Chinese Primary E.P.	Chinese Primary C.P.	Total[1]	N. of Intact Classes
Sir Ellis Kadoorie School	(K)	15	17	32	2
New Method College	(N.M.)	239	117	356	9
Raimondi College	(R)	190	79	269	7
St. Paul's College	(P)	92	144	236	6
St. Joan of Arc	(A)	80	137	217	6
St. Paul's Convent School	(C)	119	53	172	5
Total		735	547	1282	35

Table 1 shows the names of these schools together with the corresponding number of the Form I students belonging to the E.P. or C.P. categories. Excluding a small percentage of repeaters and non-native speakers of Chinese, the achieved sample size is 1,282.

The second problem confronted in this study is to differentiate E.P. from C.P. students. A commonly accepted definition of the E.P. student is one who has graduated from an Anglo-Chinese primary school. However, this definition is not useful enough in this study as there is great variation among these schools in terms of the number of subjects taught in the medium of English. To overcome this difficulty, the E.P. student is operationally defined in a quite arbitrary manner as the student who has studied in primary six at least three of the following subjects in English, or essentially in English. These subjects are:

[1] Repeaters and non-native speakers of Chinese were excluded in this study.

Arithmetic, Science, Health, Geography, History, and Social Studies. Since Arithmetic is generally given more curriculum time than the other subjects, it is weighted as two subjects. In conducting this study, each student was given a questionnaire to indicate both the language modes used by the teacher and the instructional materials used in each of the subjects in primary six. If a student indicated that there were at least three subjects in which the language mode used in the instructional materials was English or essentially English, then the student was defined as belonging to the E.P. type. If the language mode for the book materials was in English, but the mode for the teacher was completely in Chinese, then this student would still be classified into the E.P. type; but these cases were very rare. A four-category scheme was used to represent the language mode of teaching:

Language Mode	Notation
Completely in English.	E
Essentially in English, but Chinese is used for explaining the more difficult parts.	E(c)
Essentially in Chinese, but for technical terms, important words or phrases, the English equivalent is also given.	C(e)
Completely in Chinese.	C

Table 1 shows that out of the total 1,282 students participating in this study, 735 belonged to the E.P. type, and 547 to the C.P. type. *Table 2* (pg. 52) shows a comparison of the two groups of students in terms of the language mode they were exposed to at the primary six level. Each number in this table is a percentage over the total of either the E.P. or the C.P. students in all the six schools in this study. Take for instance the subject Science: 89% of the E.P. students reported that their Science instructional materials in primary six were completely in English (i.e. the E mode), 91% of the C.P. students had their materials of this subject in Chinese (i.e. the C mode). In this table, two points are to be noted. First, in the subjects of Arithmetic, Science, Health, Geography, and History, a small language of the percentage of the C.P. students reported that the language for teaching or the language of the

Table 2

Frequency Distribution in Percentages Showing the Language Mode used by the Teacher and in the Instructional Materials when the Students were in Primary Six

Subject at Primary Six	Students	Teacher				Instructional Materials			
		E	E(c)	C(e)	C	E	E(c)	C(e)	C
English	E.P.	23	69	8		94	5		
	C.P.	8	59	33		86	9	5	
Arithmetic		7	30	24	39	51	3	6	39
			3	16	81		4	14	82
Science		23	55	18	1	89	8	1	1
			4	5	88		5	2	91
Health		14	28	15	34	53	4	1	34
			1	2	92		1	1	92
Geography		18	64	12	1	86	7	1	1
			1	1	7		2		6
History		15	53	9	1	72	5		1
			1	1	3		1		3
Social Studies					86				86
					97				97

instructional materials was in the E(c) mode. This contradiction is attributable either to response errors or to the fact that a small percentage of E.P. students who failed to meet the operational criteria of having the instructional materials for three subjects in the E or E(c) mode. Second, for the E.P. students concerned, the E mode was commonly used in the instructional materials. However, in teaching, the modes with more Chinese were more frequently used. This reflects a very common teaching situation in Hong Kong where English textbook materials are explained with a fair amount of Chinese in the classroom.

Measuring Instruments

The final internal examination results of each student were collected and these data were used for comparing the two groups of students within each school. These internal subjects included: English, History, Economics and Public Affairs (E.P.A.), Geography, Science, Mathematics, Chinese, Chinese History, Art, Physical Edcuation, and Conduct Assessment. In addition, two questionnaires, two Chinese and two English language tests were constructed to assess the attitudes and language competence of the students.

Both questionnaires were in Chinese. The first one was constructed to obtain from each student three types of information. First, the student was asked to indicate the highest educational level reached by any one member of his 'family', which was defined in a very restricted sense as consisting only of the student's parents, brothers, and sisters. Educational level was measured in six categories, namely: 'No Formal Education Received', 'up to Kindergarten', 'Primary 6', 'Form 5', 'Matriculation'. and 'University or Above'. Each students was also to indicate the language(s) in which their parents wished them to become competent. Second, each student was asked to indicate the language mode of instruction in primary six. The purpose of this part of the questionnaire was to classify the student into either the E.P. or the C.P. type. The methodology used for this purpose was described previously under the section 'Sample', above. Third, the student was asked to indicate the degree of confidence in performing activities such as reading an English textbook at Form I level, listening to a lecture, or expressing himself fluently in spoken or written English. The student was also asked to indicate the degree of confidence in performing these activities when the language involved was Chinese. A six-point Likert type scale was used to measure the degree of confidence with values '1', '2', '3', '4', '5', and '6' which were defined as 'Completely Diffident', 'Mostly Diffident', 'Slightly Diffident', 'Slightly Confident', 'Mostly Confident', and 'Completely Confident', respectively.

The second questionnaire was constructed to collect the student's opinions on the language mode of instruction. With reference to a specific school subject, the student was asked to indicate the preferred language mode to be used by the teacher, the student, and the instructional materials respectively. From a prepared list, the student was asked to select the reasons (not more than four) for the choice of the language mode. The contents of this prepared list were based on the written responses given previously by a much larger sample of students who were asked similar questions in the survey stage of this study. In addition, the student was asked to indicate the degree of understanding of the language mode currently being used in the school. A five-point scale, with anchors ranging from understanding 'Practically Nil', 'A Little', 'About one half', 'More Than One Half', to 'Practically All' was used for this purpose.

In order to assess the language competence of the two groups of students, four language tests were constructed. These are summarized briefly in *Table 3*. Owing to limitation of time and resources, only the reading and listening skills of the students were assessed. Each of these skills was tested in Chinese and in English. As all the four tests consisted of different items, students are not comparable across them. However, the two reading tests are 'comparable' in the sense that both of them contain items matched in terms of difficulty, content, and objective. The matching of these items was made possible because all the test items were piloted in a much larger sample during the survey stage of this study. Constructed in a similar manner, the two listening tests are also comparable. In terms of content and objectives, the reading and listening tests are different. The two reading tests were designed to test the ability of the students to make inferences from short stories in which the information was given indirectly or implicitly, (*implicit* materials). The listening tests were designed to test the student's ability to listen to the details and the ideas explicitly stated

Table 3: Summary of Language Tests Used in this Study

Skill	Language	Test Abbreviation	No. of Items	Test Content
Reading	Chinese	RIC2	25	Implicit materials
	English	RIEI	24	Implicit materials
Listening	Chinese	LEC2	22	Explicit materials
	English	LEE1	22	Explicit materials

in the stories, (*explicit* materials). Many of the ideas and contents of these tests were drawn from the Primary Reading Assessment Units developed by Campbell, Tracy, and McErlian (1973). The original tests were developed for Canadian children in the early primary levels. By means of piloting these tests during the survey stage of this study, a number of the original items were revised to make them suitable for the local target population.

Procedures

After selecting the six schools with a balanced proportion of E.P. and C.P. students, contact was made with the heads of these schools to invite them to participate in the study. All the tests and questionnaires were administered in the June to July period of 1979. It took the students about two and a half hours to complete all the instruments. While the students were attempting the questionnaire section, they were guided through by trained test administrators item by item to ensure that misunderstanding of the questions was minimized. Using the facilities of the Language Centre of the University of Hong Kong, great care was exercised to ensure professional quality in the production and play-back of the listening tests. After all the data were collected, they were coded and electronically processed in the Computer Center of the University of Hong Kong using mainly the Statistical Package for the Social Sciences. Due to the large number of statistical comparisons to be made, all hypotheses were tested at the 0.001 level of significance.

Results

The results obtained in this study are presented and discussed under six headings: (1) Family education and parental expectation of the language competence of the students, (2) Language confidence, (3) Reading and listening competence, (4) Internal examination results, (5) Language understanding, and (6) Language preference of the students. Under each of these headings, the E.P. and C.P. students are compared within the same school. Throughout this study both groups of students of the same school are assumed to be comparable in academic aptitude at the entry point to Form I in as much as they entered the same secondary school. No attempt was made to compare students across schools as this kind of comparison would be confounded by factors such as the sex of students, differences in standard among schools, different instructional materials and teaching methods, and so on.

