The Construction and Validation of a Performance-based Battery of English Language Progress Tests

Thesis submitted by Michael Milanovic to the Institute of Education, University of London, for the degree of Doctor of Philosophy.

Abstract

The Construction and Validation of a Performance-based Battery of English Language Progress Tests

This thesis is concerned with the construction and validation of a battery of performance-based English Language progress tests.

The work is set in the context of a language teaching institute. The relationship of teaching and testing and the need for a greater degree of integration of the two as well as a greater focus on the training of teachers in testing principles and methods is a primary concern in the first part of the thesis. This is followed by a discussion of the strategies employed to overcome these problems, and a detailed description of the procedures adopted in the design of the test battery.

A hierarchical approach to the construct validation of performance-based tests is then proposed and the results of the validation procedures adopted discussed in detail.

Acknowledgements

It would not have been possible to complete this thesis without the help, encouragement and friendship of many people, and the support of the British Council.

Firstly, I would like to thank my supervisor, Peter Skehan, for his commitment and ability to inspire confidence.

Secondly, I would like to thank Peter Falvey for his vision and support, Margaret Falvey for the many hours of inspiring discussion that helped me to formulate ideas and Rex King for passing on to me some of his vast store of knowledge and experience as well as positive attitudes to testing.

Thirdly, I would like to thank all the teachers and staff at the British Council in Hong Kong for their help and willingness to understand what I was trying to do.

Finally, I would like to thank Janet Joyce, without whose example and endless encouragement I would never have been able to finish this work.

Table A7.11. B3 Demography - Occupation..... 622
Table A7.12. B3 Demography - Education..... 623
Table A7.13. C1 Demography - Sex..... 624
Table A7.14. C1 Demography - Occupation..... 624
Table A7.15. C1 Demography - Education..... 625
Table A7.16. C2 Demography - Sex..... 626
Table A7.17. C2 Demography - Occupation..... 626
Table A7.18. C2 Demography - Education..... 627
Table A7.19. C3 Demography - Sex..... 627
Table A7.20. C3 Demography - Occupation..... 628
Table A7.21. C3 Demography - Education..... 628
Table A9.1. A3 Item Statistics (Listening)..... 644
Table A9.2. A3 Item Statistics (Listening)..... 646
Table A9.3. A3 Item Statistics (Listening)..... 647
Table A9.4. A3 Item Statistics (Listening)..... 649
Table A9.5. A3 Item Statistics (Listening)..... 650
Table A9.6. A3 Item Statistics (Listening)..... 651
Table A9.7. A3 Item Statistics (Appropriacy).... 653
Table A9.8. A3 Item Statistics (Read/Write)..... 654
Table A9.9. A3 Item Statistics (Read/Write)..... 655
Table A9.10. A3 Item Statistics (Read/Write)..... 656
Table A9.11. B1 Item Statistics (Listening)..... 658
Table A9.12. B1 Item Statistics (Grammar)..... 662
Table A9.13. B1 Item Statistics (Appropriacy).... 666
Table A9.14. B1 Item Statistics (Read/Write)..... 669
Table A9.15. B2 Item Statistics (Listening)..... 673
Table A9.16. B2 Item Statistics (Grammar)..... 676

Table A9.17. B2 Item Statistics (Appropriacy).... 679
Table A9.18. B2 Item Statistics (Read/Write)..... 682
Table A9.19. B3 Item Statistics (Listening)..... 684
Table A9.20. B3 Item Statistics (Grammar)..... 686
Table A9.21. B3 Item Statistics (Appropriacy).... 689
Table A9.22. B3 Item Statistics (Read/Write)..... 692
Table A9.23. C1 Item Statistics (Listening)..... 695
Table A9.24. C1 Item Statistics (Grammar)..... 697
Table A9.25. C1 Item Statistics (Appropriacy).... 700
Table A9.26. C1 Item Statistics (Read/Write)..... 702
Table A9.27. C2 Item Statistics (Listening)..... 705
Table A9.28. C2 Item Statistics (Grammar)..... 706
Table A9.29. C2 Item Statistics (Appropriacy).... 709
Table A9.30. C2 Item Statistics (Read/Write)..... 711
Table A9.31. C3 Item Statistics (Listening)..... 713
Table A9.32. C3 Item Statistics (Grammar)..... 715
Table A9.33. C3 Item Statistics (Appropriacy).... 718
Table A9.34. C3 Item Statistics (Read)..... 720
Table A9.35. C3 Item Statistics (Write)..... 720
Table A9.36. B2 General Correlations..... 725
Table A9.37. B2 Task Correlations..... 726
Table A9.38. B3 General Correlations..... 727
Table A9.39. B3 Task Correlations..... 729
Table A9.40. C2 General Correlations..... 730
Table A9.41. C2 Task Correlations..... 731
Table A9.42. C3 General Correlations..... 733
Table A9.43. C3 Task Correlations..... 733

Table A12.1. Percentages of Variance (A3 Whole Test)..... 765

Table A12.2. Rotated Factor Matrix (A3 Whole Test)..... 775

Table A12.3. Percentages of Variance (B1 Whole Test)..... 776

Table A12.4. Rotated Factor Matrix (B1 Whole Test)..... 777

Table A12.5. Percentages of Variance (B2 Whole Test)..... 767

Table A12.6. Rotated Factor Matrix (B2 Whole Test)..... 779

Table A12.7. Percentages of Variance (B3 Whole Test)..... 769

Table A12.8. Rotated Factor Matrix (B3 Whole Test)..... 781

Table A12.9. Percentages of Variance (C1 Whole Test)..... 771

Table A12.10. Rotated Factor Matrix (C1 Whole Test)..... 784

Table A12.11. Percentages of Variance (C2 Whole Test)..... 772

Table A12.12. Rotated Factor Matrix (C2 Whole Test)..... 787

Table A12.13. Percentages of Variance (C3 Whole Test)..... 774

Table A12.14. Rotated Factor Matrix (C3 Whole Test)..... 789

Chapter I

1. Background

The work described in this thesis took place after the British Council made the unprecedented decision, in 1981, to employ a Testing and Evaluation Officer for the Direct Teaching of English Operation (DTEO) in Hong Kong. The officer held the post from January 1982 till September 1985 and the test battery described in this thesis was developed during this period.

The institute in Hong Kong is the largest British Council DTEO in the world with between 9,000 and 12,000 students registered for English language courses in any one term. Most of them are interested in upgrading their language skills in general terms, and are primarily instrumentally motivated. They come from a low to middle socio-economic background. The student body is described in detail in Chapter 4.

Testing and evaluation have not traditionally been areas of high priority in the British Council DTEO

network or most other English language teaching institutions. Up to the time that the appointment mentioned above was made, there were no testing and evaluation specialists working in any of the forty or so DTEO institutes scattered around the world. The brief of the officer appointed was:

i. to develop a suitably efficient, valid and reliable placement testing procedure to cope with the testing of up to 40,000 students a year;

ii. to develop a valid and reliable battery of progress tests for students registered with the institute to be administered to approximately 30,000 students a year;

iii. to engage in training staff in testing and evaluation principles and procedures.

Due to the unprecedented nature of the appointment, there were no customary practices to follow with regard to development procedures, training requirements, or the integration of testing and teaching to the mutual benefit of both. Similarly, there were no established test formats to adhere to. Within the institute itself, although a somewhat unsatisfactory placement procedure was in operation, there were no other testing

instruments or evaluation procedures. This meant that the testing specialist had a more or less free hand to carry out the brief.

The power of testing as an instrument of change was appreciated from the beginning of the project. The approach to course design and teaching materials was changing within the institute with a much greater focus being placed on the real-world communication needs of the students. The testing programme was seen as a necessary support to these changes. In some cases it even anticipated the changes. For this reason, it was decided to adopt a performance-based approach to the design of the test battery where possible. This meant that test items would be based on activities that students had to engage in in their day-to-day lives.

This thesis will concentrate on:

i. the issues involved in the integration of a principled progress testing programme into the life of the teaching institute;

ii. the development procedures and subsequent validation of a battery of seven progress tests, out of a total of approximately twenty-five such tests.

The relationship between testing and teaching is fraught with problems and difficulties, particularly in the field of Teaching English as a Foreign Language (TEFL). These are due in part to the fact that so little emphasis is placed on these areas in most training programmes, and also to the fact that many teachers involved in the teaching of English come from an Art's background. A comprehension of basic educational statistics is essential in the field of language testing. However, to achieve an understanding that is sufficient to carry out the fundamental requirements of test validation requires either specific training or a suitable academic background in a discipline such as psychology, or possibly one of the hard sciences. Since the majority of TEFL teachers do not have a background in either of these areas, and since the field of TEFL has been slow to realize the importance of sound testing and evaluation procedures, if only from the point of view of marketing and accountability, the introduction of a principled testing programme into a typical teaching institute has, up to now, been a low priority and open to

misunderstanding from teachers and administrators alike. This problem is discussed at greater length in Chapter II, and some of the ways adopted to overcome it are addressed in Chapter IV.

Because many institutions lack experience with large-scale testing programmes there is a danger that testing will not be well integrated into the life of a teaching institute and that tests, if they exist at all, will not match the teaching syllabus and general aims of the institute. For a testing programme to work, it must be integrated with the teaching programme, and it must reflect the needs of that programme and the students registered on it. In addition the testing instruments must be reliable and valid in the context that they are intended for at the very least, and they must reflect a view of language that is not in conflict with language teaching/learning methodology and theory. There has been a substantial, if inadequate, body of literature that has addressed these issues in Applied Linguistics and Education dating back to the fifties. Chapter 2 reviews issues related to language testing and the nature of what is to be tested. Chapter 3 considers important issues related to establishing the reliability and validity of tests and a number of

additional concerns that should be of importance to the test constructor.

While there is a paucity of adequate guidance for the construction and validation of test batteries in the context of a teaching institute, test constructors have also lacked adequate tools for the basic statistical analysis of the tests that they write. Training courses frequently focus on very time-consuming pencil and paper methods of conducting item analysis, but they are rarely applied to the real-world test construction situation. The British Council, aware of the problem, commissioned the writer of this thesis to develop a comprehensive item analysis and basic test statistics package for use on the microcomputer. This was done between 1984 and 1986, and this package (partially described in Appendix 8) was used for the initial analysis of the tests in the battery. The data from this analysis, as well as a detailed breakdown of the skills tested in the battery is presented in Chapter 5, Appendix 2, and Appendix 9.

Chapter 6 is concerned with the construct validation of the test battery. The seven tests are subjected to a series of factor analyses which investigate the nature

of the competence that is being tested. The Listening subtests are considered in greater detail than the rest because they are the most performance-based part of the battery.

2. Formulation of Hypotheses

Traditional testing methods are popularly credited with high reliability but limited validity. More recent approaches have, on the other hand, claimed high face and content validity at the expense of reliability. It is difficult to justify this position, and the apparent contradiction is discussed in greater detail in Chapter 3. The tests in this battery were designed to conform to high standards of reliability and validity while at the same time adopting a communicative format. Where possible a performance-based approach was employed in the construction of the tasks that appear in the test battery. The tasks were selected on the basis of their relevance to the students and the extent to which they reflected the course of instruction. Recent approaches to communicative syllabus design have emphasized the importance of enabling micro-skills in the successful completion of communicative events, a trend that is mirrored in the plethora of language teaching materials that have been published since the late seventies. It

was seen as a priority in the construction of the battery under discussion in this battery that current methodological positions should be reflected.

This led to the formulation of three principle hypotheses that the formed the main research interest in this thesis. These were:

i. A reliable and valid performance-based test battery could be constructed that would be of at least equivalent standard (as measured by classical test theory) to tests of a traditional format.

ii. It can be demonstrated statistically that students' ability in different language skills and areas of communicative competence are not equivalent.

iii. It can be demonstrated statistically that performance in communicative tasks is, at least partially, divisible into micro-skills.

The first hypothesis is investigated in Chapters 4 and 5. It was investigated because a popular position in the literature on language testing is that there will

inevitably be a conflict between highly content and face valid performance-based tests and the stringent requirements of a reliable measuring instrument. The writer was not able to embrace such a notion on either educational or moral grounds, and set out to prove that it was possible to produce highly face and content valid tests that were also very reliable.

The second hypothesis was investigated because there has been a body of literature claiming that a unitary competence factor underlies communicative competence. The methodological implications of such a position are that divisible approaches to both teaching and testing are in fact invalid. This seemed to the writer to be an intuitively unacceptable notion. The findings of this thesis confirmed recent research demonstrating that communicative competence was at least partially divisible.

The third hypothesis was investigated because recent approaches to both teaching and testing have implicitly assumed that communicative competence is based to some extent on enabling micro-skills. That is to say, in order to complete a complex task a language user must competently employ enabling micro-skills which have, up

till now, been isolated intuitively and left unvalidated. In order to investigate this hypothesis, the Listening section of each test was subjected to extensive factor analysis in order to see whether the individual items would group together in the same way as it was intuitively believed that they should.

Chapter 2 discusses the relationship between teaching and testing and then goes on to review approaches to language testing and teaching in order to establish a theoretical base position for the construction of the tests in the battery under discussion in this thesis. Chapter 3 considers the concepts of reliability and validity in an attempt to discern what they mean and their relevance to the work carried out. Chapter 4 provides a detailed background to the specification of test content, the nature of the student body concerned, and the integration of the design, construction and administration of the test battery with the aims of the courses, and the attitudes of the teachers and students. Chapter 5 is concerned with a detailed discussion of issues arising out of the item analysis of the test battery, and a correlational analysis of the tasks involved. It, along with Chapter 4, is concerned primarily with investigating Hypotheses One and Two. Chapter 6 is concerned primarily with the

investigation of Hypotheses Two and Three. Chapter Seven attempts to summarize the findings of the thesis overall, and to discuss their implications to language testing and teaching.

Chapter II

1. Introduction

This chapter provides a background and rationale for the design and construction of the test battery under discussion in this thesis. In the first sections attitudes towards tests and examinations and the reasons for the problematic relationship between testing and teaching are considered. The way that some of the problems mentioned was dealt with is explored further in Chapter 4.

A review of past and present trends in language testing, and the nature of communicative competence is then carried out with particular reference to the impact of these trends on the design of the test battery.

2. Some Considerations for Language Testers

2.1. Testing in the Social Context

Tests and examinations exert a powerful influence in most societies. Their effect on the individual is often traumatic and sometimes harmful. Indeed, in some countries, particularly in the Far East, it is not uncommon to read about significant numbers of suicides amongst young people which can be directly attributed to examination pressure.