Family Education

Table 4 shows a comparison of the E.P. and C.P. students within

TABLE 4

Family Education of the Students in Terms of the Highest Education Reached by Any One Close Member of the Family

	1 No Education		2 Up to Kindergarten		3 Up to Primary Six		4 Up to Form 5/Middle 5		5 Up to Matriculation		6 Up to University or above	
School	E.P.	E.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K						6	80	65	13	24		6
N.M.		1			5	7	49	53	27	26	18	11
R.	1	3	1		9	10	48	59	22	23	17	5
B.					4	6	43	56	18	19	33	19
A.					6	9	46	59	21	20	25	9
C.					3	6	48	57	21	23	24	13
Weighted % of Total		1			5	7	48	57	23	22	21	12
				z-value	-1.567		-3.113***		.276		4/279***	

x^2 = 21.006***
(d.f. = 3)

*** (p <.001)

each school with reference to the family education in terms of the highest educational level reached by any one member of the family. In this table, the abbreviations in the first column identify the schools. The six levels of family education are represented by the numbers 1 to 6 on the top row of this table. The values in the same row indicate the percentages of the corresponding E.P. or C.P. students of the same school coming from each of the levels of family education. For instance, 5% of the E.P. students and 7% of the C.P. students in school N.M. come from families whose education went no further than primary six. Empty cells in this table indicate non-existing cases or missing data as a result of non-response by some students. It can be seen that there are very few families below the primary six level of education. Under each educational level, students of the same group (either E.P. or C.P.) were pooled together, and the proportion of this sub-sample over the total number of the corresponding E.P. or C.P. students in this study is indicated in the row entitled 'Weighted % of Total'. Since there were very few cases of the families in educational levels 1 and 2, the hypothesis of independence between student type and family education was tested only for educationa levels from 3 to 6. The computed x^2 value of 21.006 with three degrees of freedom indicates that the two groups of students pooled from all schools are heterogenous with respect to family education at the 0.001 level of significance. Pairwise comparison between these two groups under each family educational level using the z-test for proportions indicated significantly ($p < .001$) a greater proportion of E.P. than C.P. families progressing to or beyond first degree level; but there is a greater proportion of C.P. families who did not progress beyond Form V level. In brief, the data collected in this study seem to suggest that the higher the educational level concerned, the greater is the proportion of E.P. to C.P. students.

Table 5 (pg.58) shows a comparison of the E.P. and C.P. students in terms of parents' expectation of the student's language competence. The data, expressed in percentages in terms of either the E.P. or C.P. students, is presented in a similar format as that in the previous table (*Table 4*). The hypothesis of independence between student type and parental expectation was tested with the x^2. Results indicated that both groups of students are homogeneous with respect to the language competence expected of their parents. Further examination of the data in the row entitled 'Weighted % of Total' indicated that more than 50 percent of the parents of both groups hoped that their children would be conversant in both Chinese and English. From this data, it seems that the majority of parents are not biased in favour of either Chinese or English. However, considering only the minority of parents favouring either Chinese or English, then for both groups of students concerned the percentage of parents favouring 'Mainly in English' (22%) is much greater than that favouring 'Mainly in Chinese' (2%).

TABLE 5

Parents' Expectations of the Language(s) in which Pupils Should Become Competent

	Mainly in English		Mainly in Chinese		Both English and Chinese		No Specific Indication	
School	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K.	20	24		6	67	65		6
N.M.	27	19	3	5	60	65	10	10
R.	18	11	2	1	51	51	14	14
B.	17	24			60	57	23	19
A.	24	31	3	4	55	50	16	13
C.	22	13	1		67	79	7	4
Weighted % of Total	22	22	2	2	59	58	13	13
Z-value	0		0		.066		0	

$x^2 = .6579$
(d.f. = 3)

Language Confidence

A comparison of both groups of students in terms of confidence in English language skills is shown in *Table 6* (Pg.60) The data in this table shows the subscale mean confidence scores of the two groups of students with respect to reading, listening, speaking, and writing. Since the subscale values ranged between 1 and 6, the mid-point value is 3.5. The greater the score the higher the degree of confidence. The overall confidence score in the last column is a summation of the subscale confidence scores. With respect to each language skill, the two groups of students of the same school are compared by means of the *t*-statistic. Comparisons that are significant at the 0.001 level are indicated by the three asterisks next to the corresponding *t*-value. The total number of significant comparisons in each column is summarized in the bottom row. Take, for instance, the value of 4 at the beginning of this rows it means that in four out of the six schools the E.P. students showed a significantly higher mean reading confidence score than that of the C.P. students. Positive values in this bottom row indicate the number in favour of the C.P. students. The data in this table seems to support two conclusions:

(1) In all the thirty comparisons between the two groups in terms of English language confidence, the E.P. students consistently obtained a higher mean score. In addition, 23 of these comparisons were in favour of the E.P. students at the 0.001 level of significance. Another difference between the two groups is that all the mean confidence scores below this value. It can be seen that the data here provide strong support for the conclusion that the E.P. students are more confident than the C.P. students in English language skills.

(2) With reference to the specific language skills, a very interesting phenomenon appears and that is that both groups of students obtained higher mean confidence scores in reading and listening than in speaking and writing. This is true for the students in all the six schools studied. The data in this table strongly supports the conclusion that students are more confident in receptive language skills, such as reading and listening, than in the productive language skills, such as speaking and writing.

However, the above two conclusions do not hold true with reference to the Chinese language confidence scores. As shown in *Table 7* (pg.61), in contrast to the data in the previous table, E.P. students do not gain any significant advantage over C.P. students. In three cases, C.P. students obtained a significantly higher mean confidence score, but the data as a whole does not provide any convincing evidence that one group is more confident in Chinese language skills than the other group. Also, when students are compared against themselves with reference to confidence in receptive and productive language skills, no convincing evidence is obtained that the confidence in one type of skill is significantly higher than in the other.

TABLE 6

Comparison of Students' Confidence in English Language Competence

School		Reading		Listening		Speaking		Writing		Overall	
		E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K	Mean	4.667	3.588	4.667	3.588	3.933	3.176	3.800	3.000	17.067	13.353
	Std.	(.816)	(1.278)	(.724)	(1.121)	(.458)	(1.074)	(.561)	(1.118)	(2.052)	(4.227)
	t	2.802		3.185		2.531		2.504		3.092	
N.M.		4.389	3.641	4.393	3.607	3.711	2.923	4.079	3.419	16.561	13.615
		(.837)	(1.004)	(.914)	(1.082)	(.942)	(1.146)	(.929)	(1.139)	(2.718)	(3.451)
		7.406***		7.165		6.892***		5.834***		8.767***	
R.		4.168	3.582	4.237	3.557	3.621	2.899	3.868	3.241	15.921	13.228
		(.792)	(.826)	(.886)	(.843)	(.779)	(1.008)	(.802)	(.990)	(2.443)	(2.778)
		5.457		5.814***		6.328***		5.439***			
P.		4.946	4.236	4.571	4.014	3.935	2.972	4.196	3.528	18.25	14.847
		(.761)	(.853)	(.805)	(.989)	(.887)	(1.01)	(.842)	(.968)	(6.948)	(3.085)
		5.500		4.510***		7.484***		5.434***		5.142***	
A.		4.280	3.783	4.432	3.860	3.795	3.25	3.873	3.547	17.699	14.457
		(1.034)	(.910)	(1.111)	(1.044)	(1.027)	(1.101)	(1.17)	(1.078)	(9.913)	(3.34)
		3.575		3.670***		3.457***		1.982		3.395***	
C.		4.277	3.925	4.647	4.019	3.630	2.736	3.983	3.132	16.563	13.792
		(.812)	(.615)	(.777)	(.820)	(.901)	(.836)	(.844)	(.785)	(2.459)	(2.307)
		2.815		4.811***		6.140***		6.236***		6.952***	
(No. of comparisons that are significant)		4		5		5		4		5	

*** (p<.001)

TABLE 7

Comparison of Students' Confidence in Chinese Language Competence

School		Reading		Listening		Speaking		Writing		Overall	
		E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K.	Mean	5.267	5.812	5.333	5.937	5.533	5.937	5.200	5.687	21.467	23.375
	Std.	(.799)	(.403)	(.724)	(.250)	(.640)	(.25)	(.862)	(.479)	(2.696)	(1.025)
	t	-2.421		-3.146		-2.344		-1.961		-2.637	
N.M.		5.402	5.647	5.703	5.664	5.736	5.466	5.515	5.534	22.377	22.310
		(.813)	(.649)	(.550)	(.632)	(.623)	(.908)	(.709)	(.828)	(2.042)	(2.511)
		-2.836		.596		3.277		-.224		.268	
R.		5.332	5.456	5.721	5.696	5.511	5.494	5.211	5.430	21.795	22.101
		(.777)	(.616)	(.564)	(.515)	(.711)	(.768)	(.834)	(.710)	(2.271)	(2.164)
		-1.263		.339		.175		-2.045		-1.020	
P.		5.337	5.573	5.685	5.706	5.435	5.336	5.130	5.366	21.576	21.951
		(.651)	(.633)	(.512)	(.567)	(.746)	(.813)	(.801)	(.785)	(2.066)	(2.516)
		-2.759		-.288		.941		-2.228		-1.194	
A.		5.014	5.477	5.522	5.695	5.290	5.500	4.928	5.414	20.681	22.18
		(1.194)	(.709)	(.964)	(.596)	(1.059)	(.813)	(1.142)	(.779)	(3.676)	(2.564)
		-3.414***		-1.554		-1.551		-3.530***		-3.350***	
C.		5.051	5.453	5.629	5.774	5.427	5.491	5.051	5.377	21.111	22.189
		(.859)	(.539)	(.583)	(.505)	(.758)	(.669)	(.981)	(.765)	(2.33)	(3.771)
		-3.136		-1.562		-.528		-2.141		-2.281	
(No. of comparisons that are significant)		-1		0		0		-1		-1	

***(p < .001)

Figure 1 (pg. 63) shows the difference between the two groups in terms of the overall confidence scores in English and Chinese language skills. A significant difference between the two groups in the same school is indicated by an asterisk. It is obvious from this figure that the profile of mean confidence scores in English language skills of the E.P. students is significantly higher than that of the C.P. students. There is no comparable difference between the two profiles in the case of the Chinese confidence mean scores. However, both groups of students consistently showed a significantly higher degree of confidence in Chinese than in English.

Reading and Listening Competence

In the participating schools intact classes in the target population were selected randomly and assigned either two reading or two listening tests, with the first test in Chinese and the second in English. The purpose of a fixed sequence with the Chinese test given first was to ensure that the students fully understand the test instructions. *Table 8* (pg. 64) shows the results of the students in these language tests. As shown in this table, there is only one school ('R') in which both the reading and the listening tests were administered. In the other schools, only one type of test was given. As in the case of the confidence scores, the differences between means favoured the E.P. students significantly only in the cases where the language involved was English. This difference between the two groups is illustrated also in *Figure 2* (pg.65) which shows that the profiles of reading and listening mean scores of the E.P. students are significantly higher than those of the C.P. students when the tests are in English. However, in the case of Chinese language tests, such differences between profiles are almost undiscernible.

Internal Examination Results

Within each school, the E.P. and C.P. students were compared with reference to the final internal examination results. *Table 9* shows the list of internal subjects used in making the comparisons. These subjects, ranging from English to Conduct Assessment, are arranged in an order with subjects requiring more English being placed first in the list. Within each school, both groups of students are compared in each of these subjects using the *t*-statistic. The three asterisks besides the *t*-values indicate comparisons that are significant at the 0.001 level. The total number of comparisons that are significant under each subject is listed in the bottom row. As in the other tables, positive values in this bottom row indicate the number of comparisons in favour of the E.P. students and negative values in favour of the C.P. students. In this table very interesting results may be observed:

(1) For the subjects of English and Western History, which require a good command of the English language, all the ten comparisons were significant in favour of the E.P. students.

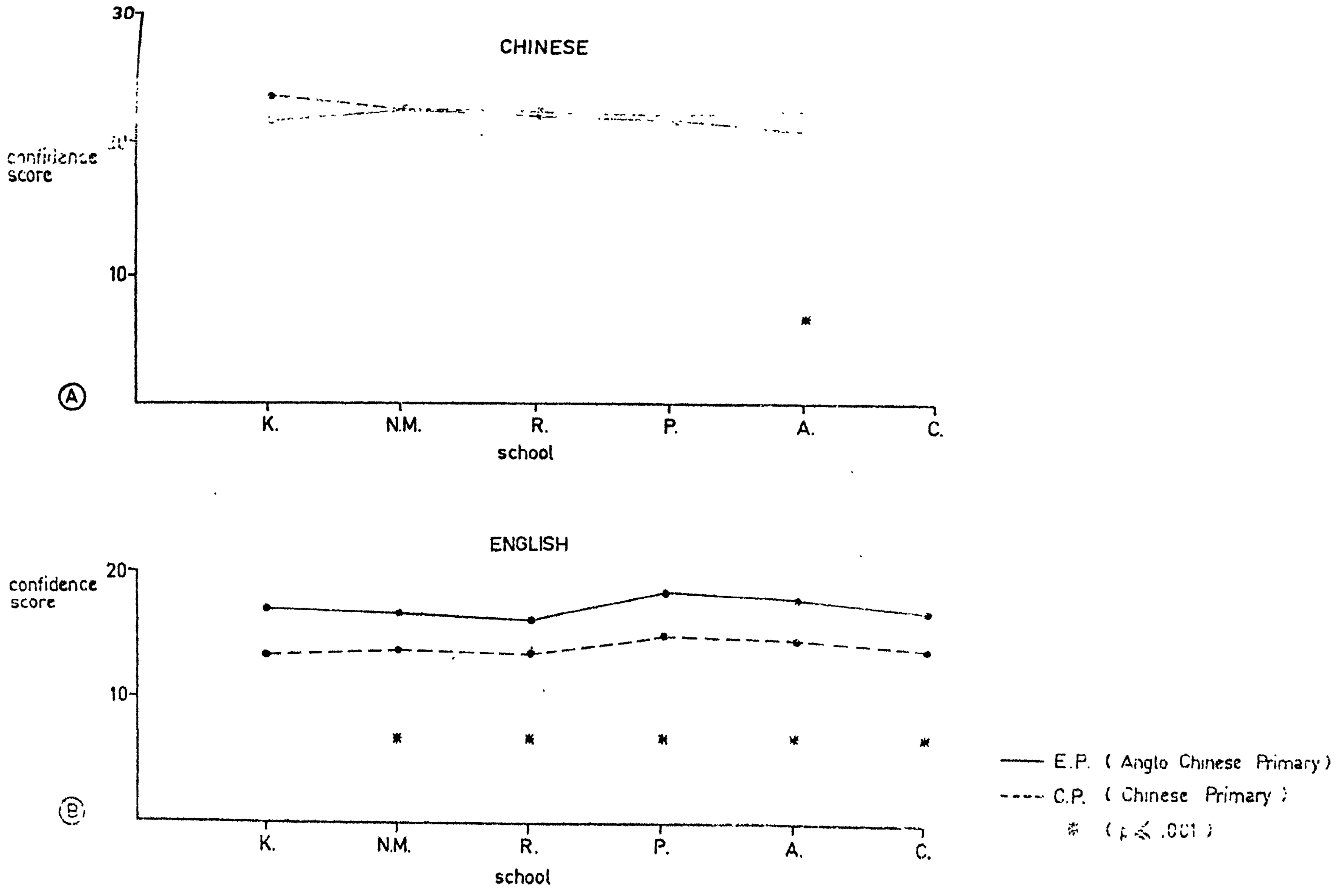


FIGURE 1 : STUDENTS' OVERALL CONFIDENCE IN LANGUAGE

TABLE 8

Language Competence of Students from Anglo-Chinese Primary and Chinese Primary Schools

School		Chinese		English		Chinese		English	
		E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K.						11.267	11.412	13.467	9.588
						(2.865)	(2.695)	(2.356)	(3.392)
						15	17	15	17
						-.147		3.706***	
N.M.	Mean	17.791	16.840	12.582	8.81				
	Std.	(2.931)	(3.722)	(3.727)	(3.855)				
	t	91	100	91	100				
		1.948		6.861***					
R.		19.936	20.381	16.157	14.50	12.575	11.946	12.0	9.919
		(2.953)	(3.177)	(3.605)	(4.175)	(2.178)	(2.788)	(3.093)	(2.793)
		110	42	108	42	80	37	80	37
		-.813		2.416		1.155		3.486***	
P.						12.772	12.803	13.293	11.931
						(2.694)	(2.766)	(3.587)	(3.473)
						92	142	92	144
						-.085		2.901	
A.		18.187	17.022	12.150	9.401				
		(3.701)	(3.880)	(4.302)	(4.761)				
		80	137	80	137				
		2.170		4.244***					
C.		19.866	20.377	17.863	16.736				
		(2.813)	(2.596)	(3.771)	(4.091)				
		119	53	117	53				
		-1.126		1.758					
(No. of Comparisons that are significant)		0		2		0		2	

*** (p<.001)

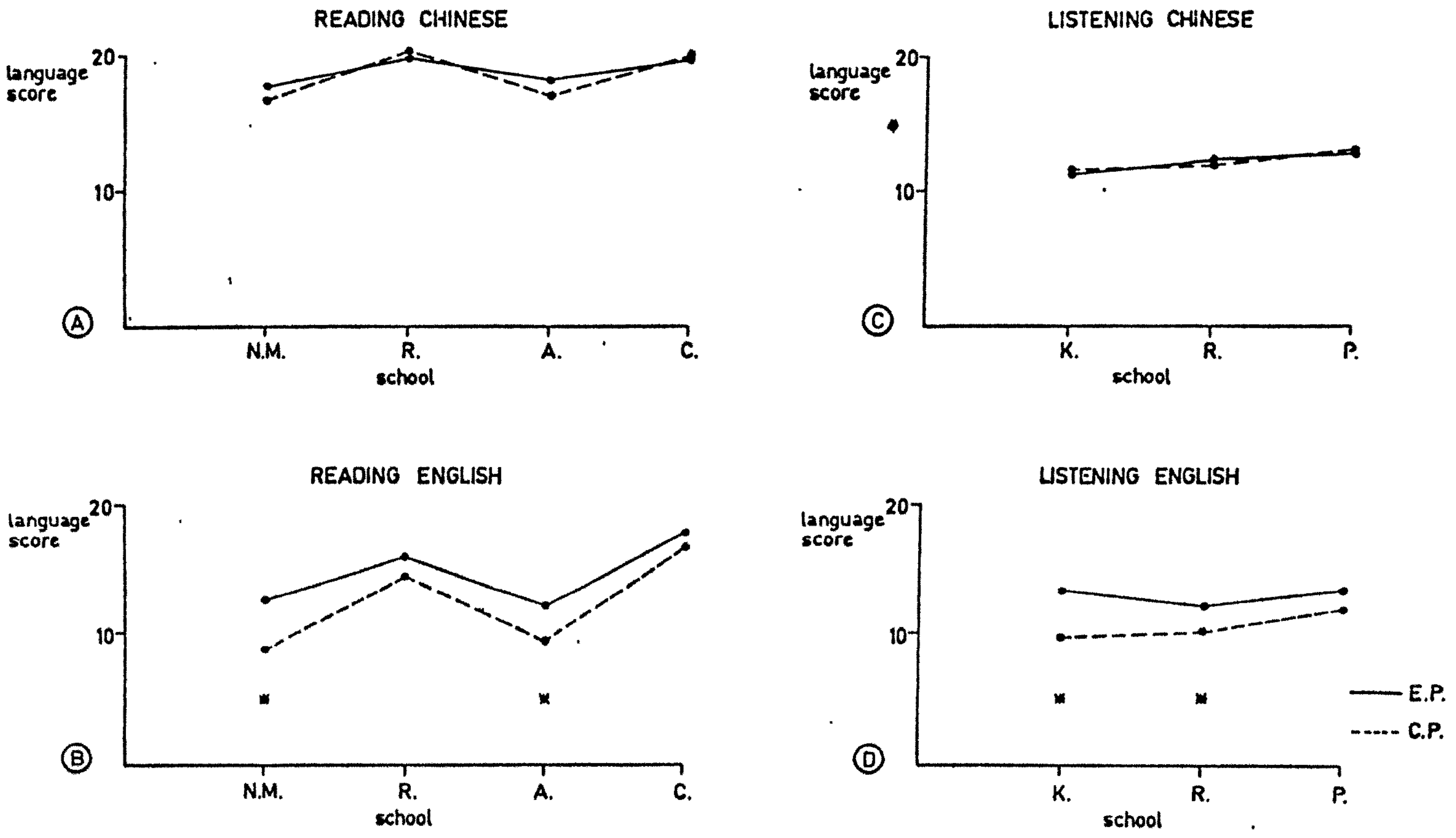


FIGURE 2 : LANGUAGE COMPETENCY OF STUDENTS FROM ANGLO-CHINESE PRIMARY AND CHINESE PRIMARY SCHOOLS

TABLE 9

Internal Examination Results of Students from Anglo-Chinese Primary and Chinese Primary Schools