Undeniably, tests carry with them unfortunate connotations, and it is impossible to conduct tests of any sort that wholly suppress these negative connotations in the wider social context. All those engaged in testing, at the classroom, institutional or public level, need to be aware that they are involved in a complex and problematic area. Individuals taking tests bring with them preconceptions expectations and personal experiences of varying sorts. Teachers involved in preparing students for tests, or even teaching a course that is tested bring with them equally variable sets of values.

The situation is at best difficult and unpredictable even if teachers and students come from the same culture and environment. If, on the other hand, the teachers and students come from different cultures, then there will, in all probability, exist, a significant mismatch of experiences, values and expectations. In most subject areas this does not happen very often. Language is the exception. English language teaching is the most common example of a discipline where teachers and students from different cultural backgrounds have to work together towards common goals, where there is often a lack of common experiences, expectations and perceptions. It is important that the test designer as well as the teacher becomes familiar with the expectations and values of the students he is working with.

2.2. Testing and Teaching

Running in parallel with the problems outlined above is the conflict between teaching and testing. Views on the nature and role of examinations differ widely. To some, they offer the only method of selecting the most

able individuals in a given population both fairly and impartially, while to others, they represent the worst form of discrimination in education. Most of the harshest attacks on testing, and its role in the educational process come from developed countries where the role of education as the major factor in upward social and economic mobility has diminished. There are many in Britain and the United States of America, for example, who write against the use, or misuse of tests. They show a very real concern for the damage that tests can cause, on the one hand, and what they consider to be the fundamental implausibility of using tests to measure the content of the test-takers' mind. John Holt (1970) is a firm critic of testing and he makes the following statement:

> "I do not think that testing is necessary, or useful, or even excusable. At best, testing does more harm than good; at worst, it hinders, distorts, and corrupts the learning process... Our chief concern should not be to improve testing, but to find ways to eliminate it ... How can we expect to measure the contents of someone else's mind when it is so difficult, so nearly impossible, to know more than a very small part of the contents of our own?"

Most people involved in education would admit that there is often perceived to be a direct conflict between the goals of the teacher and the pressures exerted on the educational process by tests and examinations. In any curriculum that is geared towards

the taking and passing of examinations, preparation for the examinations will tend to have a significant effect on what goes on in the classroom. In addition, research has indicated that most pupils feel that school should prepare them for examinations (Rutter et al., 1979; Gray et al., 1980). It is reasonable to assume therefore that learners can and will exert pressure on their teachers to prepare them for examinations and tests. To most of them the preparation will inevitably involve much practice of the item types that appear in the test. The appearance of the test is thus of great importance both from the point of view of the teacher and of the learner. This has led some (Morrow, 1979) to imply that statistical criteria should be more or less abandoned in favour of face validity. While a whole hearted acceptance of such a suggestion would be unwise it is certainly important for the test designer to take into account that pressure to prepare students for tests by practicing typical examination questions will inevitably come to bear on teachers from the learners, the school or sponsors.

The design and format of test items should reflect an appreciation of the washback problem outlined above since one of the primarily purposes of education is to prepare students to pass examinations. It is

unfortunate that this purpose has a tendency to dominate all others. Due to the role that tests play in the process of selection in most societies and the consequent need for accountability, the requirement for a reliable instrument has sometimes outweighed the effect that a test might have on the curriculum. Broadfoot (1979) puts the point well when she writes:

> "...educational assessment, perhaps more than any other aspect of education, has suffered the thraldom of 'methodological empiricism' in which questions of technique have predominated over the more fundamental issue of its effect."

Teachers are quick to point out that there is frequently a conflict between educationally desirable outcomes and certain types of test. This view is partly responsible for a great deal of the resistance that test designers often face when trying to introduce new tests into a teaching context. Holt (1981) expresses the view held by many teachers in the secondary sector:

> "What examinations, and other tests of attainment or performance, aim to do is allow conclusions to be drawn about educational activities. But there is immediately a conflict between the aim of the educator, which is to help the pupil to achieve understanding, and that of the tester, which is to distill understanding into some observable state."

The relationship between teaching and testing is characterized by this perceived conflict between the two. There seems to be no fundamental reason why there should be a conflict between the aims of the teacher, and those of the tester. It is surely more appropriate to take the position that the teacher and tester have different roles to play and different contributions to make. They can, given the appropriate context and circumstances, develop complementary roles that are to the benefit of the learner. The fact that the relationship has not always been satisfactory does not mean that it is wholly untenable. The test battery described in this thesis represents an attempt to overcome the difficulties outlined above.

3. Testing and the Teaching of English as a Foreign Language

The pressures on teachers and their own reservations about testing exist as much in the field of Teaching English as a Foreign Language as they do in the more highly developed areas of secondary and tertiary education. Resistance to testing is perhaps even stronger in the field of Teaching English as a Foreign Language (TEFL) in that language testing and evaluation

are probably the least developed field in that discipline.

It is difficult to know exactly how the tester and testing are perceived by most English language teachers since no role has been clearly defined. In General Education, Testing and Evaluation have a long tradition. In TEFL this is not the case, and much work of relevance in Education is ignored or reinvented. This may be because most of those in TEFL have not come from a training in Education. Rather they have drifted in from other often unrelated disciplines. Much of their training, at R.S.A. or Masters level, focuses either on classroom techniques, or aspects of linguistics. Many TEFL oriented Masters courses, for example, are in Applied Linguistics and organized by departments of Linguistics while most of the participants are teachers requiring a qualification that helps to equip them for teaching jobs. It seems incongruous that a discipline which is essentially concerned with educational issues should turn to a related theoretical discipline of questionable significance for one of its highest qualifications. This has resulted in an unsatisfactory situation that has meant until recently only minimal interface between work in General Education and TEFL.

In the TEFL world Testing and Evaluation takes place in three contexts that are often in conflict, with little interchange or cooperation taking place between them. The first and most common environment in which testing takes place is on a relatively informal level, in the classroom by the teacher. This may take two forms. It may involve the individual teacher actually preparing test materials for his own students in order to gauge progress or diagnose areas of weakness. Or it may involve the teacher in administering and contributing to the design of institution wide tests, at the end of the term or year, in order to evaluate and grade students' progress or achievement.

Most TEFL training courses devote minimal time to testing and evaluation, and as a result, many teachers are painfully ignorant of even the most basic ideas and concepts. This means that they often do not have the skills or the confidence to develop effective classroom tests or evaluation measures. It is not unusual in such a situation to avoid anything to do with testing. Reactions to testing from many TEFL teachers are often negative not through any firmly held moral or educational conviction, but due to a lack of adequate

training possibilities. Acheson (1977) carried out a study of teachers' awareness of testing and evaluation in the United States and Britain and discovered that few language teachers had little more than a cursory knowledge of these areas. Similarly, in a study carried out by Stevenson and Riewe (1982) it was found that most practicing language teachers do not have a coherent background in either language testing or educational and psychological measurement. The result is often that even when teachers decide to write tests, they approach the task in a very naive and unprofessional manner so that basic measurement criteria and design procedures are ignored.

The RSA Diploma course which is taken by several hundred practicing teachers every year requires approximately four hours to be spent on testing. This amounts to about four per cent of the minimum total taught course, and in many cases probably less. In most full-time one-year British-based Masters courses a frequently optional module on testing is unlikely to be allocated more that ten to fifteen hours and there may or may not be modules on evaluation and statistics. It is quite possible for a TEFL teacher to get to beyond Masters level, in terms of training, with virtually no formal exposure to the principles and techniques of

testing, not to mention evaluation in education or statistics. Given teachers' natural apprehension of these areas it is not surprising that testing along with the other areas mentioned above, remains a constant source of difficulty and misunderstanding.

The situation is no better in the field of materials production. Most published textbooks devote minimal attention to testing and evaluation. Rea (1985) makes the statement:

> "Although materials writers have assigned importance, in varying proportions, to the process of assessment, one is left with an impression this is the result of an afterthought. There is a strong element of "vague puff" about many statements on "testing and teaching", which are of doubtful value to the (overworked) practicing teacher, who has then the task of interpreting this "puff" in the form of a coherent testing programme which involves, minimally, the selection of tests and item types appropriate to the purpose(s) for testing."

Even books dealing with the history of English language teaching (Howatt, 1984) and its fundamental concepts (Stern, 1983) barely touch on the issues involved in testing and evaluation.

A greater focus on testing, evaluation and statistics in training courses, and more comprehensive attention

to assessment by materials writers would undoubtedly ease the difficulties encountered by the typical language teacher, and improve the quality of not only testing instruments and techniques, but also teaching materials.

The second major context that testing appears in is that of the professional examining body. Such organizations have a role to fulfill - to prepare formal examinations that will reliably rank candidates (although the trend towards criterion-referencing is gathering momentum in British examination boards) and provide secondary and tertiary educational bodies as well as employers with adequate information on which to base selection decisions. In order to do this successfully they often use testing devices that teachers perceive as being in direct conflict with sound teaching practice. This is partly because examination boards, while attempting to reflect current methodological and pedagogical trends are inevitably slow in implementing change. It often takes years to alter the format and content of an examination. As a result, examinations might sometimes appear to adopt an out-of-date approach.

Most examining bodies are aware of the difficulties outlined above and try to compensate for them through the involvement of teachers in the setting and moderating of examinations; focusing on meaningful testing activities; and constant attention to revisions in approach. In addition, they generally have a complex structure of subject committees with extensive teacher representation in order to ensure that educational considerations are not ignored and they engage in extensive public relations activities. And yet examining bodies are rarely perceived in a positive light by teachers and find themselves open to constant criticism in some ways providing a necessary scapegoat for teachers. Given the nature of what boards have to do in terms of accountable measurement and the time constraints that they face, it is difficult to see how they can easily improve their image with teachers in general. Yet one cannot help but feel that despite their show of professionalism, many examination boards are very lax in terms validation procedures. When questioned, they are generally on the defensive and this attitude, in addition to the problems outlined above, is responsible too for the divide that exists between the professional examining body and the average teacher.

The third role that testing plays is in Second Language Acquisition research. Tests, in a broad sense, are clearly crucial to many quantitative research projects. In Britain it is unfortunately the case that very few of the postgraduate training courses devote enough time to testing or statistics and thus much research involving statistical hypothesis testing is inaccessible to most of those involved in TEFL. Moreover, in recent years statistical techniques employed by researchers have been rather complex.

An important factor in the unpopularity of research directly related to language testing is that it makes claims that are of relevance to the classroom and yet it often does not attempt to specify exactly how they are relevant. In addition the tests that have been used in the research have been rather unimaginative for the most part e.g. TOEFL, FSI Oral Interview, cloze and dictation. There is nothing more frustrating than to be told that the results of various experiments which you do not really understand could significantly affect what you do in the classroom and then not be told how.

In addition, most testing research has not focused on educational issues at all but rather on linguistic or psycholinguistic ones. This imbalance has further increased its unpopularity. For example, theoretical positions such as the debate over the unitary competence hypothesis and the nature of cloze as a direct measure of underlying language proficiency that raged a few years ago, were not tempered by an easily recognizable regard for the concomitant educational consequences. Clearly the claims about a unitary competence factor could have important implications for methodology. However, careless use of testing techniques such as cloze which, it has been claimed, measures underlying proficiency, is bound to have a powerful and possibly negative effect on classroom practice. The research would have been more meaningful if the various allusions to the significance of the results to the teaching context had been explored more thoroughly than they were.

For a test design and implementation project to succeed attention needs to be paid to the fact that testing is often disapproved of by teachers and that many of them have minimal training in test design and analysis. The approach adopted to cope with these difficulties in the design and implementation of the test battery under

discussion in this thesis is explored at greater length in Chapter 4.

4. Testing past and present

Faced with the practicalities of test design and implementation in the context of a language teaching institute test designers inevitably encounter a number of problems that need careful consideration. The first problem is the extent to which research in language testing and the nature of what is to be tested should affect test design and appearance. Arising out of this is the problem of the effect that tests have on what is taught in the classroom and the expectations of students and teachers about what should be tested. Finally, there is the interpretation of test results and the use to which they will be put. Test results are used by teachers, mainly for diagnostic and formative purposes (Cohen, 1980) in that they provide data for amendments to the teaching syllabus. They are used by institutions and society normally for selection or deselection. And they are used by individual students both as a measure of competence, generally in relation to their peers, as well as to provide information on their progress, strengths and weaknesses in order that

they may, for example, make decisions about their own learning strategies - either amending them, or leaving them as they are.

Language Testing has inevitably reflected teaching practice to some extent. When attitudes to teaching were highly structural, so was testing. As attitudes to teaching encouraged a more communicative approach testing followed suit. In the next section the interaction that has taken place over the last thirty years or so is reviewed.

4.1. The Pre-scientific Period

Various trends and approaches have dominated language testing in this century. The prescientific (Spolsky, 1975), sometimes referred to as traditional (Bird and Dennison, 1987) trend, is said to have characterized language teaching and testing prior to the 1920's (Valette, 1977). In fact, it would be true to say that in foreign language teaching in the United Kingdom at least, this trend persisted through to the 1960's and later. Indeed, Bird and Dennison (1987, p. 13) contend that it probably exists almost to this day:

> "Interestingly, PGCE students, when asked in November 1986 how they had learnt their foreign languages, reported that this process had involved many of the traditional features, though the majority of these students had been in the fifth form only six or seven years previously!"

The same may also be true in the United States. Valette in her handbook (1977) writes:

> "The first edition of this handbook, 'Modern Language Testing' (1967) was written primarily to help teachers without formal training in measurement to move from this "prescientific" method of evaluation to the more objective evaluation techniques of the "psychometric-structuralist" trend."

The concern we now have for objectivity, reliability, statistical validation and analysis was not an issue with prescientific language testers. Language was taught and learned in order to improve the analytical, intellectual capacity of the student and to equip him with the necessary tools to understand and appreciate the literature of the language he was studying. It was, according to Bird and Dennison (1987, p. 13) ...

> "... highly teacher-centred, based on written work and used grammatical explanations, together with translation, as the learning medium."

The focus was to a large extent on the written language as opposed to oral communication. Perhaps the reasons

for this are linked to the fact that Latin was studied by most grammar and public school pupils. The study of Latin focused by necessity on literature and the written mode. Oral aspects of language were only focused on as a way of learning grammatical patterns and as a quick and easy way for the teacher to assess whether the students had mastered the rules and vocabulary required. In foreign language teaching the same was generally the case. It is true that there has long been an oral component in public examinations such as the GCE 'O' and 'A' levels, however, its influence on foreign language teaching was minimal for many years.