		English Total		Western History		E.P.A.		Geography		General Science		Mathematics	
School		E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K	Mean Std. t	282.667 (50.845) 4.913***	176.824 (68.35)			(Social Studies) 73.667 (16.408) 4.059***	47.706 (19.380)			124.2 (21.462) 3.563	92.235 (28.281)	137.133 (41.614) .553	130. (31.087)
N.M.		5.556 (1.027) 5.914***	4.880 (.984)			(Social Studies) 5.410 (1.280) 4.224***	4.838 (1.017)			5.238 (1.343) 2.318	4.923 (.853)	5.017 (1.423) 1.128	4.846 (1.164)
R.		182.837 (25.402) 9.445***	143.127 (42.577)	58.889 (15.599) 6.825***	43.759 (18.682)	62.979 (16.121) 4.688***	51.722 (21.716)	63.158 (13.328) 6.677***	50.215 (16.951)	54.721 (14.67) 6.928***	40.329 (17.401)	132.621 (28.984) 2.34	122.873 (35.776)
P.		51.141 (12.808) 6.437***	40.431 (12.244)	49.043 (12.601) 4.782***	40.931 (12.778)	47.263 (13.129) 2.346	42.983 (12.266)	48.395 (12.039) 3.901***	41.488 (12.134)	49.185 (13.309) 4.603***	41.590 (11.722)	42.293 (13.314) -2.046	45.833 (12.544)
A.		(General English) 4.8 (1.174) 7.027***	3.610 (1.218)	4.287 (1.593) 5.82***	2.978 (1.598)	(Economics) 4.850 (1.592) 5.981***	3.412 (1.770)	4.575 (1.348) 6.534***	3.221 (1.538)	4.212 (1.209) 6.157***	3.081 (1.356)	4.8 (1.297) 2.749	4.294 (1.312)
C.		202.517 (30.779) 8.687***	156.472 (34.757)	64.714 (18.713) 3.539***	53.679 (19.266)	71.681 (11.24) 2.4	67.208 (11.394)	68.622 (13.825) 3.135	61.245 (15.164)	65.252 (17.344) 3.508***	54.792 (19.577)	65.918 (17.226) 1.184	62.585 (16.603)
(No. of Comparisons That are significant)		6		4		4		3		4		0	

***(p < .001)

TABLE 9 (Continued)

Internal Examination Results of Students from Anglo-Chinese Primary and Chinese Primary Schools

School		Chinese Total		Chinese History		Art		Physical Education		Conduct	
		E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K.	Mean Std. t	208.6 (30.211) .803	199.941 (30.611)			82.267 (7.401) .156	81.882 (6.566)	66.333 (4.791) 1.682	61.941 (9.045)		
N.M.		4.723 (1.334) -6.006***	5.556 (.978)	5.013 (1.247) -.892	5.137 (1.196)	4.975 (1.566) -2.618	5.402 (1.16)			7.456 (.671) -1.146	7.538 (.550)
R.		118.311 (12.22) -.262	118.747 (12.882)	53.089 (11.214) -.1	53.244 (12.169)	60.9 (6.805) -.524	61.367 (6.301)	2.968 (.423) -1.383	3.051 (.504)	6.437 (1.707) -.196	6.481 (1.616)
P.		(Chinese Test) 10.391 (3.268) -3.233	11.715 (2.934)			49.217 (7.95) -3.201	52.889 (8.981)	51.913 (8.593) -1.112	53.111 (7.722)	2.739 (1.3) 0.02	2.736 (1.044)
A.		(Chinese Language) 4.312 (.836) 2.357	4.015 (.927)	4.037 (1.335) 2.221	3.603 (1.416)	5.062 (1.194) 1.387	4.838 (1.117)	4.375 (.891) -.66	4.456 (.860)	6.9 (.851) .437	6.854 (.681)
C.		190.008 (31.05) -1.837	198.981 (25.804)	58.319 (16.295) -.457	59.547 (16.253)						
(No. of Comparisons that are significant)		-1		0		0		0		0	

***(p < .001)

(2) For subjects such as E.P.A., Geography, and General Science, which require, in addition to the command of the English language, other skills such as the interpretation of graphic materials, skills in mathematics and laboratory work, sixteen comparisons were made; and out of this number, eleven were significant in favour of the E.P. students.

(3) For the subject of mathematics, competence in which is the least affected by English language skills, and for the non-language related subjects such as Art, Physical Education, and Conduct, no significant difference between the two groups was detected in any of the nineteen comparisons.

(4) For subjects such as Chinese and Chinese History, which require a good command of the language of the other type, ie. Chinese, no significant difference was detected between the two groups, except in one case which favoured the C.P. students.

In conclusion, it seems that instruction in the medium of English has favoured E.P. students, particularly for subjects requiring a good command of the English language. The greater the amount of English required, the greater the advantage to E.P. students.

Language Understanding

The amount of understanding of the language used in teaching was measured on a five-point scale ranging from 1 to 5, to represent the anchors as described previously in the section on *Measuring Instruments*. The mid-point scale is '3' which means the understanding of 'About one half' of the language used. *Table 10* (pg.69) shows a comparison of the two groups of students in terms of the degree of understanding of the instructional medium as reported by the students. As expected, all the twelve comparisons were in favour of the E.P. students: out of this number, eight comparisons were significant at the 0.001 level. *Figure 3* (pg.70) shows clearly that the profile for the E.P. students to be consistently higher than that of the C.P. students across the different schools, the actual language mode of teaching was closely similar to that preferred by the students. Perhaps this suggests that teachers have flexibly adjusted the language mode to suit the standard and requirement of the majority of the students. *Figure 4* (pg.71) illustrates the overall preferences of the students as indicated in the bottom row of *Table 11* (pg. 72). In this figure, it can be seen that the majority of students preferred the E(c) mode for teaching and instructional materials. *Figure 4 (b)* (pg. 71) shows clearly that instructional materials in the C(e) or C mode are not preferred by the students participating in this study. While the students were making a choice of the language mode, they were also asked to indicate the reasons of their preference of the language. As the majority of students chose the E(c) mode to be used by the teacher and in the instructional materials,

TABLE 10

Students' Understanding of the Language Used in Teaching and in the Instructional Materials

School		Degree of Understanding of the Language Used in:			
		Teaching		Instructional Materials	
		E.P.	C.P.	E.P.	C.P.
K.	Mean	3.60	2.588	3.286	2.588
	Std.	(.910)	(.712)	(.611)	(.939)
	t	3.525		2.392	
N.M.		3.845	3.761	3.610	3.198
		(.868)	(.916)	(.788)	(.906)
		.842		4.366***	
R.		3.857	3.128	3.373	2.667
		(.960)	(1.21)	(.936)	(1.124)
		5.214***		5.255***	
P.		4.402	3.667	4.198	3.769
		(.647)	(.893)	(.582)	(.719)
		6.830***		4.781***	
A.		4.077	3.578	3.835	3.584
		(.864)	(1.003)	(1.055)	(.905)
		3.675***		1.846	
C.		4.449	4.019	4.359	3.830
		(.516)	(.635)	(.701)	(.612)
		4.683***		4.735***	
(No. of Comparisons that are significant)		4		4	