Tests, during the prescientific era, were mainly essays, dictations and translations. It would be unfair to claim that considerations such as standardization of marking, for example, were ignored. Common sense tells us that markers would have felt the need to agree on standards. However, the emphasis on such considerations was not as important as it was to become in later years. It seems clear however, that the tests used reflected classroom practice and the aims of the teaching syllabus fairly well. They were direct tests of proficiency or achievement, and as such perhaps more valid educationally than what was to follow.

4.2. From Psychometrics to Psycholinguistics

Until the last ten years or so, the predominant approach to the design of language tests was laid down by Lado and others during the 1960's. Just as the formal structural analysis of language provided the main focus for language teaching materials during this period, (and still does in many parts of the world), the structural syllabus generated by the structural approach in its various forms provided the main source for language test content. The main design principles for language tests of this kind, based on behaviourist psychology and structuralist theories in linguistics, are well known and illustrated in works by Lado (1961), Valette (1967), Harris (1969), and Heaton (1975). Much testing was indirect in that productive language skills such as writing or pronunciation were tested receptively using the multiple-choice format. Psychometric-structuralist test items are characterized, for the most part, by an emphasis on objectivity of marking which is achieved by using carefully written discrete-point multiple-choice items and an emphasis on statistical techniques that conformed to high standards of reliability and

concurrent validity. The tests that were produced implicitly adopted a hierarchical view of language proficiency in line with the structural linguistic view of the nature of language i.e. from phoneme to morpheme to word to sentence. A cursory glance at Lado's "Language Testing" (1961) illustrates this point clearly. However, there were few attempts to define language proficiency explicitly, even though Lado went some way towards a definition when he described the process of language acquisition as the internalization of a series of habits of communication. He wrote:

> "These habits involve matters of form, meaning and the distribution of layers of structure, namely those of the sentence, clause, phrase, word, morpheme, and phoneme." (1961, p.22)

Although discrete-point tests dominated at this time, there were critics even then. Carroll (1961) noted that a major limitation of discrete-point tests was that they tested only one element of language at a time. He argued that this did not reflect real language use in most cases. He suggested the use of types of tests that focused on the communicative effect of an utterance rather than discrete-point components. Carroll called such tests 'integrative'. He (Carroll, 1968, p.58) provided a clear statement on the integrative position:

> "Since the use of language in ordinary situations calls upon all these aspects (of

> language), we must further recognize that linguistic performance also involves the individual's capability of mobilizing his linguistic competence and performance abilities in an integrated way, i.e. in understanding, speaking, reading or writing in connected discourse."

The term 'integrative' was defined by Oller (1973) to include all tests which were not discrete-point. Since integrative tests are supposed to test the learner's ability to apply many language skills all at the same time, Oller proposed that tests like cloze and dictation would be best suited to the task. He spent many years developing a theory and attempting to prove that it was correct. Clearly, a difference of opinion existed between those who favoured a discrete-point approach, and those who favoured an integrative one. A third position favoured an eclectic approach (Rivers, 1968; Clarke, 1972; Heaton, 1975; Davies, 1978) and held that language tests should be a combination of the first two approaches. The overall impression remains, however, that much of the discussion had retreated into a domain of specialist testers which did not have strong connections with the requirements and demands of instructional programmes.

Although many essential techniques in language testing were initiated during this period, two major

weaknesses are apparent. Firstly the intricate interaction of teaching and testing was not dwelt on. Emphasis was placed on the relationship between linguistic and psychological models of language and tests as opposed to the equally important issue of learning, teaching and testing. Secondly, the role of language as the facilitator of communication was not really considered in depth. A typical statement on the matter is found in a paper by Ingram (1968), which, while accepting the integrative nature of language performance did not explore the implications:

> "The purpose of language is communication. The so-called language skills describe different modes of communication. Speaking, listening, reading and writing are not intrinsically separate, they describe communication in terms of sending and receiving, in spoken and written language."

Questions such as who is communicating with whom, for what purpose, in what setting, and the extent to which performance is based on underlying skills (enabling skills) was not a major concern in language testing, although it would be unfair to say that they played no part at all. The linguistic aspects of language proficiency were easier to isolate and thus more testable. Moreover, linguistics itself was at this time focusing mostly on the Chomskian approach to the analysis of language that placed its primary focus on

the ideal speaker-listener. Consider the statement below (Chomsky, 1965, p. 3):

> "Linguistic theory is concerned primarily with an ideal speaker-listener, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance."

The focus in linguistics started to shift noticeably in the late sixties and early seventies with some of the most influential work being done by Hymes. His article "On communicative competence" (1970) has been credited with playing a major role in the broadening of the scope of linguistics and language teaching. However, at the time that Ingram wrote the article cited above, this influence was not apparent. An example of the language test specification she produced is illustrated below. It was typical of the period.

Language Skills

Components	Listening	Speaking	Reading	Writing
Phonology/ orthography		✓		✓
Structure				✓
Vocabulary			✓	
Rate and general fluency	✓	✓		

Check marks indicate components measured by the final examination. Shaded areas indicate components not emphasized in the course.

Although Ingram's paper does not define the term 'communication' the issue had been considered in the testing literature to some extent. However, detailed discussion was generally dismissed. Lado justified the dismissal when he wrote:

> "The situations in which language is the medium of communication are potentially almost infinite. No one, not even the most learned, can speak and understand his native language in all the situations in which it can be used ... even if we could pick only valid situations and even if we could be sure that understanding these situations occurred through the language used, we would still have the problem of the great variety of situations which must be sampled. The elements of the language on the other hand are limited, and it is more profitable to sample these elements than the great variety of situations in which the language is used."

Using such a justification, test writers could produce fairly abstract test items which while satisfying measurement and linguistic criteria, did not make any serious attempt to provide a valid external context. That is to say, the rationale for the construction of what may have been a somewhat bizarre test item was that it attempted to test a point that appeared on a list of structures rather than that it reflected language use in real life situations. However, since much language teaching methodology was equally structural and inward looking, the tests were considered valid. We might now question their validity on the grounds that the language segments sampled for test items were neither adequate nor authentic and that the relationship between use and usage was left unexplored (Alderson, 1981). On the other hand, the important contributions made to test design during the psychometric-structuralist era must not be forgotten. Contributions like, the emphasis on statistical analysis, reliability and validity, the careful planning of test content and the development of the discrete-point multiple-choice item. All of these were of importance to the work described in this thesis.

4.3. The Psycholinguistic-sociolinguistic Period

4.3.1. Communicative Language Testing

Communicative language testing arose out of a shift in language teaching/learning theory and methodology away from a predominantly structural focus to one that emphasized the importance of language in use. This shift of focus began in linguistics and was continued and modified by developments in related fields such as sociolinguistics. It was Hymes (1967) for example who developed the notion of the speech event, a term used to refer to language activities that are governed by rules of use. He pointed out that different speech events demand different sets of rules of use and that their structure can be defined by breaking them down into constituent factors such as participant, setting, purpose, topic, channel etc. In language teaching in Britain such ideas were developed and discussed by Widdowson (1978), Munby (1978), Littlewood (1981) and Brumfit (1984) amongst others.

Munby's (1978) specifications and guidelines for communicative syllabus design are very detailed and

were one of the factors that sparked off the development of a more communicative trend in language testing. His approach is based on the premise that the language to be taught should be related as closely as possible to the learner's immediate and future needs, that the learner should be prepared for authentic communication, and that the language taught should have a '**high surrender value**' (Wilkins, 1976).

While psychometric-structuralist language tests paid relatively little attention to defining the dimensions of language proficiency and communicative competence, the same cannot be said of the work conducted in language teaching and testing since the mid-seventies - a period that Spolsky (1975) called the psycholinguistic-sociolinguistic era. Cziko (1982) has made a useful distinction with regard to the research in language testing during this era. He divides the research into two main categories which he calls descriptive and working models of communicative competence. Descriptive models are ones which attempt to describe:

> "... all the components of knowledge and skills that a person needs to communicate effectively and appropriately in a given language."

Working models are defined as attempts to:

> "... show how components of communicative competence are interrelated psychologically to form a set of independent factors."

This is a useful distinction in any discussion on language testing. Descriptive models are illustrated by the work of Canale, Swain and Cummins while Oller, Palmer, Bachman and others focus on working models.

4.3.2. The Canale and Swain model

Perhaps the best known descriptive model of communicative competence is the one put forward by Canale and Swain (1981; 1983) According to them, communicative competence encompasses four components - grammatical, sociolinguistic, discourse and strategic. **Grammatical competence** is concerned with the mastery of vocabulary, and the rules of word formation, sentential grammar, linguistic semantics, pronunciation and spelling. In short, the sorts of things that provide the content of tests described earlier in this chapter. **Sociolinguistic competence** contributes to the individual's ability to communicate appropriately. It is the extent to which utterances are produced and understood appropriately in different sociolinguistic

settings depending on factors such as the purpose of the interaction, the status of the participants and so on. It involves an awareness of the do's and don'ts of social interaction that are culture specific. **Discourse competence** refers to the mastery of the ways in which grammatical forms and meanings combine to achieve unified spoken or written texts. Finally, **Strategic competence** refers to the mastery of verbal and nonverbal communication strategies to compensate for breakdowns in communication. Such strategies might include things like repetition, paraphrase and slower speech. Strategic competence is different from the other competencies postulated by Canale and Swain in that it interacts freely with the others.

5. Validating the Canale and Swain model

Not a great deal of research has been attempted in order to validate this model. Some fairly informal work on the model as a whole was done by Schmidt while Farhady investigated the construct of sociolinguistic competence.

5.1. The Wes study

An interesting if fairly informal piece of work was carried out in Hawaii by Richard Schmidt (1983). He traced the development of Wes's (a Japanese artist) communicative competence over a four year period basing his definition of communicative competence on the Canale and Swain model. Schmidt found that Wes's communicative competence developed in an unexpected way. In the whole of the four year period there was hardly any improvement at all in the area of grammatical competence. That is to say that Wes was making very much the same mistakes at the end of four years as he was at the beginning. On the other hand, Schmidt was able to detect significant improvement in sociolinguistic, discourse and strategic competence.

The Wes study has interesting implications for Second Language Acquisition research, but it has equally interesting implications for language testing. It indicates that if our aim is to assess communicative competence as defined by Canale and Swain, it may well not be adequate to expect to be able to infer overall competence from the results of a grammar or vocabulary test, for example, because such a test will supply us with very limited information. Wes, in Schmidt's study,

would have scored very low in grammar tests and a school would have been obliged to say that he had made no measurable progress. However, there had been significant progress in other areas and Wes was accepted as a competent user of English by the native speakers that he mixed with. Schmidt's research supports the view that there are different components in communicative competence and that ability in one of the components does not necessarily reflect ability in any or all of the others. It also indicates that the development of the components might well be staggered as opposed to uniform. If we are to produce meaningful tests, meaningful to teachers, students and the institute, then it seems reasonable that the tests need to focus on features of communicative competence over and above the level of grammatical competence as defined by Canale and Swain. Schmidt's research seems to indicate that there is not necessarily a direct relationship between the different components. Consequently, a common claim made of certain language tests - that overall competence can be inferred from one type of competence, is highly questionable. Test designers need to experiment with items that aim to measure a broad range of the facets of communicative competence. Equally they need to be aware that an individual's ability to use language in order to communicate will be multi-faceted and that the

profiling of different aspects of communicative competence may be a more appropriate approach than simply producing overall scores that are a combination of these facets. On the other hand, it is not clear what the relative importance of the different aspects of a profile may be.

5.2. The investigation of sociolinguistic competence

Farhady (1981) also used the Canale and Swain model of communicative competence in the design of a functional ESP proficiency test to measure sociolinguistic competence. His test is based on two functions from the Van Ek (1975) taxonomy. Situations were carefully chosen for their relevance to the test group and pretested at some length with native speakers as well as a sample of the test population. As a result, Farhady was able, to his satisfaction, to specify degrees of appropriateness recognizable by native speakers as the criterion of sociolinguistic competence.

A similar approach was adopted at the level of the large-scale public examination in Hong Kong with the

Junior Secondary Education Assessment (JSEA) in English (described in Milanovic, 1987). In this test there is a section which attempts to focus on functional/ sociolinguistic competence. Situations are selected from an analysis of the types of interaction that the pupils who take the test are most likely to encounter. Grammatical accuracy is not supposed to be a criterion in the marking of this section of the JSEA. Items are discrete-point in the sense that they stand alone, as the example below shows:

A friend asks you:

"What are you doing this evening?"

You reply:

a. "I'm a student."

b. "I'm feeling well."

c. "I'm fine thanks."

d. "I'm going out."

The emphasis is clearly different from the conventional type of grammar based discrete-point item. However, producing items of this sort is not very easy because it involves a needs analysis approach to the selection of appropriate situations, role relationships and content. It is much simpler to look up a list of structures at the back of a text book. In addition, it is important to keep items pure in the sense that they

should not test grammatical accuracy rather than appropriateness. It is often difficult to separate the two with this type of indirect test. However, items of this type should be investigated to determine whether they can be shown to test something like sociolinguistic competence. An attempt is made to do this later in this thesis.

5.3. Concluding comments

The descriptive model of communicative competence proposed by Canale and Swain has provided some useful direction to the test designer. It has encouraged research into the areas that they define, leading to the consideration of a different focus for test items. The traditional focus on the four skills has been shifted. However, there do not appear to be any clear guidelines for test constructors, the implication being that constructors must also be researchers. Many questions are left unanswered. Are the areas defined to be tested separately or are the items to be integrative? If they are to be tested at all, what is the relationship between them and what is the relative importance of the different components? In other words, how should they be weighted? The model is intuitively

appealing nonetheless, and should be experimented with in test design and the measurement of communicative competence. It was influential in the design and construction of the test battery discussed in this thesis.

6. Cummins Model of Communicative Competence

An important descriptive model of communicative competence which may have an influence on the design of tests and interpretation of results was developed by Cummins (1979; 1983). This model has undergone some changes since he first presented it. The two major versions are discussed below.

6.1. Version 1

Cummins' first model of communicative competence drew a distinction between cognitive/academic language proficiency (CALP) and basic interpersonal communication skills (BICS). While everybody is supposed to possess BICS the same is not true of CALP, which is strongly related to literacy skills, and which

Cummins equates to Oller's global language proficiency factor. BICS is thus a species minimum competence, while CALP is acquired through education which is why, according to Cummins, it takes language minority students much longer to attain grade/age appropriate levels in English academic skills than it does in face-to-face communication.

6.2. Version 2

The BICS/CALP distinction suggests that there are two types of language proficiency. However, Cummins has since revised his position (Cummins, 1983). He has suffered harsh criticisms from Edelsky et al. (1983) for the potentially detrimental consequences of his theory for minority children and defended himself vigorously (Cummins and Swain, 1983). This is an important dispute but one which I shall not go into here since it is not directly relevant to the discussion.