*** (p<.001)

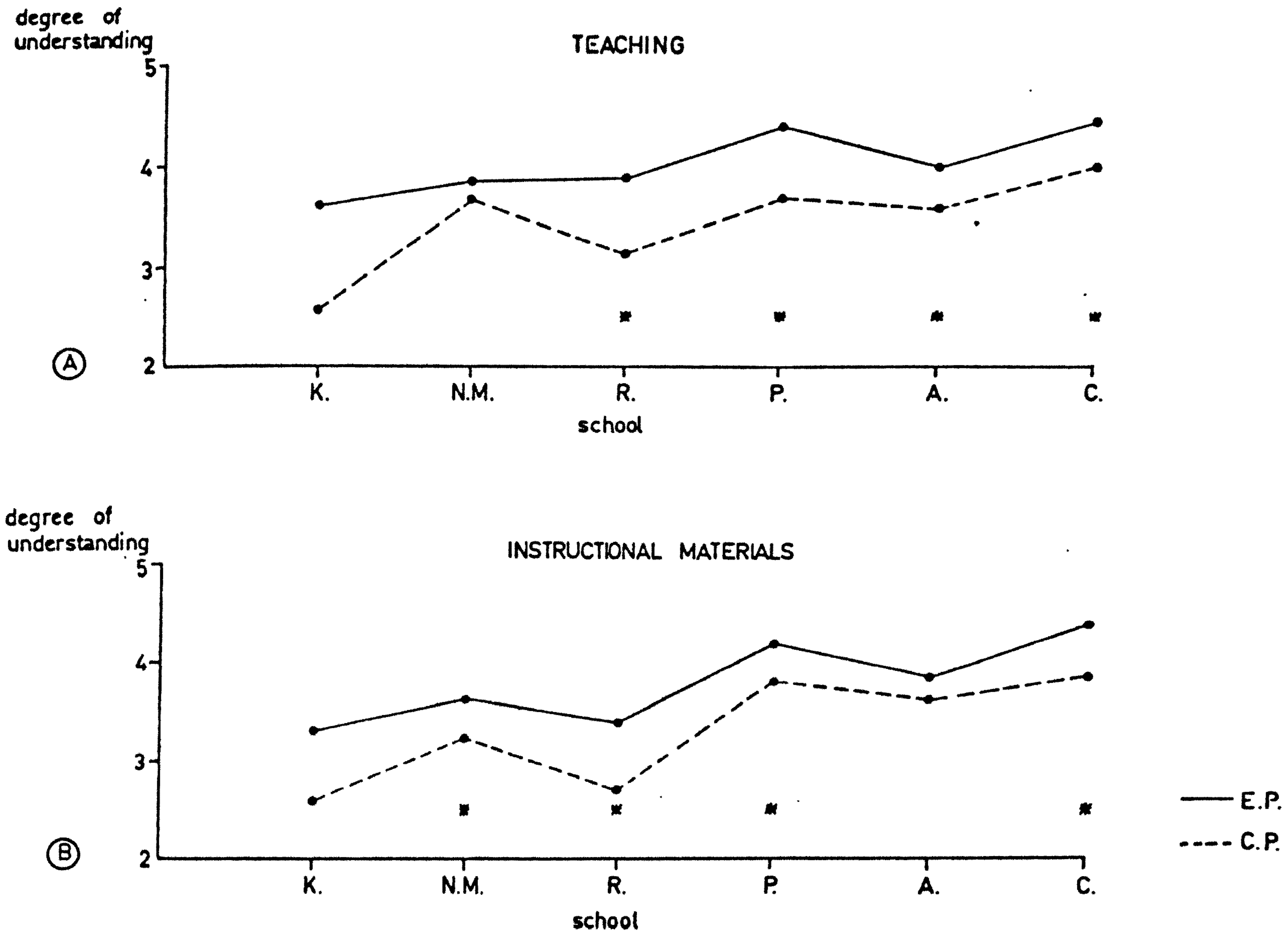


FIGURE 3 : STUDENTS' UNDERSTANDING OF THE LANGUAGE USED IN TEACHING AND IN THE INSTRUCTIONAL MATERIALS

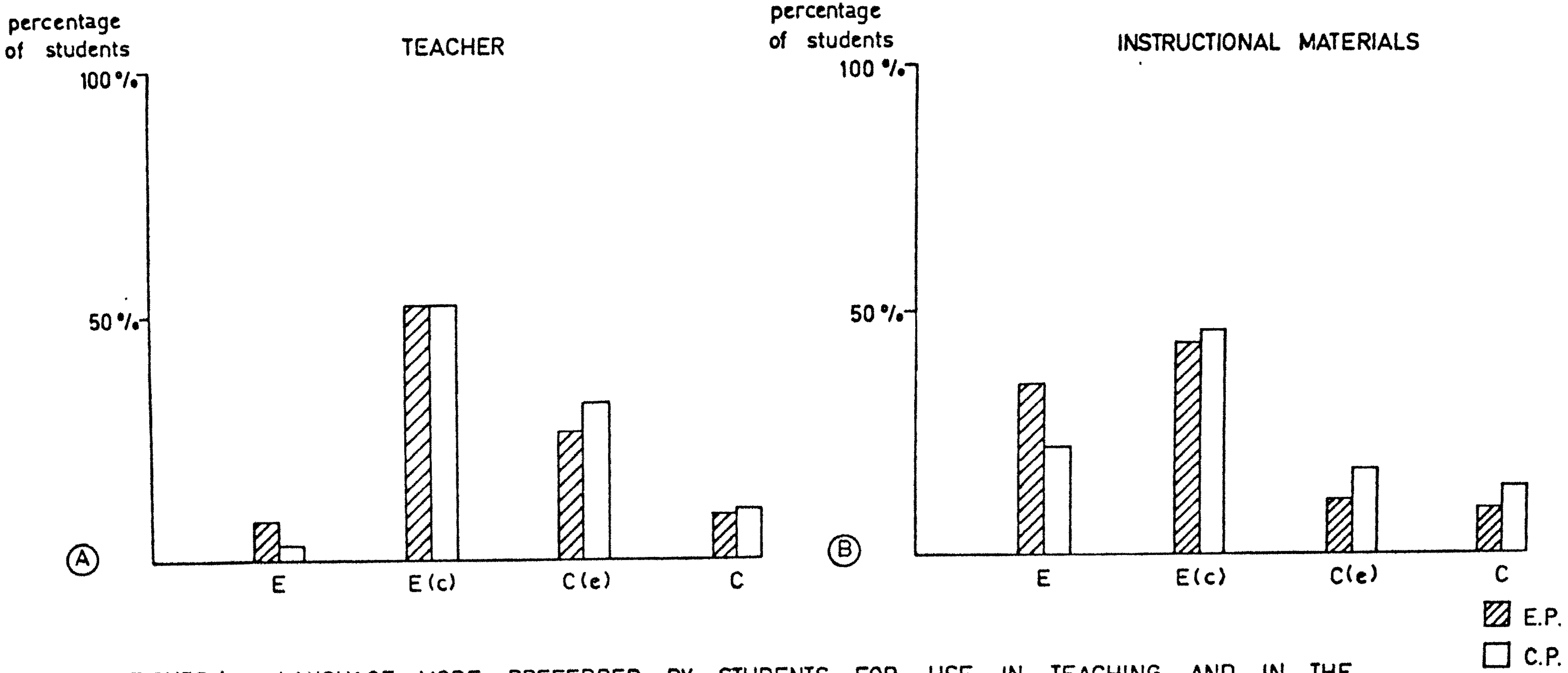


FIGURE 4 : LANGUAGE MODE PREFERRED BY STUDENTS FOR USE IN TEACHING AND IN THE INSTRUCTIONAL MATERIALS

TABLE 11

Percentage of Students Preferring the Language Mode to be used by the Teacher, Student, and in the Instructional Materials

School	Subject Rated	Student	Teacher				Student				Instructional Materials			
			E	E(c)	C(e)	C	E	E(c)	C(e)	C	E	E(c)	C(e)	C
K	Integrated Science	E.P.	13	#67	20		20	47	33		33	53	13	
		C.P.		47	47	6		24	53	24		53	24	24
N.M.	Mathematics	E.P.	8	34	#41	12	5	26	44	26	40	32	15	13
		C.P.	3	43	48	6	3	27	48	21	36	26	22	12
R.	History	E.P.	2	39	#34	11	4	22	46	27	17	46	19	12
		C.P.	1	20	41	18	3	15	42	41	10	33	24	32
P.	History	E.P.	14	#77	5	3	17	47	23	13	58	37	2	3
		C.P.	1	74	18	8	7	33	38	21	30	49	11	10
A.	Geography	E.P.	15	#60	15	9	19	49	19	14	34	51	4	1
		C.P.	9	50	28	12	7	39	28	24	15	60	15	9
C.	Economics & Public Affaires	E.P.	8	#82	3	4	21	49	11	18	38	58	1	4
		C.P.	2	70	23	6	6	55	28	11	17	62	13	8
Weighted Average Percentages		E.P.	8	52	26	9	11	34	33	21	35	43	11	9
		C.P.	3	52	32	10	5	33	37	24	22	46	17	14

#Mode of language actually used by the teacher in instruction.

the reasons for their selection are summarized in *Table 12* (pg.74). It can be seen that the majority of those students from both groups choosing the E(c) mode put down reasons '1', '2', '3', and '6' as the explanation of their decision. From the students' responses in this table, it seems that the E(c) mode is an intermediate step to help them to adapt to instruction in English. The majority of the students believed that, rightly or wrongly, instruction in the medium of English can help them to improve their English.

Although the students of both groups showed a preference for the E(c) language mode, it does not follow that students in this study perceived English as a more important language than Chinese. The data in *Table 13* (pg. 75) shows the relative importance of English and Chinese as preceived by the students. The numbers in this table are given as percentages. Over 60 percent of both groups of students pooled from all schools perceived English and Chinese as equally important. The computed x^2 value of 16.133 with 3 d.f. indicated that the two types of students are probably homogeneous with respect to the perception of the language of importance.

CONCLUSION

In the above, the E.P. and C.P. students within each school are compared under the five headings of family education, language confidence, language skills, internal examination results, and language preferences. Although teachers in these schools have adjusted the language mode to the standard of the students, results of this study provide strong evidence that the shifting of instructional medium has put the E.P. students in a favourable position compared with the C.P. students in the first year of secondary education in subjects requiring competence in English. The higher the degree of English language loading in a particular subject, the greater is the relative advantage for the E.P. students. For the subjects such as Mathematics, Chinese, Chinese History, Art, Physical Education, and Conduct, no convincing evidence is obtained that one group is significantly better than the other. Although an axiom of bilingual education is that the best medium of teaching is the mother tongue of the students (Saville and Troike, 1971), the kind of instructional medium the E.P. students experienced in the primary school has placed them in a favourable position at least in the first year of Anglo-Chinese secondary school where the instructional medium is essentially in English. In terms of family education, there is a greater proportion of E.P. to C.P. students progressing to or beyond the first degree level. However, the majority of families of both groups do not go beyond the Form V level. One might argue that the higher academic performance of the E.P. students is due to the fact that they have come from a selected group. This may be true but it is not easy to show the evidence. On the other hand, many of the Chinese primary schools are also very selective in the intake of students. In this study a strong assumption is made that, when E.P. and C.P.

TABLE 12

Percentages of Students Responding[1] to Each of the Reasons for Preferring the E(c) Mode to be Used by the Teacher and in the Instructional Materials

Reasons	Teacher			Instructional Materials		
	E.P.	C.P.	Z	E.P.	C.P.	Z
* 1. Cannot master English well, so the inclusion of some Chinese is needed.	59	68	-2.46	62	67	-1.43
* 2. Want to know the Chinese translation of some English terms as well.	50	40	2.75	55	52	.56
* 3. Can improve my English.	67	60	-.45	52	49	.81
4. My English standard is higher than that of my Chinese.	18	21	-1.06	11	10	.29
5. Used to learn mostly in English.	38	24	3.8***	29	15	4.17***
* 6. The inclusion of some Chinese helps me gradually to adapt to English instruction.	67	64	.82	66	68	-.61
7. The content matter of this subject is easy to understand.	3	4	-.69	4	6	-.81
8. English should be used in Anglo-Chinese schools.	22	18	1.36	19	21	-.72
9. The status of English is high, and English provides better career prospects.	27	30	-.85	21	26	-1.37
10. My friends and teacher think that English should be used, and I don't want to be different.	8	12	-1.70	5	5	-.22

[1] Each student marks not more than four responses

TABLE 13

The Language of Importance as Perceived by the Students

School	Mainly in English		Mainly in Chinese		Both English and Chinese		No Specific Indication	
	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.	E.P.	C.P.
K	27			6	73	82		12
N.M.	20	20	8	18	70	56	3	7
R.	6	9	8	30	57	52	4	5
B.	39	24	2	3	52	69	4	3
A	24	15	6	10	61	66	5	2
C	24	8	6	9	68	72	1	6
Weighted % of Total	20	16	6	13	63	64	3	4
Z-value	1.699		-3.698***		-.130		-1.026	
x^2 = 16.133 (d.f. = 3)								

*** (p<.001)

students are admitted to the same secondary school, both groups of students do not differ significantly in academic ability at the intake level. It must be admitted in conclusion that the truthfulness of the findings in this study depends on the degree of validity of this assumption.

REFERENCES

Campbell, E., Tracy, P. and McErlain, E. (1973) Primary Reading Assessment Units. The Ontario Institute for Studies in Education.

Engle, P.L. (1975) 'Language medium in early school years for minority language groups', Review of Educational Research. 45, No. 2, pp. 283-325.

Lado, R. (1964) Language testing. New York: McGraw-Hill.

Lambert, W.E. and Tucker, G.R. (1972) Bilingual education of children: the St. Lambert Experiment. Rowley, Mass.: Newbury House.

Macnamara, J. (1976) 'The bilingual's linguistic performance - a psychological overview', Journal of Social Issues. 23, 58-77.

Saville, M.R. and Troike, R.C. (1971) A handbook of bilingual education. Washinton, D.C.: Teachers of English to speakers of other languages.

Stern, H.H. (ed.) (1972) Foreign languages in primary education: the teaching of foreign or second language to younger children: Report on an international meeting of experts, April, 1962. International Studies in Education. Hamburg: UNESCO Institute for Studies in Education.

E.L.T.S: THE ENGLISH LANGUAGE TESTING SERVICE OF THE BRITISH COUNCIL AND THE UNIVERSITY OF CAMBRIDGE LOCAL EXAMINATIONS SYNDICATE

Peter Falvey
British Council
Hong Kong

1. Background

1.1 This battery of tests is designed to measure the language skills of overseas students who need to provide evidence of their English Language Proficiency. It has been developed jointly over the past five years by the British Council and the University of Cambridge Local Examinations Syndicate.

1.2 For a number of years before the development of the ELTS the British Council used the 'Davies' Test (produced by Alan Davies of the University of Edinburgh) in order to test the proficiency of students from overseas who were required by British universities to produce evidence of their ability in English. The Davies Test was designed primarily to assess the proficiency of potential postgraduate students. It was created to meet a demand from universities which arose because of a large increase in the number of students from overseas in the 1960's and 1970's. This growth reflected the emphasis given by developed countries to overseas aid which manifested itself in the awards of scholarships by the British Ministry of Overseas Development and various agencies of the United Nations (W.H.O.; F.A.O.; U.N.D.P.; I.L.O.; U.N.E.S.C.O.).

1.3 The 'Davies' Test which eventually reached Mark IV in Version 'D' was a reliable predictor of a student's ability to satisfy the language requirements of British universities for Postgraduate courses. It was not designed to predict the ability of candidates going on short courses, attachments, visits, working holidays, and courses which did not require such a high level of proficiency.

1.4 In the 1970's it was noticeable that more and more students were coming to Britain for undergraduate/further education/ technical courses particularly from oil-rich countries and from those countries where Overseas Aid was aimed at technician and technologist training rather than tertiary education/ training. The 'Davies' Test, which was not designed for this type of student, was not able to predict accurately the proficiency of these students in English.

2. Development of the ELTS

2.1 In order to cope with the growing demand for a testing instrument covering a wider range, the British Council asked Brendan J. Carrol[1], one of its officers, to begin designing a battery of tests. It was eventually agreed that the new series of tests, to be known as 'The English Language Testing Service', would be run under the auspices of the University of Cambridge Local Examinations Syndicate. The aim of the service is to provide a profile of each students' language ability which indicates:

a) how well that student will be able to meet the demands of any course of study or training in Britian, (or anywhere else where the teaching is done through the medium of English) and

b) those areas or skills in which further language training is needed.

2.2 After some years of development and piloting, the test was launched in 1979 and 1980. By 1980 the service was available in 48 countries and, by the end of 1981, it will be available in further 28 countries[2].

2.3 Brendan J. Carrol left the British Council in 1980 and, since then, responsibility for testing and liaison within the British Council has been under taken by the ELTS Liaison Unit, headed by Ian Seaton, advised by Alan Moller, Senior Consultant, English Language Services Department. The Unit is based in the English Language Division of the British Council (10, Spring Gardens, London SW1A 2BN). Cambridge University is represented by Dr. David Shoesmith, Director, Test Development and Research Unit of the Local Examinations Syndicate.

3. The Testing Service - Composition of the Test Battery

3.1 The current battery of tests assesses the competence of the candidate in the areas of language and study needed for courses of study[3]. The Test is made up of two main sections:

3.1.1 A General (G) section which tests the receptive skills of reading and listening, and reflects general ability in these skills in non subject specific areas.

3.1.2 A Modular (M) Section which tests study skills used in reading, writing, listening and speaking, and is related to a specific subject specific areas. Use is made in this section of authentic material from books and journals.

3.2 Modules are available for the following subject areas:

Life Sciences

Medicine

Physical Sciences

Social Studies

Technology

If no module is available or if the student is doubtful as to which module is closest to his own subject a General Academic Module is available[4].

3.3 The General Section consists of two parts, G1 and G2. G1 is a 40 minute, multiple-choice test of Reading and Usage. G2 lasts approximately 30 minutes and consists of multiple-choice questions on Listening Comprehension.

3.4 The Modular Section consists of three parts.

Module	Test	Time	Type
Module (M1)	Study Skills	55 mins	Multiple choice
Module (M2)	Continuous Writing	40 mins	Essay
Module (M3)	Oral Proficiency	10 mins	Interview

For each subject area there are three Modules. The modules are different for each of the five subject areas and for the General Academic Module.

3.5 The rationale behind the large number of sub-tests which include a structured, subject-related interview is that there will emerge, eventually, a language profile of the candidate. Candidates too react well to the battery because they feel that the tests are relevant - they have face validity. There is no pass or fail - no one off mark. The battery of tests will, it is hoped, gradually build up a picture of the student's proficiency in English and his ability to cope with whatever he is facing when he eventually reaches his destination.

4. A Description of the Tests[5]

The tests must be done in the order listed:

G1, G2, M1, M2, M3.

4.1 G1 Reading Test

The G1 Reading Test tests students ability to read and understand English. Answers are recorded on a multiple choice answer sheet. The texts and questions are given in booklet form. There are three sections to G1. An example is provided for each section (see 4.1.1 below).

4.1.1 Section One consists of 11 questions. Students have to select from 4 statements the one statement which is closest in meaning to a given sentence e.g.

Jack is taller than Jill

A. Jack is as tall as Jill.
B. Jill is shorter than Jack.
C. Jill is as tall as Jack.
D. Jack is shorter than Jill.

The questionsgrow in difficulty and cover a variety of syntactic items and items testing inferential meaning.

4.1.2 Section Two consists of a cloze test with 24 cloze items. For each item the student selects from four possible choices.

4.1.3 Section Three consists of three newspaper articles, each article dealing with the same event. Specific fact finding multiple choice reading comprehension questions are asked after each passage; 4 after passage A, 3 after passage B and 4 after passage C. After the 3 passages and multiple choice questions have been dealt with the candidate is required to answer 5 further questions which test his ability to search for information amongst the 3 passages he has just read and to be able to discrimate between the way information is presented in each passage.

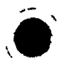
4.2 G2 Listening Test

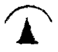
G2 tests students' ability to understand spoken English and consists of four sections.

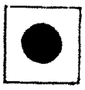
4.2.1 Section One consists of ten questions. Each question refers to a diagram. After each question the student has to select one of four diagrams, e.g.

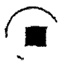
Student hears: Which shape consists of a circle with a square in it?

He sees:









The questions become more difficult as the test progresses

4.2.2 Section Two consists of an interview. The students are warned that they will hear the interview only once and that they will then be required to answer questions on what they have heard. After listening to a short interview they are given 6 multiple-choice questions to answer involving awareness of facts, inference, drawing conclusions and assessment of attitude from tone of voice.

4.2.3 Section Three consists of 10 questions given on the tape-recorder. For each question posed one of four possible replies is selected. The student must choose the one he thinks is most appropriate. This sub-test tests the student's ability to provide coherent replies to spoken discourse. E.g.

He hears: "Where are you going now?"

He reads: A. I'm going there next week.
B. I thought I might call to see Jim and May.
C. Yes, indeed.
D. I've been to the shops.

4.2.4 Section Four tests the listening skills needed for participation in a seminar. The candidate hears a seminar discussion between three people in the course of which plans are referred to. The discussion is in two parts. After the first part there is a pause during which the candidate answers four multiple-choice questions. After the second part five multiple-choice questions have to be answered which are referred to in the discussion. The candidate is provided with plans to which he may refer while listening to the tape.

4.3 Modular Tests

For each Modular Test the student is provided with a Source Booklet for the subject he has chosen. Each Source Booklet consists of authentic text, diagrams, pictures, graphs, an index, a glossary and a bibligraphy. They are, in fact, mini textbooks of between 10-13 pages.

4.3.1 Modular Test 1 (M1)

The M1 Study Skills Test requires the candidate to answer 40 multiple-choice questions in 55 minutes.

In order to answer the question the candidate must read the instructions in his question booklet, read the section referred to in the source booklet, and answer the relevant questions on his multiple-choice answer sheet.

The source booklets are usually divided into 4 or 5 sections together with glossaries, indices and booklists. The questions deal with the various sections one by one and require, in addition to the ability to comprehend written text, the ability to understand graphs and diagrams; infer meaning from pictures; search for information in glossaries and indices; and make bibligraphical searches. In addition, at the end of the question booklet there are general questions which test the same ability as that tested in Section 3 of the G1 Reading Test - namely the ability to make post-referential searches for information.

4.3.2 Modular Test 2 (M2)

The M2 Writing Test takes 40 minutes to administer and consists of two questions, both of which must be answered. Each question refers to a section of the Source Booklet. After reading the first question and referring to the relevant section, which may be a written text, a series of graphs, or some diagrams, the candidate is required to give opinions, discuss advantages and disadvantages, defend an argument, argue against a point, or give reasons for a choice.