Cummins now postulates that language proficiency can be conceptualized along two continua:

> "First is a continuum relating to the range of contextual support available for expressing or receiving messages. The extremes of this continuum are described in terms of "context-embedded" versus "context-reduced" communication. They are distinguished by the fact that in context-embedded communication the participants can actively negotiate meaning (e.g. by providing feedback that the message has not been understood) ... context reduced communication, on the other hand, relies primarily (or at the extreme of the continuum exclusively) on linguistic cues to meaning and may in some cases involve suspending knowledge of the "real" world in order to interpret (or manipulate) the logic of the communication appropriately."

Cummins claims that interpersonal communication is normally context-embedded while context-reduced communication occurs in situations where linguistic precision is of great importance. Of course, the extent to which something is context-embedded or context-reduced is dependant to a large extent on the individuals concerned in the communicative event (Munby, 1978) and there can be no hard and fast rules. However, the implication is that the less context there is the greater the effort to communicate will need to be. This position has implications for the types of task that one might include in a test in relation to the amount of context that the individual brings to the test and endorses the view that tests can never be fair to everybody. In practical terms this is something that test designers have to live with but it is an important point to bear in mind in the

selection of test materials and tasks. It supports the position that if test items do not provide a familiar context for students, then the items will be more difficult. In the battery of tests described later, an important design principle, one of familiarity, is based in part on Cummins argument.

The second continuum in the Cummins model relates to the amount of cognitive involvement in a task or activity. Cummins defines cognitive involvement as ...

> "the amount of information that must be processed simultaneously or in close succession by the individual in order to carry out the activity."

Language tasks are categorized in relation to the amount of context supplied and the degree to which they are cognitively demanding. However, it is difficult to define the notion of context very precisely because it will be different from one individual to the next. Furthermore, what might be cognitively demanding at the beginning of a learning cycle may not be at a later stage.

It may be that one of the most important implications of Cummins' model to language testing is that it

requires the test designer to focus more on the learner as an individual. In order to use the contextual and cognitive dimensions it is inevitable that the test designer will consider the background characteristics of the people being tested. Unfortunately, this is a feature of content validity that is generally ignored, often to the detriment of the test takers.

7. Working Models of Communicative Competence

I do not propose to spend much time on working models of communicative competence here since they are discussed in greater depth in Chapter 6. Oller (1974) and others have attempted to measure a hypothesized underlying linguistic competence based on the notion of a grammar of expectancy. They have used the cloze technique as one of their major testing instruments and this has led to a number of research papers based on the cloze test that have made far reaching claims. The effect on test designers has not always been positive in that there has been a tendency to use the cloze as a panacea (Alderson, 1982), not only in the testing of proficiency but also in achievement testing. In consequence, important considerations of validity have frequently been ignored. Claims were made that the

cloze was automatically reliable hence tedious statistical procedures could also be ignored. Text effect was argued to be of no consequence so it could be discounted too. The current position held by Oller and others is now far weaker than it was (Oller, 1985) due to criticisms of their findings and research methodology by Alderson (1978), Vollmer and Sang (1983), Farhady (1983), Lee (1984) and Sang et al. (1986). However, indiscriminate use of cloze with unrealistic claims is still a problem with some test designers.

Palmer and Bachman (1980, 1981) have focused on the a posteriori validation of the construct of communicative competence principally through the use of factor analytic techniques. They used TOEFL subtests in many of their validation studies and although their findings are of importance, they are of debatable direct value to the test designer. In essence their findings tend to support the hypothesis that there are a number of factors at play in language proficiency, in contradiction to Oller's early findings, and that there may be an item or task effect, that has generally been ignored in validation studies, which is potentially at least as powerful as the skills effect. Bachman and Palmer's research (along with the research of others in

the same area) suggesting the partial divisibility at least of the construct communicative competence, and their work on task effect were influential in the design of the test battery under discussion in this thesis.

8. Communicative Language Testing in Britain

8.1. The Morrow approach

Research in language testing in North America and Canada has either attempted to describe the construct of communicative competence, or statistically isolate factors underlying it. This is not the case in the field of language testing in Britain, for the most part. Morrow (1979) does not really attempt a definition of communicative competence. Instead, he lists some of the features he hypotheses to be part of authentic communication that should be taken into account in communicative language test design in order to make tests valid. These features are:

i. Communication is **interaction based**, in that what is said or written by an individual depends crucially on what is said or written to him/her.

ii. Communication is **unpredictable** and data has to be processed in real time.

iii. Communication requires a **context** that will be situational as well as linguistic.

iv. Communication is **purposeful** in that an individual must be able to recognize why utterances are addressed to him/her and produce relevant responses that will achieve the desired purpose.

v. Communication requires **performance**, that is, the ability to use language in real situations.

vi. Communication involves the use of **authentic texts**.

vii. Communication is **behaviour-based** in that it has an outcome.

These features are explained in greater detail in Morrow's article 'Communicative language testing:

Revolution or evolution.'(1979) Although this is a useful list and of value in communicative test design, it has been criticized (Alderson, 1981; Weir, 1981; Moller, 1981). Morrow does not define terms like communicative proficiency, language competence, performance test and behavioural outcome nor does he explain adequately how the seven features outlined above are to be taken into account in the design of communicative language tests and how they should be measured or weighted.

The approach has had an influence on test design, primarily in Britain. It has encouraged the production of tests that look more appealing and realistic, and that bring teaching and testing materials closer together. Morrow's approach influenced the design of the progress test battery discussed in this thesis in the ways outlined above.

The most widely used public examination to be influenced by the Morrow approach is the Royal Society of Arts **Examination in the Communicative Use of English**. It represents a major change in emphasis in language testing terms and as such, merits discussion. In addition, it was of influence in the design and

production of the test battery presented in this thesis.

8.2. An Examination in the Communicative Use of English

The best known public examination in Britain to be developed along communicative lines is the Royal Society of Arts "Examination in the Communicative Use of English". The specifications are different from those of earlier large-scale public examinations in several ways and based to a large extent on the Morrow approach. The designers have accepted the fact that people taking the examination may have different levels of competence in different skills and that this in itself should not prejudice their chances of passing the examination. As a result, candidates are allowed to take components at different levels. They may choose the intermediate writing component, the basic listening and speaking components and the advanced reading component. This allows for the creation of a student generated profile of competence and a type of flexibility not available in other public examinations.

The tasks are not assessed on accuracy criteria alone. The range of considerations used include, size of text, complexity, range, speed, flexibility and the amount of repetition required. The input is as authentic as possible while the tasks to be carried out on the input are graded. A list of possible topics is also specified for the listening, reading and writing sections of the examination.

An interesting approach has been adopted in the oral section. Rather than conduct an interview with one assessor and one candidate as is generally the case in this type of examination, two candidates are examined together. The nature and quality of their interaction as well as their production of language and comprehension is the subject of the assessment. The examiner does not play a participatory role in the interaction. Instead, an interlocutor is used. His role is to initiate and, where appropriate, guide the interaction. The approach attempts to replicate real communicative events much more so than the traditional approach to oral examining. As such it makes a valuable contribution to the format of oral tests.

It is important to note however, that the criteria established for this examination have been based on intuition for the most part and there has been little attempt to validate them. Reliability and various types of validity have been rejected in favour of face validity which appears to be the main justification for format much of the time. This weakens the impact of its contribution to language testing quite considerably.

8.3. Proficiency Tests of English for Academic Purposes

Test design in Britain for the last few years has focused extensively on the activities that the people taking the test need to perform as the source for test items. Variations on Munby's (1978) needs analysis model have been employed by Carroll (1978) and Weir (1983) in the design of large-scale proficiency tests for English for Academic Purposes.

The ways in which the test takers are likely to use language have always been a central concern and it would be unfair to suggest otherwise. However, the rigour with which these purposes are identified, the techniques developed for sampling and the detail with

which the tasks are specified is much greater now than previously. We have always known, for example that prospective university students need to write essays, we have not, on the other hand, always attempted to describe the range of skills necessary to facilitate the production of adequate written work.

A focus on needs analysis has also led to the identification of what Munby calls communicative events. It is claimed that these are composed of a series of enabling skills the existence of which is widely accepted by test designers though not really verified through empirical research. Although this might sometimes lead to a fairly conventional test format, it is important to note that the rationale for the selection of testing points is significantly different from that adopted by tests in the past. Since many of the test tasks in the battery under discussion in this thesis were based on the assumption that they were made up a range of enabling skills, and attempt was made to discover whether they could be isolated statistically, thus providing evidence of the construct validity of this intuitively appealing approach to language testing.

9. Performance-based Language Testing

The approaches outlined above reflect a trend towards performance-based testing that has been developing in recent years. According to Wesche (1987):

> "In performance-based testing, examinees must demonstrate their second language proficiency through tasks whose content and contextual features represent the situations in which the second language will eventually be used."

In the minds of many teachers, performance-based testing is confused with criterion-referencing (discussed in greater detail in Chapter 3). They may indeed have much in common, but they are certainly not synonymous terms. Wesche argues that a performance-based approach would not generally be used to establish overall proficiency levels, nor to establish specific aspects of language knowledge, but rather to determine the extent to which an individual can carry out certain specific activities that are directly related to the uses to which he will put the target language. She also maintains that such tests will be used primarily with relatively advanced examinees mainly because:

> "... second language acquisition theory as yet provides no principled way of assessing interlanguage performance at early stages of acquisition in terms of the requirements of complex, real-world verbal tasks."

A performance-based approach to the construction of many of the tasks in the battery under discussion in this thesis was adopted, despite the reservation pointed out by Wesche above. It was adopted because of the positive washback effect of a valid performance-based test, which indicates to students as well as teachers that the purpose of language instruction is actually to prepare students for the world outside the classroom. This was found to be of significant motivational value in the study under discussion in this thesis. However, as Wesche (1986) points out:

> "Performance-based test construction requires considerable advance of 'front end' work: careful specification of objectives, identification and sampling of appropriate discourse types, content and tasks, and consideration of scoring criteria and procedures."

The methodology adopted to deal with these factors, as well as the way in which the test battery discussed in this thesis was subjected to a variety of analyses to establish its reliability and validity are discussed in greater detail in the following chapters.

10. Conclusion

In this chapter a number of important considerations in language testing theory and practice have been considered. The first one concerned the powerful influence that tests and examinations exert on teachers and students, an influence which is often perceived to be harmful. Teachers' attitudes towards testing in the field of TEFL were shown to be negative, and some of the blame was attributed to inadequate training. An important feature of the work described this thesis was the attempt to compensate for the problems mentioned above through a focus on teacher involvement in test design and construction, and the development of more training possibilities. In addition, students' attitudes to testing and what appeared in the tests were always a consideration. A fuller discussion of these aspects of test design takes place in Chapter 4.

Past and present trends in language testing were reviewed in order to establish

i. the context in which the tests discussed in this thesis were developed;

ii. why the tests in the battery under discussion appear as they do.

A number of decisions concerning the design of the test battery were made on the basis of past and present approaches to language testing:

i. For reasons of washback on classroom practice, and the highly functional uses to which the students concerned needed to put the English language a predominantly task/performance-based approach was adopted.

ii. A needs analysis approach (similar to ELTS and TEEP) to the specification of test content was adopted.

iii. Where possible the discrete-point approach was rejected in favour of an integrative one.

iv. Items were designed in such a way as to allow for a systematic and detailed analysis using classical testing theory.

v. The Canale and Swain model of communicative competence was of importance to the format and design of the tests in the battery.

vi. Morrow's descriptors of what constitute authentic communication were of use in the design of items.

Chapter III

1. Introduction

In Chapter 2 the relationship between teaching and testing was explored, and approaches to language testing over the last three decades reviewed. Chapter 3 investigates a number of issues of major relevance to the test constructor and their interaction with the teaching process. These issues include the classification of tests, how tests can be referenced, the importance of reliability and validity, how tests may be validated and the validation procedures adopted in this thesis. The chapter concludes on a cautionary note. It is pointed out that test performance can be influenced by many factors and that those interpreting test results should at least be aware of such factors even if they are powerless to do much about them.

2. The Classification of Language Tests

Many writers on Language Testing (Harrison, 1983; Harris, 1969; Heaton, 1975; Finocchiaro and Sato, 1983;

Davies, 1977) and testing in general education (Brown, 1981; Hopkins and Antes, 1985) offer definitions of different types of test. Davies (1977) for example, discusses four types of language test. The **Achievement** or **Attainment** test is summative (Bachman, 1981). It is concerned with attempting to establish how much has been learnt after a course of study. It can be school-based, an end of term or year examination, or it can be system-based like the GCSE examinations in the United Kingdom. **Proficiency** tests attempt to establish how much an individual "knows" regardless of teaching input. The TOEFL and Michigan are examples of proficiency tests. **Aptitude** tests attempt to establish how successful an individual might be in learning a language. The Modern Language Aptitude Test (Carroll and Sapon, 1959) and the Language Aptitude Battery (Pimsleur, 1966) are the best known generally available language aptitude tests. **Diagnostic** tests, generally teacher prepared, are formative (Bachman, 1981) and supposed to be used as a source of information for remedial action on the part of the teacher. Heaton, (1975) also mentions the **Progress** test which is intended to establish the amount of progress that an individual has made up to a specific point in any given course. The results may be used, like with the diagnostic test, as a possible source of information for teaching input.

While it is generally made clear that the distinctions between the different test types revolve around their function, it is fairly easy for the casual reader to conclude that the way in which tests are constructed and the appearance of the various test types will be significantly different by virtue of their function alone. This need not be the case. Indeed, it is quite possible for the same test to be used for more than one of the above purposes. The distinction between these test types is not one of construction procedures or content, it is largely one of the use to which the test is being put and the same test can be viewed in a number of different ways. The tests in the battery under discussion in this thesis therefore can be viewed in at least four different ways, depending on the nature of the questions that are to be asked about the tests themselves and the students taking them. For example, they can be categorized as a battery of achievement tests in that they all occur at the end of a course of study of either one, two or three terms and attempt to quantify the achievement of students. They can be classified as progress tests, in that the results are intended to measure progress over the course of study, and provide information for teachers which they may be able to feed back into the course.

They could also be classified as diagnostic tests since they are timed to occur three weeks before the end of term and teachers are encouraged to use the results as a source of information for remedial action. Alternatively, if we make use of Clarke's (1978) definition of a proficiency test, which is that it focuses on the ability of students to use language effectively for real life purposes, then the battery tests proficiency, regardless of its association with the teaching process.