The second question again requires reference to the Source Booklet. It tests the candidate's ability to describe processes, outline reasons, describe causes or provide factual narrative.

The writing test thus provides an opportunity for the candidate to display his skill both in narrative and explanatory writing as well as in exposition and discussion.

4.3.3 Modular Test 3 (M3)

The M3 Test consists of an interview which lasts approximately 10 minutes.

Both the interviewer and the candidate refer to their own copies of the Source Booklet.

The interview is structured and consists of three separate phases although the interviewer is instructed to try to keep the interview flowing. After each section the interviewer unobtrusively makes an assessment. After phase one, a broad assessment is made. This is narrowed in phase two and a final decision is made after stage three.

Phase One - Introduction is the 'settling-down', 'putting at ease' phase which lasts two or three minutes. The interviewer asks personal and professional questions designed to draw out the candidate and relax him. Questions about the weather, recent events, the region, the family are asked and the candidate is given some indication of how the interview will develop. During this phase, the interviewer circles three of the nine bands 1-9.

Phase Two This section, which centres around the Source Booklet, is introduced naturally, usually by a question about the candidates feelings about the test, e.g. which sections were easy, difficult, familiar or unfamiliar. Eventually the interviewer asks the candidate to refer to his Source Booklet and questions are asked to elicit from the candidate detailed descriptions of a process or operation which he has read in the Source Booklet which is open in front of him. The student's ability to be explicit, to describe, explain and clarify details in English is being assessed here and towards the end of this phase (about 4 minutes) the interviewer narrows the banding from three to two bands.

Phase Three - Extended Dialogue is introduced by moving from the Source Booklet to questions about the student's plans for his future and the relevance of his former studies to his future plans. This phase (3/4 minutes) is designed to allow the candidate to speak at length. The interviewer is mainly an auditor, interrupting only to force the candidate to refer back to a particular point which he is then asked to exemplify or explain or justify.

At the end of Phase 3 the interviewer narrows the two band assessment of Phase 2 to a final band.

5. Scoring

5.1 Multiple Choice Tests G1, G2 and M1

These tests are scored by using a template and the raw scores are then converted, by conversion tables, to bands. (The bands will be analysed and discussed below).

5.2 M2 Writing Test

This test is marked subjectively in two sections. The marker has available for reference a brief performance description for each band together with samples of previously assessed writing (which match the band ratings) against which he can compare his own judgement.

e.g. Band 7 Good Writer

> 'Matter presented in a well-ordered, intelligible manner with well-structured main and subordinate themes and relevant supporting detail. Generally accurate in language and appropriate in style, but with more frequent lapses than the Band 8 performer in accuracy, relevance and style. The reader has the impression of a functionally efficient writer.'

5.2.1 In assessing question 1 the marker decides on a whole band only after comparing the answer with the Scale (Band) and the written samples provided.

5.2.2 Question 2 of Module 2 is used to modify the first assessment. Question 1 is of a more descriptive nature while question 2 tests accuracy, correctness of English and the communicative value of the writing. In modifying the assessment of question 1 the band may not be moved up or down by more than <u>half a band</u>.

e.g.		Question 1	- Band 6
	if -	Question 2 (worse)	- Band 5.5
	if -	Question 2 (better)	- Band 6.5

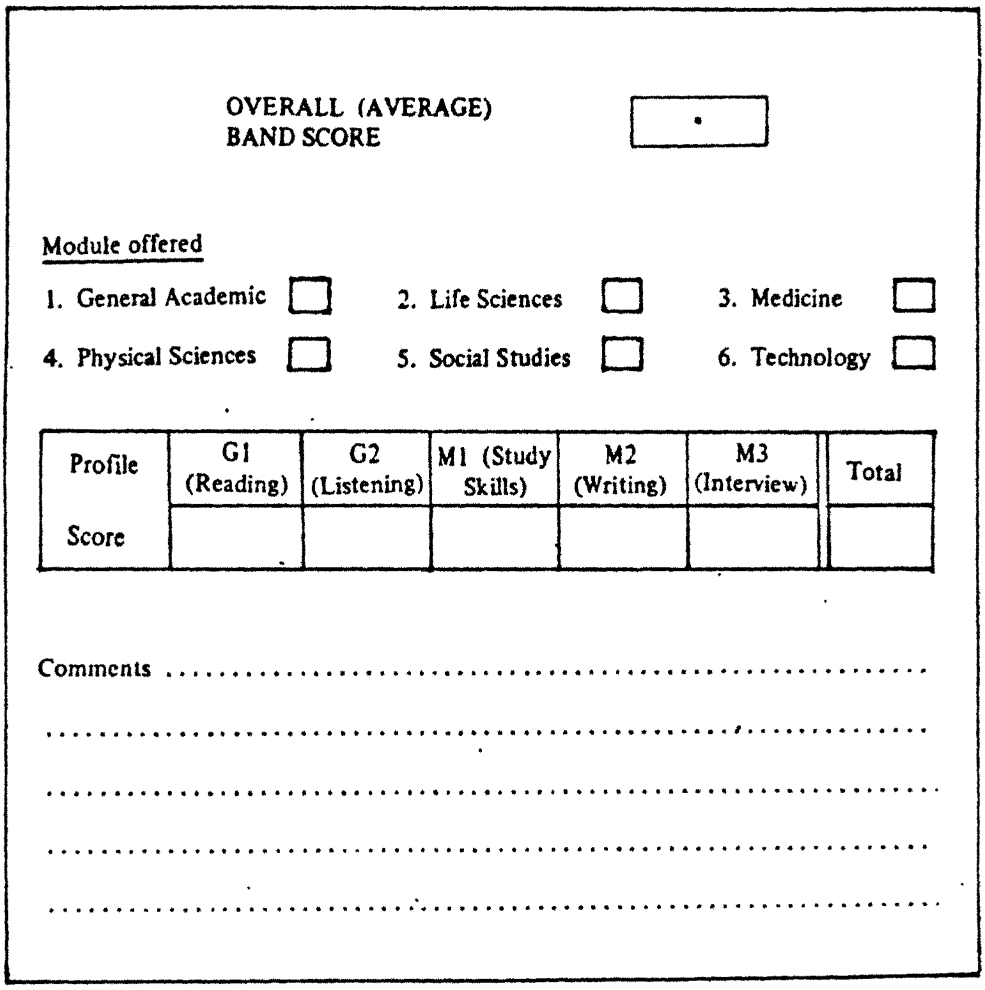
5.3 <u>M3 Interview</u>

Scoring for M3 has already been discussed in 4.3.3 above.

6. Presentation of Results (Profile and Banding)

6.1 After all the tests have been scored it is possible to present the candidate's language profile by means of a Test Report form which presents his results and notes for Band Interpretation.

The profile and test results are presented as follows:



OVERALL (AVERAGE) BAND SCORE □

<u>Module offered</u>

1. General Academic □ 2. Life Sciences □ 3. Medicine □

4. Physical Sciences □ 5. Social Studies □ 6. Technology □

Profile	G1 (Reading)	G2 (Listening)	M1 (Study Skills)	M2 (Writing)	M3 (Interview)	Total
Score						

Comments ....

....

....

....

....

6.2 The creation of the 'profile' enables the recipient of the report to see where the student's strengths and weaknesses lie. The profile provides a wider perspective than, say a TOEFL score of 502 or a 'Davies' Test score of 42.3 as it covers a variety of language skills many of which focus on the student's speciality.

7. Reliability and Validity of the ELTS

Three areas are currently being researched to increase the internal reliability, content validity and predictive ability of the ELTS.

7.1 Internal reliability

The following studies are currently being undertaken:

7.1.1 a study of the mechanical accuracy of scoring tests and of calculating and recording bands.

7.1.2 a re-marking study of the writing test M2 in order to evaluate the reliability of marking and to refine any judgements about standards.

7.1.3 a standard and routine item analysis of the tests.

7.1.4 a correlational analysis to see to what extent the various elements appear to be measuring the same or different traits.

7.1.5 a study of standards.

7.2 Content validity (Internal and External Studies)

The texts and items in the two General tests and each Modular Study Skills test are being redescribed in terms of language skills and the relation of that description to earlier skill descriptions in 1978 and the original specifications of 1977. This study will contribute to the control of the continuing process of specification, test realisation, validation and respecification.

7.3 External Studies - validation of profile (predictive validity)

All candidates who were tested in the first two years of the test will be followed up with questionnaire which will gather information in order to validate:

band profile

tuition hours prediction

placement target levels

These questionnaires will be answered by universities, language schools and British Council departments responsible for handling overseas students.

APPENDIX A

BRITISH COUNCIL/CAMBRIDGE UNIVERSITY EXAMINATIONS SYNDICATE

English Language Testing Service (ELTS)

1. ELTS is at present available in the following 48 countries:

Algeria
Argentina
Australia
Austria
Bahrain
Bangladesh
Belgium
Brazil
Burma
Colombia
Cyprus
Eduador
Ethiopia
Egypt
Finland
France
Greece
Holland
Hong Kong
Indonesia
Iraq
Israel
Italy
Japan
Jordan
Kenya
Kuwait
Malaysia
Mexico
Morocco
Nepal
New Zealand
Oman
Pakistan
Peru
Philippines
Portugal
Qatar
Saudi Arabia
Singapore
Spain
Sri Lanka
Sudan
Sweden
Thailand
Turkey
Venezuela
West Germany

2. In April 1981 it will be available in an additional 28 countries. The introduction of ELTS overseas has been a staged one and in the case of some of the following countries ELTS will be available before the spring of 1981:

Botswana
Cameroon
Canada
Chile
China
Denmark
Ghana
India
Korea
Lebanon
Lesotho
Malawi
Nigeria
Norway
Poland
Senegal
Sierra Leone
South Africa
Swaziland
Syria
Tanzania
Tunisia
United Arab Emirates
Yemen
Yugoslavia
Zaire
Aambia
Zimbabwe

3. In addition, in the UK, it is available at present in the Birmingham, Bristol, Cambridge, Cardiff, Edinburgh, Leeds, London, Manchester, Newcastle and Oxford offices of the British Council.