While it is important to provide a framework by which tests may be categorized, since they will clearly serve different functions in different contexts, it is equally important to understand that this framework is not rigid. There are not necessarily any intrinsic differences between the form and structure of one test type and another. The same test can serve different purposes in different circumstance or alternatively, once administered, can reasonably be viewed in different ways. Indeed, in the context of a teaching institute it is more than likely that tests will serve several rather than one purpose. This situation will be further complicated if there is also a research element involved. While the project under discussion in this thesis was being conducted, it was found that if this

variability of function was made explicit it tended to confuse and antagonize the teachers who were required to administer and mark the tests. In general, therefore, it was considered expedient to present the tests as fulfilling a single function for the most part rather than a possible variety of functions.

The tests in this battery were classified as progress tests because their most obvious and pedagogically most beneficial function was to measure and provide some information about the progress of the students. Reinforcing the concept of "progress" was particularly appropriate in the type of teaching context in which the test battery occurred. Progress does not give an impression of finality and in an institute where most students' level of proficiency was rather low, finality was a concept to be avoided. It was important that both students and teachers perceived the tests as part of an ongoing process rather than as any sort of culmination. This was the main reason that the tests were not classified as achievement or attainment tests although they clearly played that role as well. From the research point of view they were also considered as proficiency tests, even though they arose out of a teaching environment. This classification was justified because many of the tasks in the tests were based on

real-world activities. An ability to carry out these tasks successfully was as much a measure of the proficiency of the student in using the language as it was a measure of his attainment after a course of study.

3. The Issue of Reliability and Validity

3.1. Fundamental Considerations

Regardless of the purpose of any test, there are some basic common sense conditions that it must satisfy. Firstly, a test must be consistent and trustworthy in its measurement. There must be minimal doubt that if a test were to be used again with the same students, it would provide more or less equivalent results. This condition is as important in the research context as it is in the educational context. In the former case because it is very dangerous to base any theory that might arise from a set of results on untrustworthy base data, and in the latter case because people's lives are affected by test results and there is a moral as well as an educational responsibility on the part of the test constructor and institute concerned to ensure that

these results are influenced as minimally as possible by matters attributable directly to chance. Any test therefore must be a reliable measuring instrument. Secondly, any test must test what it is intended to test, and must be seen to do so. In other words, it needs to be a valid measure. The two conditions of reliability and validity will be discussed at greater length in the sections below.

3.2. The Condition of Reliability

A test's reliability commonly refers to the consistency with which it is measuring whatever it is supposed to be measuring (Popham, 1981; Mehrens and Lehmann, 1984; Walsh and Betz, 1985). However, the condition of reliability is not essentially a statistical one and should not be perceived as such. Unfortunately in the eyes of many teachers it is seen as exactly that; an esoteric and somehow inhuman concept intended in some way to pervert the natural course of education. It is important to ensure that a test is reliable, that the results can be trusted, because many decisions of consequence may be based on test results. On the societal level progress to higher education and numerous training courses are based directly on test

results. It is important to be confident that a minimal amount of arbitrariness is responsible for these results. On the school level, streaming, if it exists, will be based to a very large extent on internal examination results or continuous assessment. Which ever may be the case, it is vital that the measures used are reliable. While one may object morally or educationally to the streaming practice, for example, if it exists then it should be as conducted as reliably as possible. Finally, on the classroom level, the teacher needs to be confident that the decisions made and the attitudes developed about students are not purely arbitrary. It is important that they are arrived at in a reliable manner.

The need for reliability of measurement therefore, is clearly an educational issue and not a statistical one. It is not required as the whim of a testing specialist nor as some perverse method of destroying effective teaching and it applies at all levels in the educational process. Statistical considerations are of consequence only when we try to establish the reliability of any measure.

Unfortunately, in the EFL field at least, although the point applies to a greater or lesser extent to all fields of Education, the testing specialist has been accused of allowing the need for reliability to overshadow sound educational practice (Morrow, 1979; Broadfoot, 1979) and this may indeed be the case in some instances. For example, strict adherence to a testing methodology such as multiple-choice, to the exclusion of other more realistic testing devices, simply because it allows for marking objectivity and ease of scoring is a potentially harmful approach to testing. The need for reliability of measurement while of vital importance, is only one of the conditions that a test needs to satisfy and though it cannot be ignored, it must not be perceived to dominate the rationale for test construction. The test constructor must be seen to produce items that are educationally acceptable in the sense that their washback effect does not influence teaching negatively. It is never adequate or necessary to justify an item that teachers see as potentially harmful to their classroom practice simply on the grounds that it is a reliable measure.

3.3. Estimates of Reliability

3.3.1. Standard Error of Measurement

The theory of reliability is based on the notion that although a subject may have a true score on any test, a test can only provide an estimate of that true score, which is unobservable. The true score is that portion of the individual's observed score not affected by random error. Since we cannot test an individual repeatedly it is necessary to have some way of estimating the possible variability of a person's scores on numerous testings. This is frequently done by calculating the standard error of measurement for a test (the formula used in the Item Analysis Programme developed for the analysis of this battery is presented in Appendix 8). The standard error of measurement provides a range of marks on a test within which an individual's true score is likely to fall around his observed score. There is a 68% likelihood that the true score will be in the range of one standard deviation above or below the observed score, and a 95% likelihood that it will fall within two standard deviations above or below the observed score.

The concept of standard error of measurement is an important one since it demonstrates the volatility of test scores, and the need to interpret them with care. The temptation to blindly equate true scores with observed scores should be avoided where possible. Unfortunately, while the testing specialist may be aware that a degree of latitude should be allowed for in the interpretation of test results, once a test comes into use there are pressures from teachers, administrators and students alike to attribute clear-cut distinctions to results. For example, there is generally a requirement for a passing score. When the battery under discussion in this thesis was introduced, no passmark was specified to the consternation of many teachers. A passmark was not specified precisely because it was accepted that there may be variation between true scores and observed scores, and that teachers should have some say in the final grading of students since they had spent a whole term teaching them. Even though the tests were demonstrated to conform to high standards of reliability (see Chapter 5), they were not and could not be absolutely accurate. The attempt to share the responsibility for final grading met with some resistance from teachers, and was often perceived as a sign of weakness rather than strength. While it will be impossible in certain

contexts, such as a public examination, to allow for this type of flexibility, within a teaching institute it is desirable practice.

The concept of standard error of measure of measurement allows for a numerical realization of the range of possible variability of true and observed scores. It should not be presented as a statistical notion to teachers since this will arouse needless antagonism. It should simply be made clear that test results are not an absolute measure, and that some flexibility in their interpretation may be required and, based on the standard error of measurement, some crude estimates of the range of this flexibility may be provided.

3.3.2. Measures of Equivalence and Stability and Internal Consistency

The reliability of a test can be estimated in a number of ways, and the method selected depends on the context and requirements of any given situation.

Reliability can be estimated by establishing the extent to which a test is equivalent to another test that is purportedly measuring the same things in the same way. Two equivalent tests can be administered in close proximity to the same group of students and the results correlated. By this method we are determining how confidently we can generalize a person's score to what he would receive if he took a test composed of similar but different items. While this is not a very convenient way of estimating reliability, equivalent measures can have important uses in the determination of concurrent validity, for example.

Reliability can also be estimated by establishing the stability of a measure. This procedure is generally called the test-retest method, where the same test is administered to the same group of subjects on two separate occasions and the results are correlated. The higher the correlation, the higher the reliability of the measure. This method is somewhat unsatisfactory because the results can be affected by practice effect, and how much a subject remembers from the first administration. It is also a very cumbersome procedure.

The third and most satisfactory way of estimating reliability is to determine the internal consistency of a test. Conceptually the simplest estimation of internal consistency is the split-half method. This involves splitting the test in some way, normally odd and even question numbers, and then correlating the two halves as if they were two separate tests. The higher the correlation, the higher the reliability. Although the split half method is theoretically the same as the equivalent forms method, it is generally considered a measure of internal consistency because the two equivalent forms are part of the same test. Since the two tests are shorter than the whole test the Spearman-Brown prophesy formula may be applied to the results to compensate. If no better method of estimating reliability is available, then the split-half method may be used although it is not as sensitive as other measures of internal consistency due largely to the difficulty of demonstrating that the two halves are actually equivalent, even though they come from the same test.

The Kuder-Richardson formulae, commonly used as estimates of reliability, are based on estimating the internal consistency of a test. There are two formulae - KR-20 and KR-21, the distinction between them being

that the latter assumes all items to be of equal difficulty. If this assumption is not met, then KR-21 will give a slightly lower estimate of test reliability (Mehrens and Lehmann, 1984). Since KR-20 makes use of the information available on the difficulty of individual items it is a more sensitive formula to use, and was employed in the Item Analysis Programme developed for the test battery under discussion in this thesis. The formula itself is presented in Appendix 8.

The Kuder-Richardson formulae are used with dichotomously scored data. The Coefficient Alpha (Cronbach, 1951) is an alternative for items that are not dichotomously scored. It is a useful formula to use with essay type questions for example, when the scoring might be on a continuous scale.

3.3.3. Factors Influencing Reliability

Several factors can affect reliability and they need to be borne in mind in the interpretation of reliability estimates.

Firstly, reliability is generally higher with a longer test. This is so because random positive and negative errors within the test have a better chance of cancelling each other out with a longer test thus making the observed score closer to the true score. The most usual way of compensating for this is to use the Spearman-Brown prophesy formula. This formula was not used with any of the reliability estimates generated in this thesis since they were all sufficiently high, despite sometimes short test length. Had it been used, then the reliability estimates would have been higher.

Secondly, tests are sometimes classified as either speed tests or power tests (Hopkins and Antes, 1985; Mehrens and Lehmann, 1984). The results of a speed test are dependant on the number of items a candidate attempts, where the expectation is that nobody will finish all the items in the test. All the items will be very easy, and so the score is a result of processing time rather than knowledge per se. A power test is one where every candidate has time to attempt all the items but because of their relative difficulty no one generally obtains a perfect score. Most tests are in fact a combination of both of these test types. However, reliability is affected if many candidates do not have time to finish. The test battery under

discussion was designed as a series of power tests although inevitably, some candidates did not finish. This may have had a marginal effect on the reliability estimates. The problem could have been avoided to some extent had the order of some items been reversed. However, under the circumstances this was not possible.

Thirdly the homogeneity or otherwise of a group of testees affects test reliability. In principle, the more heterogeneous a group, the higher the reliability of a test will be. The students taking the different progress tests in the battery under discussion were pre-selected on the basis of a placement test and were thus intended to be a fairly homogeneous group. However, due to a number of unavoidable factors, they appeared to be rather more heterogeneous than anticipated.

Finally, reliability is affected by the difficulty of the individual items in the test. A very easy test, for example, will inevitably produce little variability amongst the scores as will a very difficult test. This will consequently reduce reliability. The tests in the battery were all of moderate difficulty which resulted in a better environment for higher reliability.

3.4. The Condition of Validity

The concept of validity is generally defined as the extent to which a test measures what it is supposed to measure (Pratt, 1980; Popham, 1981; Priestly, 1982; Carroll and Hall, 1985), or the extent to which it provides information relevant to the decision that is to be made on the basis of the test results (Thorndike and Hagen, 1977). We can attempt to establish validity in a number of ways: by comparing test content with the syllabus on which it was based; by comparing the results of a test purportedly measuring a particular skill or trait with another already established test measuring the same trait; by comparing test performance with eventual success in a given area; and by attempting to establish whether there are grounds to claim that the skills being tested do indeed reflect the competence theory that underlies a given test.

3.4.1. Is there as Reliability/Validity Conflict?

Various writers on Language Testing (Morrow, 1979; Underhill, 1982; Davies, 1978; Weir, 1983) suggest that there will be an inevitable conflict between reliability and validity in test construction, particularly with reference to oral and written tests (Underhill, 1982) and communicative tests (Hawkey, 1982).

From an educational and moral perspective, a focus on either reliability or validity at the expense of the other is quite unacceptable. It is true that a test can be reliable and yet completely invalid. It is equally true however, that a test cannot be valid if it is not reliable (Walsh and Betz, 1985). Unless consistency of measurement can be established, none of the various validation procedures involving statistical techniques can be used with any degree of certainty. It may be that the content of a test can be shown to be relevant, on the basis of a syllabus or some such inventory, however, it cannot necessarily be demonstrated that the syllabus itself is appropriate on any grounds other than intuitive ones. This is not to say that intuition has no place in the teaching and hence testing process,

since it clearly does. But intuition is cheap, in the sense that everyone has it, and it differs significantly from individual to individual, teacher to teacher, student to student, or materials writer to materials writer and thus it is very dangerous to follow the Morrow (1979) line when he writes:

> "Reliability, while clearly important, will be subordinate to face validity. Spurious objectivity will no longer be a prime consideration ..."

Achieving acceptable standards of reliability is difficult no matter what sort of test is constructed, and no item or item type can claim to be reliable simply because it is widely used. Recent approaches to Language Testing have moved towards unfamiliar formats where there is not a large body of literature and experience to draw on, and consequently item types and construction procedures have to be proven to be effective. There is no reason why this should not be done, indeed it is imperative that it should be. To suggest, as Morrow does, for example, that his approach is valid almost by default because it looks good, seems to be more of a defensive strategy than a statement of justifiable conviction. The claim that validity is ultimately not dependant on reliability allows the test constructor total freedom to produce whatever he wants to and justify it. This is surely more dangerous than

the status quo that Morrow identifies and criticizes so vehemently.

In preparing the test battery under discussion in this thesis, it was accepted that achieving reliability and validity was bound to be difficult due in part to the untried nature of most of the item types and in part to the fact that satisfying these two conditions is always a problem. It was also accepted that both of these conditions had to be met on both educational and moral grounds. Therefore, extensive moderation and pretesting procedures were employed to achieve both reliability and validity at the expense of neither.

3.4.2. Aspects of Validity and how they can be Established

Many different ways of establishing validity have been put forward and the literature on the subject is sometimes confusing and contradictory. This being the case, the initial discussion here will use the basic guidelines laid down by Standards for Educational and Psychological Tests (1974) and supported by Popham (1981), Cronbach (1970), and Thorndike and Hagen

(1977), which suggests three basic types of validity: **content**, **criterion-related** and **construct validity**. In addition to these, much attention has focused recently on **face validity**.