APPENDIX B

What the Test Contains

GENERAL SECTION

G1 (Reading)

Time allowed: 40 minutes.

Questions: about 40. All are multiple-choice, i.e. for each question you will be given a number of possible answers marked A, B, C, D. You are asked to choose the one that you think is the *best* answer and mark your choice on the Answer Sheet (see page 6).

G1 has a number of sections, each of which contains one type of question. You will be told in the test exactly what you have to do for each type.

EXAMPLE
You may have to read a sentence and then choose from four other sentences the one that is closest in meaning to it.

Choose the sentence A, B, C or D which is closest in meaning to the sentence on the left.
A lorry is larger than a motor cycle.

A Lorries and motor cycles are the same size.
B Motor cycles are larger than lorries.
C A lorry is as big as a motor cycle.
D Lorries are bigger than motor cycles.

Correct Answer D

EXAMPLE
In another section you may have to read a passage. However, some words have been left out and their positions are marked by short dotted lines. You have to choose the word which you think best fills each gap.

Read this passage, and for each of the numbered blanks choose the one word that best fills the gap.

The real test of a news story is the effect upon the reader. If he feels he is getting something new and fresh or something
1 interesting and different then.....is satisfied. This means clear crisp reporting, a story that moves and is efficiently told,
2 ... keeps to the point and carries the reader along without wasting his time and
3 .....attention.

A	B	C	D
she	he	him	it
though	that	however	so
many	the	much	his

Correct Answers 1B 2B 3D

EXAMPLE
In another section you may have to read short passages, all about the same subject but by different writers with different views and styles of writing. The questions then test your understanding of the passages and the differences between them.

G2 (Listening)

Time allowed: 30 minutes.

Questions. about 35. All are multiple choice.

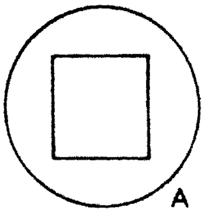
G2 has a number of sections, all of which are recorded on tape. As you listen, you will have in front of you a booklet which contains the possible answers. When you hear each question you have to choose the best answer from the booklet and mark your choice on the Answer Sheet (see page 6). You will hear the tape *once only*.

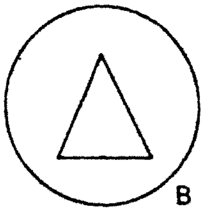
Each section contains one type of question. You will be told in the test exactly what you have to do for each type.

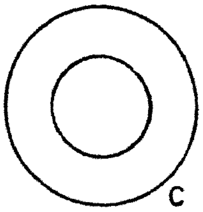
EXAMPLE
In one section you may hear a series of descriptions which you are asked to match up with diagrams printed in the booklet.

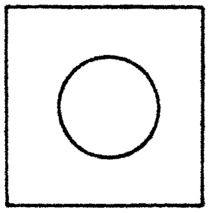
(You hear on the tape) 'Which figure consists of a square with a circle inside it?'

(You see in your booklet)


A


B


C


D

Correct Answer D

EXAMPLE
In another section you may hear a series of questions and you have to choose the best reply.

(You hear on the tape) 'What are you doing now?'

(You read in your booklet)
A I'm looking for my passport.
B I'm going there next week.
C It's a very large dog.
D Yes indeed.

Correct Answer A

EXAMPLE
In another section you may hear extracts from an interview or conversation between two people or parts of a seminar in which the voices of several people will be heard. The questions will test how well you have understood the main points, information, comments, etc of the discussion.

MODULAR SECTION

M1 (Study Skills)

Time allowed: 55 minutes.

Questions: about 40. All are multiple-choice.

You will be given a Source Booklet containing texts taken from books, journals, reports, etc related to the specific subject area you have chosen. The booklets also contain the kind of features to be expected in study texts: contents pages, bibliographies, appendices (where necessary), indices, etc. You will also be given a Question Booklet and Answer Sheet.

Here is an example of the kind of text you could read in a Source Booklet (in this case General Academic). This is followed by examples of questions about the text that might appear in the Question Booklet.

EXAMPLE

Section 3: Education and Society

The urgency of the tasks related to the expansion of educational systems has led planners to concentrate mainly on the quantitative aspects of the effort required, of the obstacles to be overcome and the disequilibria to be corrected. But we are becoming aware that the essential questions — questions of substance — concern the relationships between education and society, education and the learner, education and knowledge, between aims avowed and aims achieved.

Education is both a world in itself and a reflection of the world at large. It is subject to society, while contributing to its goals, and in particular it helps society to mobilise its productive energies by ensuring that human resources are developed. In a more general way, it necessarily has an influence on the environmental conditions to which it is at the same time subjected, even if only by the knowledge about these which it yields. Thus, education contributes to bringing about the objective conditions of its own transformation and progress.

When history is seen in sufficient perspective, there is no reason why this dialectical process — and this optimistic view of it — should not be quite clear, with its relationships of cause to effect and effect to cause.

Section 3: Education and Society

Read quickly through Section 3 in the Source Booklet and then answer these questions.

14 The writer's attitude, in the first paragraph, is that he
- A agrees with the planners' main point.
- B condemns the planners for their attitude.
- C considers that the planners' disequilibria must be corrected.
- D places the emphasis on different issues from the planners.

15 The word 'its' (at the end of line 8) refers to
- A the world at large.
- B society.
- C education.
- D a reflection.

16 The last sentence (lines 11-13) of the second paragraph serves as
- A a summary.
- B an illustration.
- C an explanation.
- D a counter-argument.

17 The words 'this dialectical process' (lines 14-15) primarily concern
- A the progress of education.
- B the objective conditions.
- C a sufficient perspective.
- D environmental conditions.

Correct Answers 14D 15B 16A 17A

APPENDIX C

BAND INTERPRETATION

BAND 9 EXPERT USER:

Fully functional command of the language.

BAND 8 VER GOOD USER:

High level of mastery. Possibly occasional inaccuracies or inappropriate usage.

BAND 7 GOOD USER:

Generally accurate and appropriate use of language, but more frequent lapses than for 8 and 9.

BAND 6 COMPETENT USER:

Reasonably good communicator, although somewhat deficient in fluency and with occasional blocks to communication.

BAND 5 MODEST USER:

Conveys and understands gist of the topic, but may lack clarity, with several inaccuracies and inappropriate usage. Lacks interest and variety.

BAND 4 MARGINAL USER:

Frequent lack of clarity, with inaccuracies of both vocabulary and grammar. Frequent lapses in style. Barely competent as communicator.

BAND 3 EXTREMELY LIMITED USER:

Below level of functional competence so that one has to strain to make out his message. Constant blocks in communication.

BAND 2 INTERMITTENT USER:

Unable to communicate effectively although message comes through intermittently. Constant strain for reader/listener to make out the message.

BAND 1 NON-USER:

Either has little or no knowledge of the language, or does not provide sufficient evidence for assessing his language competence.

BAND 0 Insufficient response for assessment to be made. Reason stated in Comments section.

NOTES

(1) See Brendan J. Carrol Testing Communicative Performance Pergamon 1980.

(2) See Appendix A for a list of countries.

(3) It has been pointed out that a 'non-academic' model is needed for those going on attachments and visits. The Service hopes that such a module will be available in the Autumn of 1981.

(4) The General Academic Module, together with Test Gl is often used, for short-stay professional visitors to the UK. This has replaced the 'Subjective' Test of the 'Davies' Test which consisted of an interview only.

(5) Appendix B contains further samples of Tests Gl, G2 and Ml, taken from the "user Handbook" of the English Language Testing Service. The handbook is available at all British Council offices where the test is held and can be requested at offices where the test has not yet been run.

LANGUAGE CENTRE ON-GOING ACTIVITIES IN LANGUAGE TESTING

Project on:

I. Placement and Language Proficiency Tests -

  a) for students studying in the Faculty of Arts (co-ordinator: Dr. Angela Fok)

  b) for students studying in the Faculty of Medicine (co-ordinators: Dr. Angela Fok and Lee Yick Pang)

  c) for students studying in the Faculty of Engineering (co-ordinators: Graham Low and Lee Yick Pang)

II. Self-assessment and Achievement - Peter Pan and Peggy Ng.

III. Theoretical and Practical Aspects of Second Language Proficiency Testing - Lee Yick Pang.

Recent Publications and Presentations

Fok Chan, A.Y.Y. (1981) 'The Testing of Listening Proficiency at the Tertiary Level', in John A.S. Read (ed.) Papers on Language Testing, Occasional Papers No. 18, SEAMEO Regional Language Centre, Singapore, 36-40.

Lee Y.P. (1981) 'Evaluation and Measurement of Communicative Competence: The Case for Direct Language Tests', in John A.S. Read (ed.) Papers on Language Testing, Occasional Papers No. 18, SEAMEO Regional Language Centre, Singapore, 86-98.

Lee Y.P. & Low G. (1981) 'Classifying Tests of Language Use', Paper to be read at the International Conference in Applied Linguistics. Lund, Sweden, July.

UNIVERSITY LANGUAGE RESEARCH

UNIVERSITY OF BIRMINGHAM

Department of English Language and Literature

MA

APPLIED ENGLISH LINGUISTICS

A two-year sandwich course for experienced teachers of EFL (and in particular ESP), combining work in-post with one term's intensive study at Birmingham each year. Ideal for those able to obtain short-term release from EFL institutions at home and abroad.

B PHIL (ED)

TEACHING ENGLISH AS A FOREIGN LANGUAGE

A one-year full-time post-experience advanced degree course in Applied Linguistics. Open to both graduate and non-graduate certificated teachers of EFL.

ENGLISH LANGUAGE RESEARCH JOURNAL

The Journal is concerned with English Language teaching, learning and research, and covers a wide variety of topics including ESP, discourse analysis, language planning and syllabus/materials design. Articles, reviews, research reports and news items are invited.

Further information on courses, and Journal subscription details from: Mrs Anne Preston, Secretary, English Language Research, University of Birmingham, PO Box 363, Birmingham B15 2TT.