3.4.2.1. Content Validity

Content validity, sometimes referred to as the principle of "inclusiveness" (McCormick and James, 1983) concerns the extent to which a test covers the content that it is supposed to. This aspect of validity is very important in the context of achievement and progress tests at the classroom and school level (Deale, 1975). If we are considering an achievement test, the content to be covered will be found in the syllabus and course books. Content validity can be established by comparing what is in the syllabus, for example, and what is in the test. In theory, the closer the match. the better the content validity. If proficiency is being tested, then it tends to be the test constructor who defines the content of the test. Moller (1982, p.37) gives some thought to this issue:

> "Content validity, together with reliability, will ensure that a test adequately reflects the objectives and linguistic content laid down in the syllabus. In the case of a proficiency test, however, the test

> constructors themselves decide the 'syllabus' and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe. Thus the evaluator looking for content validity is really assessing the test constructors' definition of proficiency."

In fact, the same argument can be used with regard to achievement testing. Somebody, somewhere, generally designs a syllabus and, on the basis of this, materials writers prepare textbooks. The appropriacy of the content of the syllabus, and the extent to which the materials then reflect it are subjective decisions for the most part. Thus, the Strategies Series, for example, is based on the Council of Europe syllabus. As a course it may or may not adequately reflect the syllabus. Institutions will then use the Strategies series and may even write tests to accompany it. These tests may be valid in the context of the materials, but there is no guarantee that the materials are actually valid in the context of the syllabus nor indeed that the syllabus itself is valid.

The integration of published materials with an institute specific approach is discussed in greater detail in Chapter 4. Course outlines at the British Council, Hong Kong, were ultimately designed with attention to the real world needs of the students as

well as the published texts that originally served as the major source of teaching materials. The battery of progress tests was introduced at the same time and reflected the changing focus of the institute's approach to the teaching of English. The tests were in line with the thinking of some teachers, but it was anticipated that they would provide the impetus for the majority to make some changes to what and how they taught. This type of situation is mentioned by Thorndike and Hagen (1977, p.59):

> "... the relationship between teaching and testing is typically intimate. Test content has been drawn from what has been taught, or what is proposed to be taught. The instructional program is the original source of test materials. Sometimes the thinking in a test may lead the thinking underlying a local course of study, as when a group of specialists have been brought together to design a test corresponding to some emerging trend in education."

It seems therefore, that content validity can also be based on what may be expected to be taught as well as what is actually taught, and that a test battery can have content validity if it is intended to help in a change of teaching focus. To some extent this was the case with the test battery under discussion in this thesis.

3.4.2.2. Criterion-related Validity

Criterion-related validity refers to the extent to which a test is predictive and/or concurrent i.e. measuring the same thing as another, already proven test. The most usual way of establishing criterion-related validity is by correlation.

To establish the concurrent validity of a test it is normal that students' results on another test purportedly measuring the same trait as the test under investigation are correlated with that test. The problem arises in that it is difficult to show conclusively that two tests are indeed measuring the same things in the same way. This has long been perceived as a difficulty but recent research into trait and method effect (Bachman and Palmer, 1983; Shohamy, 1984) for example have made it very clear that there can be no guarantees that methods are equivalent. Thus, when new item types are developed, it is very questionable that traditional concurrent validation procedures are applicable.

It is sometimes the case that concurrent validity can be demonstrated by correlating test results with either teacher ratings or grades (Chaplen, 1970;). Ingram (1974) goes as far as to say that teachers' ratings are the best method of establishing the concurrent validity of a test. Students' scores on tests and subtests under discussion in this thesis were in fact correlated with four types of teacher assessments as well as placement test scores in order to gain some indication of the extent to which the measures agreed. The correlations, discussed more fully in Chapter 5, were generally moderate. This was not a surprising finding in the sense that defining criteria for teachers to use is generally a problematic issue (Moller, 1982, p.52), as is the standardization of teachers' views of what those criteria mean.

If establishing concurrent validity is fraught with difficulties then so is determining predictive validity. Many studies have focused on the academic context since this is where most students take proficiency tests, and where most data is available for investigation. Moller (1982) reviews a number of predictive validity studies and makes the point that most of them have used non-linguistic criteria, such as Grade Point Averages in their work. Numerous factors

can affect the results of a predictive validation study such as the amount of exposure that subjects get to the target language, their willingness and ability to improve, the extent to which language issues are focused on by their tutors and so on. In addition, in many predictive validity studies the samples used are unavoidably biased in the sense that studies can only be carried out on subjects who achieve the sort of score that allows them to go to a university in an English speaking country. All those who do not achieve the appropriate grade are excluded. Whether some of them would have succeeded is open to question.

Establishing the predictive validity of the test battery under discussion in this thesis was not considered a feasible course of action since determining the criteria that should be used was perceived as an insurmountable problem. No further reference is therefore made to the issue of predictive validity.

3.4.2.3. Construct Validity

To satisfy the condition of construct validity a test must be shown to measure the psychological constructs that it is hypothesized to be testing. It is a basic assumption that tests are intended to provide us with information on some real world phenomenon, characteristic or behaviour. They are generally an indirect or operational way of attempting to describe the extent to which individuals possess a theoretically postulated characteristic or construct.

The process of construct validation, according to Walsh and Betz (1985) may be broken down into three stages:

> "First, the construct of interest is carefully defined and the hypotheses regarding the nature and extent of its relationships to other variables are postulated. Second, an instrument designed to measure that construct is developed. Third, after the degree to which the test is reliable has been examined, studies examining the relationship of the test to other variables (as formulated in the hypotheses about the construct of interest) are undertaken."

Numerous statistical techniques are available to investigate construct validity, although one of the most popular ones in recent years has been the factor

analysis of the intercorrelations of subtests or items in order to establish how many dimensions or traits may be required to summarize or explain test performance. For example, if a test is designed to measure more than one trait and yet a factor analysis yields a single "general" factor then, from this point of view at least, the test could not be said to possess construct validity.

In language testing research construct validation attracted relatively little interest until the late seventies when suddenly a spate of research, mostly in the United States, began to focus on it. Most of this research can be viewed , according to Weir (1984, p.65):

> "... principally as the a posteriori statistical validation of whether a test has measured a construct which has a reality independent of other constructs. The concern is much more with the a posteriori relationship between a test and the psychological abilities, traits, constructs, it has measured than with what it is that it should have elicited in the first place."

As important as the a posteriori validation of a test is the a priori establishment of the appropriateness of test content. That is to say, there needs to be a clear definition of what is to be tested before we can

attempt to establish that it has been tested. Ebel (1983) points out in fact, that a major reason for problems of test validation is an overemphasis on the need for empirical validity data, and a failure to recognize the primary importance of explicit verbal definitions of what the test is intended to measure.

A greater focus on a priori validation inevitably leads to an overlap between content and construct validity. With the test battery under discussion in this thesis, for example, it was hypothesized that a students ability to use English was not necessarily the same with different traits, and that additionally, performance was micro-skills based. In order to validate these hypotheses, very explicit descriptions of test content had to accompany test construction and after administration the tests had to be subjected to a series of correlational and factor analyses.

3.4.2.4. Face Validity

Although the concept of face validity is an important one, it is not generally included as one of the three

major types of validity due to the questionable role that it plays in the validation process.

Face validity, the extent to which a test looks as if it is testing the right thing in the eyes of students, teachers, and sponsors has also received much attention in discussions about the validity of language tests. There is no generally accepted procedure for determining whether a test has face validity, and this has led some to argue that it should have no place in the discussion of test validity (Bachman et al., 1981).

Stevenson (1985), while supporting the movement in language testing towards more performance based formats warns that:

> "Face validity is the mere appearance of validity to the metrically-naive observer. It provides the psychometrically unsophisticated self-assurance that allows someone simply to look at a test and, without further technical examination, conclude: "I know a valid test when I see one." "

Stanley and Hopkins (1972, p. 105) also point out that face validity is very much on a naive and superficial level and that it is dangerous to attribute too much importance to it. While a test with good content

validity will generally have face validity, the reverse is much less likely to be true.

There can be no doubt that a test should look as if it is testing the right things in the right ways, whatever that may mean. However, there can be equally little doubt that it is inappropriate to approach test validation primarily from the angle of face validity which should naturally arise out of a real concern to ensure that the content of a test is valid.

4. The Problem of Referencing a Test

A distinction is drawn in educational measurement between the norm-referenced test and the criterion-referenced test. The basic distinction between these two approaches to the referencing of test scores is quite straightforward. A norm-referenced test is one where the ability of a student is measured in terms of his peers, whereas a criterion-referenced test is one where the main interest focuses on whether or not a student can successfully carry out a task or series of tasks. No reference to the ability of his peers is required. Unfortunately, a great deal of confusion has

arisen as to the difference between these two approaches.

Norm-referenced measurement arose out of a tendency on the part of those interpreting test scores to treat them as absolute values (Mehrens and Lehmann, 1984). For example, if Student A gets a score of 62% on a grammar test, where the passmark has been set (probably arbitrarily) at 60% then his performance can be categorized as adequate. On the other hand, if he gets a score of 58% on a reading test with the same passmark, then his performance would be classified as unacceptable. While this type of inference is common, it is inadequate because it assumes that the test scores in themselves have some absolute value, which they probably do not. Two assumptions are made: that the amount of knowledge needed to pass a test can be determined; and that it is the same in the case of both tests. Each of these assumptions is open to question.

Norm-referencing offers a solution to the type of problem mentioned above. A score is given meaning by comparing it to the scores of other students taking the same test. So, in the case of the example stated above, the score of 62% may have put the student in the 45th

percentile, whereas the score of 58% may have put him in the 65th percentile. The knowledge of an individual is quantified in terms of the knowledge of other individuals in the same peer group.

The norm-referenced solution to the situation described above seems reasonable enough. However, it has often been criticized. A critic of norm-referencing might claim that the use of percentiles alone does not of itself guarantee to provide us information on how much a student has actually learnt or knows, and few would argue the point. In fact, in most educational contexts, with the exception of large-scale standardized batteries, test scores are usually interpreted in more than one way. We may indeed say that Student A is in the top ten percent, but we are also likely say that he achieved a score of 75%, and we are further likely to assert that this score is adequate, minimally adequate or excellent. In this sense, we would be interpreting the results from a criterion-referenced as well as norm-referenced point of view. Most teachers interpret test scores on both ways.

In the minds of many teachers, and some test constructors a norm-referenced test is "traditional" in

the sense that it always uses multiple-choice items or something similar whereas a criterion-referenced test is performance-based, in that the items used reflect some sort of real-world type of activity. In other words, many people believe that there is a fundamental difference in appearance as well as purpose of tests being used in a norm-referenced or criterion-referenced way. However, as Glaser and Nitko (1971, p. 654) point out:

> "The distinction between a norm-referenced and a criterion-referenced test cannot be made by simple inspection of a particular instrument."

The same test can be used for either purpose: to present student ability on a continuum or to separate students on the basis of whether they can or cannot perform a particular task.

Glaser (1963) was the first to use the term 'criterion-referenced test'. Ever since that time there has been some dissatisfaction with the word 'criterion' because of its ambiguous meaning. The first interpretation might be called the 'criterion-as-a-desired-behaviour' conception while the second would be the 'criterion-as-a-level' conception (Popham, 1981) where criterion refers not to behaviour but rather a desired level of

proficiency. In the first sense, criterion-referencing is clearly distinguishable from norm-referencing and has an important educational contribution to make. In the second sense, it does not offer any substantial advantage over norm-referencing. It is quite possible, for example, to transform any norm-referenced test into a criterion-referenced test simply by setting a specific proficiency level.

This problem over the interpretation of the word 'criterion' has led to the suggestion that it be replaced by 'domain' which does not carry with it the unfortunate connotations concerning levels of proficiency. However, while most experts agree that the term domain-referencing is more appropriate (Popham, 1981; Mehrens and Lehmann, 1984), there has not actually been an agreement to adopt it and abandon the term criterion-referencing.

The test battery described in this thesis may be described as criterion or domain-referenced in the sense that the items were selected on the basis of an investigation of the domains of language use that the students encountered inside and outside the classroom. However, the item analyses drew on norm-referencing

techniques. With regard to the interpretation of results, it may be said that they were criterion-referenced in the sense that we were not particularly interested in the relative ability of the students to each other but were rather concerned with the level of proficiency. In addition the completed scripts were intended to serve as diagnostic measures for the use of teachers in a formative sense.

5. The Characteristics of a Good Test

From the discussion in the sections above it is apparent that for a test to be considered a good test it needs to meet various conditions of reliability and validity. However, there are a number of practical considerations that are of importance that will be discussed briefly below.

It is important that a test be economical and practical to administer. The amount of printed matter and any other additional materials associated with the test need to be kept to a reasonable minimum otherwise costs can be very high thus arousing the antagonism of administrators.

Marking time needs to be carefully considered so that it is not unnecessarily excessive. If it is, then it will arouse antagonism on the part of markers, who will in all probability be teachers in many instances.

What students have to do needs to be considered carefully so that they do not feel that they are being made to jump through unnecessary hoops. In addition it is unwise to make a test too long, since then variables such as fatigue and boredom will have a tendency to affect students' scores in unpredictable ways. Test administration needs to be as practical and efficient as possible to avoid antagonism from both teachers and students and also reduce the possibility of mistakes being made by those administering the test.

6. Additional Considerations for Test Constructors

The interpretation of test results is one of the most troublesome aspects in testing. Attempting to satisfy the conditions of reliability and validity discussed above provides a degree of confidence but numerous

other factors interfere which, while they may be difficult to deal with, should at least be considered. For example, how important is the testing method with regard to student performance, to what extent does the background of the learner affect how well he performs, is cognitive style significant, how should we deal with the variability of input and output, and what is the effect of test anxiety on performance? These matters are considered below.

6.1. Item and Task Format

Quite a lot of research has been carried out over the last few years on the effect of different testing methods on the performance of students.

Two major factors might be said to interact in the process of language testing. These factors are generally referred to as trait and method. **Trait** pertains to the knowledge that a test is trying to measure such as writing, grammar, listening etc., whereas **method** is the way in which this knowledge, skill or trait is being measured (multiple-choice, cloze etc.). There are many methods and procedures

available to test any given trait. However, the effect of these methods on the trait being measured and consequently students' scores is open to question. In the worst possible case, the results generated by a test could be based largely on the effect of method rather than on the trait that we are trying to measure. Thus, one characteristic of a good test is that the method has little or no effect on the trait.

One way of finding out if, and to what extent testing method affects performance is to use multiple methods to measure the same trait and then compare the results in order to see if there is a method effect. The multitrait-multimethod matrix proposed by Campbell and Fiske (1959), is a common way of doing this. Research using this model (Bachman and Palmer, 1983) has shown that method can and sometimes does have an effect that is more powerful than trait effect. In other words, we might on occasion be testing the ability to deal with multiple-choice questions rather than reading, for example.

Research into the cloze procedure (Alderson, 1983) has shown that students perform differently depending on the deletion rate, the difficulty of the text, and the

way in which words are deleted, thus indicating that different students may be favoured by different methods. Bachman (1983) investigated performance variations when deletions in cloze passages were either fixed-ratio or rational. He was able to show that difficulty levels and factor structures were not the same.

Shohamy (1983) looked at the effect of method in the assessment of oral proficiency. She found that the elicitation procedures used tended to affect the students' scores on an oral interview speaking test and that there was a low probability of a student scoring the same mark with on an identical method if two different interviewers were used. In addition, when the speech style was changed from interviewing to reporting, scores varied even more drastically.

Shohamy (1984) attempted to find out if there were differences in reading comprehension testing methods. She designed an experiment where several groups of students took a reading comprehension test using two methods, open-ended and multiple-choice. She also set the questions in English and Hebrew which was the

native language of the students concerned, ending up with what was in effect four separate tests.

The results showed that the multiple-choice tests were consistently easier than the open-ended ones. This may be due to the fact that different skills are required of the students. For the multiple-choice test a student had to understand and select, whereas for the open-ended test the student had to understand and produce. It was also the case that when the questions were in Hebrew, students did better. However, Shohamy was able to show that the language used for the questions, and the test method were much more significant factors at lower levels of proficiency than they were at higher levels.

Test constructors need to be aware that different methods of testing the same trait can yield different results. In practice, however, it may be difficult to gauge in what way and to what extent this is the case. A compromise solution to this problem, for the person involved with testing at the practical level, is to try and make sure, where possible, that a variety of methods are used in testing the various traits, in the

hope that the effects of different methods will cancel each other out.

6.2. The Learners' Background

The learners' background as a factor in test performance was pointed out by Carroll as early as 1961 but has been consistently ignored by test designers ever since due to the enormous complexity of considering background variables in the interpretation of test results. Farhady (1982) carried out some research in this area. He compared the test performance of university students with a number of background variables such as sex, university status (graduate or undergraduate), major fields of study and nationality. He found that there was no significant difference in test performance whether the test takers were male or female. On the other hand, university status did make a difference. Graduates performed better on cloze, grammar and reading subtests than did undergraduates while on the listening subtest, undergraduates outperformed graduates. Farhady accounts for this by hypothesizing that graduates would have had more practice in areas like grammar and reading than undergraduates. The fact that undergraduates performed

better on the listening subtest is accounted for by factors such as age, length of stay in the United States and recent changes in educational systems which place more emphasis on oral/aural skills. Farhady also found significant differences in test scores between students in different fields of study and from different nationalities. He claims this may be due to different educational policies in different countries. One of the conclusions that Farhady draws is that:

> "... ignoring all these factors, simply by defining language proficiency as a concept independent of learner variables, seems unjustified ... it could be assumed that test taker characteristics were factors which resulted in different performance patterns. Thus, if some of these variables could be incorporated in testing programs, it would be a step in the right direction."

Cziko (1982) defines the term language background very narrowly to refer to the type of contact that test takers have had with the English language and the amount of opportunity they have had to acquire the various skills in English. He writes:

> "... the pattern of results of language tests administered to a group of second-language learners can be meaningfully interpreted only in the light of the language background of the group. Instead of focusing solely on the pattern of test results, patterns of test results should be compared with the language background pattern of the group. If this is done, then we may well find that what is often taken as evidence of either a one-factor or multi-factor working model of communicative competence may instead be

> simply an indication that the pattern of language proficiency one acquires is related to the type of exposure to the language one has had."

A certain amount of preliminary research was carried out on the relationship between background variables such as sex, age, education, occupation and length of enrollment at the institute and test performance using data generated by this thesis. This research will not be discussed in detail, since it is beyond the scope of the current investigation, however, the major findings and their implications will be mentioned.

An analysis of variance of placement and progress test scores with the background variables mentioned above revealed that the sex and age of the student did not produce significantly different results. However, students with a higher level of education did perform significantly better on all tests. This information was useful from the placement testing point of view. In addition, there were significant differences in performance depending on occupation. The job that students do is often a reflection of their level of educational and so it was not surprising that there should be differences in performance. Finally, it was found that students who had been registered with the

institute longer tended to perform less well than students who came straight into a particular level.

The background of the learner is an area that is still largely neglected in test design. Exactly how the test designer is to incorporate language, educational, social or cultural background in the design of tests and the interpretation of results is not clear. What is clear however, is that the test designer needs to be aware that background variables are likely to have an effect and that if possible, this should be borne in mind when constructing tests and interpreting results.

6.3. Cognitive style and Language Testing

Common sense informs us that different people think in different ways and that this fact may have an influence on test performance and the ways that results should be interpreted. Field dependence-independence refers to individual differences in preferred ways of perceiving, organizing, analyzing, or recalling information or experience. It is claimed that a field dependent person is one who has a tendency to rely on external frames of reference in cognitive activities and foster skill

in interpersonal relationships. On the other hand, a field independent person is thought to rely heavily on internal rules or strategies for processing information and have more developed mental restructuring abilities. Most researchers in language testing have not seriously considered the degree to which a cognitive style construct such as field dependence-independence may affect the test takers' performance on language tests for the very good reason that making use of any information that might be generated would be extremely complicated. Stansfield and Hansen (1983) have done some work in this area. They compared the degree to which field dependence-independence had an influence on test performance as measured by a variety of proficiency tests including cloze. Working models of communicative competence have consistently used the cloze test as a source of data in order to establish the existence of a global language proficiency factor. Stansfield and Hansen found that the field independent cognitive style was associated with a higher level of proficiency on all the measures of second language proficiency that they used. However, they also found that cloze test performance was influenced to a greater degree by a field independent cognitive style than were the other measures that they used. This research indicates that test performance may be influenced by

cognitive style and that this may not be reflected in a person's ability to communicate in other situations.

There is some evidence therefore, that cognitive style may affect test performance and in consequence, could affect the results of studies that try to attribute certain factor solutions to underlying processing mechanisms. The effects of cognitive style are not investigated in this thesis. However, it is accepted that cognitive style may play a role in performance and that this might ultimately need to be taken into account in the interpretation of test results even though it is unclear exactly how this may be done at present.

6.4. Variability and Language Testing

Students' language is variable dependant on context and interlocutor. Of this there can be little doubt. The extent and systematicity of this variability is not as yet very clear. However, variability is bound to have an effect on performance to some extent, and is thus a relevant consideration in the interpretation of test

results. Like the other areas mentioned above, it is difficult to gauge how it can be taken into account.

Let us take as an example the role of accommodation theory in second language acquisition with regard to the nature of the variability of the language produced by second-language learners. Accommodation theory suggests that people adjust their speech in order to express their values and their intentions to their interlocutors. In other words, if speaker X, for example, wants to win speaker Y's approval, speaker X will sample from speaker Y's speech, and from it infer Y's personality characteristics and values. Assuming Y approves of these characteristics, X (largely unconsciously) chooses from his repertoire aspects of speech to project Y's characteristics (Giles and Powesland, 1975). Speakers can adjust towards the interlocutor, in which case, the shift is called "convergence". If they shift away (to maintain or assert distinctiveness) it is termed "divergence".

Beebe (1983) carried out some research into the implications of accommodation theory to second language acquisition and while the findings are tentative they indicate that more research is required in this area.

Beebe gathered data from a group of third grade Puerto Rican children who were enrolled in either the bilingual or monolingual program of an elementary school on the edge of New York city. Each subject was interviewed in English three times, each time by a different interviewer. The interviewers, close in age, were all middle class women, born in the United States, and had the same amount of education. One interviewer was monolingual, native English-speaking, one was English-dominant Hispanic, and one was Spanish dominant Hispanic. Beebe found that not only was there less talk with the English-dominant Hispanic but also that certain phonological features appeared in the subjects speech only in interaction with this interlocutor. Beebe speculates that this is due to less convergence with the English-dominant interlocutor because the subjects did not identify with her to the same extent as they did with the other two. She may have been seen as a sort of traitor by the subjects. Beebe carried out another study along the same lines with Thai and Chinese speakers, with similar results.

In a second-language context such as Hong Kong the degree to which students do or do not accommodate to interlocutors, be they oral examiners or teachers, and the extent to which this affects performance should be

of concern. Hong Kong currently has colonial status, however, in 1997 it will be reunited with China. This reunification is regarded ambivalently by many of the Hong Kong residents. While we can be certain that the students registered with the British Council perceive the need to improve the standard of their English, their conscious and subconscious attitudes towards that language and the people who speak it are by no means clear. Similarly the effect that this may have on test performance is equally unclear. The test constructor, as well as the teacher, need to be aware that there are potential difficulties in this area although how they can be quantified and or neutralized is open to question.

6.5. Test Anxiety

Anxiety, an emotional reaction, has been the focus of much research in trying to establish its effect on test performance (Speilberger, 1966; Madsen, 1982; Trungamphai, 1982). Experts divide anxiety into two types. **Trait anxiety** is a fairly stable personality characteristic, and it is not dependant on the type of examination or test, but rather on the personality of the individual. **State anxiety** tends to fluctuate in

response to different stimuli. Thus state anxiety can be affected by the type of task or question that occurs in an examination. Studies on the effects of anxiety produce a range of results. For example, girls tend to manifest higher test anxiety than boys, and persons with low anxiety tend to outperform those with high anxiety.

However, anxiety should not necessarily be considered as having simply a negative effect. It has been established that there may facilitating as well as debilitating anxiety (Alpert and Haber, 1960). Studies have investigated these two constructs and it seems that they can be related to a student's general outlook on life and performance at school.

The fact that time constraints can have an effect on test performance is clearly indicated by Hill (1983). Hill found that when children were under time pressure, those identified as high test anxious made three times as many errors and took twice as long as low anxious children. However, when time limits were removed, the high test anxious children performed as well as their low test anxious peers and completed the test in approximately the same amount of time. Hill points out

that the performance of the high test anxious children in the first testing situation was most probably limited by anxiety resulting from time constraints, and not by low achievement in the subject being tested.

Madsen (1980) tried to establish the amount of anxiety produced by different types of test question. He showed that the oral interview created the least amount of anxiety, while reading comprehension and cloze the most. He argued that the anxiety generated by the reading test appeared to be based on the complexity and difficulty of the items, in particular the distractors.

In another study, Madsen (1982) examined the debilitating effect of anxiety on test performance on ESL examination batteries. He concluded that the anxiety generated by a high anxiety-producing subtest was due to the complexity and difficulty of items. He commented further that anecdotal accounts and research indicated that in addition to the form of the exam (reading comprehension, cloze, etc.) faulty instructions, lack of face validity, difficulty level, insufficient time and cheating by other students caused anxiety on the part of many students.

The implications of this type of research to teachers and test writers is that there is a need to be aware that people are different and that differences in personality and personal circumstances affect the amount of anxiety that they may bring to a testing situation. This anxiety can be debilitating or facilitating. Test constructors, administrators and teachers can try to make sure that the way they approach testing does not disadvantage students who are adversely affected in their performance by anxiety. We can try not to make time a significant factor in the test. This is not to say that a test should go on forever, but it is clear that if there is too much in a test we are not necessarily giving the able students a chance to show how good they are. Anxious students will perform worse not necessarily because they know less but because they are put off by the time factor. We can try to make our instructions as clear, correct and as uncomplicated as possible. The test should look good, and as if it is testing the "right" thing. In other words, face validity should be a consideration. We need to be aware that certain types of test item, such as multiple-choice reading comprehension and cloze, may produce higher anxiety levels than some other types of item. With multiple-choice items we should try to make sure that the distractors are not too complex and

tricky simply for the sake of it. Such considerations were considered important in the design and administration of the battery under discussion in this thesis.

6.6. Concluding Comments

Research in areas related to language testing is of importance to the test constructor. There may not always be ways of incorporating the ideas and implications generated by the research into tests at the moment and maybe there never will be. However, research in related fields indicates strongly that many factors are at play when it comes to interpreting the results of language tests. The language tester and researcher always need to be aware of the frailty of results and the scope of the problem that they are facing. Perhaps most important of all, the test constructor and teacher need to be aware that test takers are individuals and as such they are different. Tests tend to bury these differences and make them seemingly less important. Individuality is submerged. While this is unavoidable, we should at least be aware of the fact.

7. Conclusion

The first three chapters of this thesis have attempted to establish the basic principles and considerations underlying the construction of the test battery. The writer is fully aware that he is dealing with a complex set of problems and issues that underlie test construction. Statistical techniques are used in subsequent chapters as a means of clarifying some of these. However, it is fully accepted that they are fragile and very limited in their scope. This does not of course invalidate them in any way. Any research project, particularly in the field of Applied Linguistics, is relevant to a specific context. There is no reason why it should necessarily have wider relevance.

CHAPTER IV

1. Introduction

In this chapter I will provide a detailed description of the context in which the tests discussed in this thesis were developed. The chapter will include a survey of students who were registered with the British Council Institute at the time. This is followed by a discussion on the nature of the course development which precipitated the construction of the test battery. A review of the student placement procedure is then carried out in order to show that it was both reliable and valid in the context. Next there are two sections that deal with the design of the battery, and the selection of test items and general format. Section eight gives a detailed account of the content of one of the seven tests which is representative of the others. This is followed by an account of the stages of test preparation with particular reference to the important role that the training and familiarization with testing principles and practice play in the context of a language teaching institute. Finally, Section 10 describes the implementation of the tests and Section 11 the teacher assessments. These were introduced in

order to temper the objective test scores with a grade from the teacher and as a possible means of validating the tests themselves.

2. Background to the design of the test battery

The design of language tests and test batteries has received much attention over the years (Lado, 1961; Harris, 1968; Ingram, 1968; Heaton, 1975; Davies, 1976; Morrow, 1979; Farhady, 1981 ; Carroll, 1982; Allen, 1982; Canale, 1985; etc). However, test format has remained fairly static until quite recently. New test item types have been introduced from time to time such as 'cloze' (Oller, 1973, 1975) and the 'C-test' (Klein-Braley, 1983) but the traditional formats such as multiple-choice, essay and dictation have continued to attract the greatest number of adherents. The Test of English for Foreign Learners (TOEFL) for example, taken by over 500,000 candidates in 1987, was exclusively multiple-choice until very recently (1986) when an essay component was introduced. Most tests and examinations coming out of the United States are very similar to TOEFL in format and design - multiple-choice and norm-referenced.

In the late seventies, as the communicative approach to language teaching gathered momentum it inevitably had an influence on test design, primarily in the United Kingdom in the first instance. Language courses were overtly designed with the purposes of the user in mind and various taxonomies of functional situations appeared, the most well known arising out of the work of the Council of Europe in the early and mid-seventies (Van Ek, 1975; Van Ek & Alexander, 1977). Munby (1978) also exerted a considerable influence on the design of specific purposes courses. With the change in emphasis in course design, many teachers began to feel uncomfortable with the apparent mismatch between new approaches to course and materials design, and more traditional approaches to testing.

One of the first major moves away from a more traditional approach to language testing came with the development of the English Language Testing Service (ELTS) examination by the British Council (Carroll, 1980). ELTS replaced the Davies Test as the main diagnostic instrument used to assess the competence of foreign students entering Britain to attend post-graduate courses for the most part. Modifying the approach laid down by Munby (1978) the ELTS design team

produced a modular examination focusing on skills and text-types considered to be relevant to students of different academic disciplines. The examination comprised four main components (listening, reading, writing and speaking) with the results being reported in a profile format. ELTS has been criticized for its lack of empirical research in isolating the various skills and domains of use appearing in the test. A more thorough piece of work was produced by Weir (1984) with the Test of English for Educational Purposes (TEEP) sponsored and administered by the Associated Examining Board.

When the project described in this thesis began there were no testing instruments at work in the institute other than a placement test, which was a haphazard collection of items brought together from a number of commercially available language tests and test booklets. The courses were based almost exclusively on commercially available textbooks and no records of student achievement were kept, in large part because no testing instruments were available. It became policy within the institute to:

i. increase the relevance of course materials by engaging in a needs analysis approach to the description of the student population and their behaviour in the real-world and using the results as a basis for in-house materials development;

ii. introduce a series of performance-based progress tests that would support the implementation of (i) above, providing a reliable and valid method of evaluating students' progress, the effectiveness of the courses, and identifying students who were having significant difficulties coping with the courses.

In short, an attempt was being made to introduce a more communicatively based approach to the teaching of English through syllabus design, teacher training, and effective testing.

Underlying the communicative approach to language teaching is the recognition that learners need to be able to use language in situations outside the classroom that they are most likely to encounter. It is therefore important that the teaching materials reflect this need and in consequence the testing materials must do so too. Classroom materials are frequently changed

in the light of pedagogic fashion, whereas testing materials tend to be much longer lived. While there may be a conflict between learners' expectations of language teaching materials and those that are current under the influence of the communicative approach, there is also a conflict between testing materials appropriate to the communicative approach and those that are familiar to the teacher. Both sets of expectations are based primarily on previous experience.

The battery of tests discussed in this thesis were written in conjunction with a major reorganization of the teaching materials in the British Council Institute in Hong Kong. They were designed to support the move towards a more relevant approach to the teaching of English where the needs and characteristics of the learners were taken as a fundamental feature of the production of new course outlines and materials.

It is frequently the case that testing instruments are added almost as an afterthought in the process of course design. Such was not the case with this project. From the beginning the tests were seen as a powerful force in the implementation of change. However, it

should also be noted that while there was a major shift in focus in the approach adopted to the selection of teaching aims and teacher designed materials, a series of commercially available textbooks (Strategies 1-3) were issued to the students. These textbooks formed the course as far as the students were concerned even though it was recognized by the design team and some of the teachers that they were only partly relevant. Teachers were encouraged to try and develop ideas and materials, linked to the course books that had greater relevance to the students. This involved a degree of investigation on the part of the teacher and an acceptance that some form of negotiation with students would be necessary. A series of descriptive statements of behaviour were drawn up to help teachers and students towards an understanding of the need for greater specificity of materials and approach (see Appendix 5). Testing materials drew on areas covered in the course outlines (see Appendix 6) and on areas outlined in the descriptions of behaviour. The testing programme was intended to help influence a change in teaching approach. The behavioural descriptions and tests will be discussed at greater length below.

3. About the British Council Institute in Hong Kong

The British Council language institute in Hong Kong is the largest of its kind in the world. There are between 9,000 and 12,000 students registered in any one term. In the region of 80% of the students are registered in what are loosely called General English courses with the remaining 20% following specific examination oriented and business skills oriented courses. The seven tests discussed in this document form part of a larger battery of twenty five tests. They were designed for a subset of the General English courses. These courses are divided into four basic levels, each lasting for three twelve week terms. The levels are called A, B, C, D and subdivided into terms 1, 2 and 3, where Level A1 is for the lowest proficiency and Level D3 for the highest. The tests under discussion here are those for Levels A3, B1, B2, B3, C1, C2 and C3. These tests was selected because they were taken by the largest proportion of the General English students. It was established through a simple survey, which involved asking students what grade they had achieved in the School Certificate and then comparing it to the placement test score, that the average A3 student would have failed the Hong Kong School Certificate in English Examination (the locally accepted standard of English

proficiency), while the average C3 student would probably have achieved a grade D pass. In school terms, the range was roughly equivalent to Form III - Form V.

As mentioned above each course used a commercially available textbook. It was felt however that the textbooks lacked relevance and that to supplement them teachers should have detailed information as to the precise nature of the student body in addition to detailed descriptions of the type of language behaviour that the students were likely to engage in in order to develop materials themselves that would be more appropriate to the students they were teaching. Through the use of a questionnaire (see Appendix 7), general characteristics of the student body were established. These characteristics are summarized below:

3.1. Age

Most of the students fell in the age range 19 - 26.

Table 4.1.

Below 18	14%
18 - 22	26%
23 - 33	46%
34 and above	15%

3.2. Sex

There was roughly an equal number of men and women.

3.3. Educational Background

In Hong Kong secondary education only became compulsory in 1978. Officially the medium of instruction in 90% of secondary schools is English with only 10% of the schools claiming to be Chinese medium. In theory this should mean that the overall standard of English is high, particularly as all Hong Kong School Certificate[1] examination papers, with the exception of Chinese Literature etc., are written and supposedly answered in English. Unfortunately the pressure for English medium instruction is social for the most part rather than

educational - English is seen to be vital to upward social mobility. With the rapid expansion of secondary education, standards are inevitably seen as having fallen and a system which worked well when education was elitist is now groaning under the pressure of an unrealistic focus on English. This situation has created a great demand for additional English language tuition satisfied to some extent by the British Council. Listed below is a breakdown of the educational background of the student body as a whole. A more detailed breakdown by level is available in Appendix 7.

Table 4.2.

Primary	11%
Form III	25%
Form V	48%
Matriculation	9%
Graduate	4%
Post-graduate	3%

Primary education lasts for six years from the age of six. Secondary education is compulsory to Form III when there is an examination to select the most able 70% who are funded till Form V. Any other students wishing to continue till Form V have to pay for themselves. It is

not surprising therefore that 36% of the students are educated to Form III standard or less.

3.4. Occupation

Fourteen categories of occupation are listed on the registration form. The distribution is illustrated below:

Table 4.3.

Clerical Workers	26%
Students	20%
Trade and Technical Workers	10%
Factory and Construction Workers	6%
Housewives	8%
Teachers	4%
Salesmen/ladies	2%
Shop Assistants	3%
Hotel/Restaurant Workers	3%
Policemen/women	1%
Medical Workers	2%
Government Employees	4%
Management/Finance	3%
Others	10%

(N.B. In Hong Kong the term 'student' most frequently refers to young people at secondary school.)

Most of the working students fall into two main categories: those engaged in clerical/office work of some sort and those working manually, either skilled or unskilled. Some of them already use and need English in their work, but if this is the case then it is likely that they can more or less cope with the demands placed on them. However, all of them believe that they will need more English in the future since better jobs generally require a higher standard of English. Thus many students are anticipating a need. They come to the British Council in the hope that they will improve their English and so improve their promotion prospects.

4. The Approach Adopted to Needs Analysis and Course Design

The main focus of the approach adopted to course design was to take into account the relevant behaviour of students outside the classroom in the specification of learning outcomes. This behaviour was described in a series of real-world performance referred to as

objectives. The objectives were not of the conventional type such as those illustrated in works by Mager (1962), Valette and Disick (1972) or Jarvis and Adams (1979). In language teaching, a major weakness of many courses, both structurally based and functionally based courses, is that too little emphasis is placed on exactly what the real-world relevance of micro teaching objectives is. It seems frequently to be the case that outcomes are established on the basis of the syllabus which is often derived from a linguistic analysis, be it structural or functional, and not to any great extent on the actual reasons why students need to use the language. The end results of a course of study are thus often inaccessible to teacher and student alike. Learning outcomes need to be expressed in language that is readily comprehensible, illustrating features of language use that are familiar to learners, teachers and testers. Statements of learning goals phrased like the one below are so vague that they are virtually meaningless for all concerned:

> "Students will develop the ability to communicate orally in the language." (Jarvis and Adams, 1979 p.13)

A major aim of the specification of learning outcomes in the institute's reorganization of the curriculum was to make these outcomes readily comprehensible to all

concerned, and to bring the real world into the language teaching classroom.

In order to gather the data required to specify the type of language use that the students were likely to engage in, the following strategies were used:

i. questionnaires;
ii. interviews with students and employers;
iii. teachers comments;
iv. previous experience of needs analysis for specific purposes courses in Hong Kong.

From these sources were established the age, sex, nationality, educational background, occupations previous language learning experience, domains of use and students' stated purposes for learning English.

Three main reasons for the study of English were apparent. These were:

i. to communicate in English effectively at work;
ii. to study through the medium of English;
iii. to communicate in English effectively for social purposes.

Specific groups of students' needs related to one or more of the three reasons outlined above.

Those needing to communicate effectively at work fell into two main categories. They were predominantly office workers between the ages of 18 and 30 holding clerical posts where they needed to communicate in English with foreign colleagues or more commonly foreign superiors or customers. The foreigners came from a range of first language backgrounds, but were predominantly English and Japanese speaking. The second smaller category consisted of manual workers between the ages of 18 and 30 holding skilled or semi-skilled posts which involve contact with foreign customers for information, service and maintenance.

Those needing to study through the medium of English also fell into two main groups. The first aspired to professional improvement through the medium of English.

Hong Kong is a bilingual community but most of those who wish to gain professional qualifications have to take examinations in English, and to study through the medium of English. About 85% of the students registered with the institute were educated to Form V level (17 - 18 years of age) or below, and many of them continued with part time studies, normally of a professional nature, in order to improve their employment prospects. They frequently lacked the skills to study effectively by themselves. The second group comprised full time students - still at secondary school and constitute 20% of the institute's student body.

There was not really a clearly defined group of individuals needing to improve their English for social purposes. They came from a variety of backgrounds. The perceived need to use English socially was perhaps based on some form of integrative motivation. It was not uncommon for those needing to improve their English for work or study purposes also wanting to use English for travel, leisure reading, films or simply to get to know and mix with foreigners.

Learning outcomes were specified in terms of real-world performance intended to motivate the creation of

teaching and testing materials alike. They were specified on two levels, one fairly general and the other more specific. They were not intended to be definitive, in the sense that it was understood that the contexts specified would and could not be of equal relevance to all students. Their aim was to provide meaningful examples of possible behavioural goals to both teachers and students.

Below is an illustration of two general statements aimed at office workers:

1.

1.1 Describe how things work in the office or related work areas.

- the function and purpose of equipment e.g.
    - typewriters;
    - photocopiers.
- the function and purpose of procedures e.g.
    - filing systems;
    - forms and documentation;
    - regulations.

2.

2.1 Describe how to do things in the office or related work areas.

- how to operate equipment e.g.
  - photocopiers;
  - typewriters.
- how to follow procedures e.g.
  - operating filing systems.

Such descriptions are fairly bare. While they have some meaning it is difficult to imagine quite how they can be of great assistance in the language classroom. In order to make them more meaningful example situations were cited to illustrate real-world behaviour. These examples were not intended as a constraint to teachers or test designers. They were intended to increase the amount of freedom available by broadening their horizons. Listed below are five examples:

1.1.1. A clerical worker explaining to a superior officer the relative merits of two typewriters in order to demonstrate why a particular model is preferable.

1.1.2. A bank clerk explaining to a customer or client the difference between a savings account and a current account.

1.1.3. A secretary explaining to a newly arrived senior member of staff the function and purpose of filing systems/office regulations.

1.1.4. A clerical worker writing a response to a letter of enquiry, explaining the difference between two products apparently very similar.

1.1.5. A clerical worker, with responsibility for the filing system, writing an internal memo that includes an explanation of one aspect of the filing system in order to ensure that letters initiated within the office are correctly filed.

Each description attempts to capture an aspect of reality which may constitute in effect, a realistic goal and context for language learning. However, it will be noted that the learner is essentially in the role of initiator in the examples cited above. In fact, each set of descriptive behaviour also puts the learner in a more passive, receptive role. This does not mean, however, that the learner will not have to contribute to the activity. Below is a parallel example to 1.1 cited above:

1.

1.2 Read actively and/or listen actively to descriptions of how things work in the office or in related work areas:

- the function and purpose of equipment e.g.
    - a new collating machine;
    - a multi-purpose photocopier.
- the function and purpose of procedures e.g.
    - filing systems;
    - forms and documentation.

The nature of the interaction is different from the examples cited earlier because the role of the learner changes from 'knower' to 'non-knower'. Two examples are given below:

1.2.1. An expatriate member of staff describing features of a piece of equipment which is unfamiliar to the worker so that the worker can get brochures and estimates.

1.2.2. An expatriate member of staff explaining reasons why a change in the system is needed, giving the principles and criteria underlying the new system and describing envisaged problems so that the clerical worker understands why the new system is being implemented.

(For a complete set of these descriptions of behaviour see Appendix 5.)

In some ways, the approach adopted to the analysis of language needs and specification of objectives was a fairly standard ESP approach. What differentiated it, however, is the fact that the course for which the materials were intended was not an ESP course. It was a General English course. As was clear from the profile of the student body, there were considerable variations in specific features of background and needs.

General English courses pose a serious problem to the course designer, materials writer and testing specialist in that everything usually has to be so bland that it is of little interest or significance to teachers or students. In a commercial environment, where the consumer evaluates the worth of a course very carefully, and if dissatisfied simply does not return, relevance and interest are of primary importance.

5. The Placement Procedure

Prior to entry to any course offered by the institute students underwent a placement procedure. This procedure was unavoidably fairly complex and needed to be as thorough as possible in the time allowed. Due to the large numbers of students involved in the placement procedure, up to 5,000 in any one day, it was imperative that a student could be tested and placed in two hours or less. The Placement Procedure (see Appendix 4) was in four stages which are summarized below.

5.1. Stage 1

The students take a 40 minute objectively marked multiple-choice test. This test comprises a series of rational deletion multiple-choice cloze passages. The text types are varied, (a letter, an article, a short story) and graded in difficulty. The reliability of the test, as measured by the Kuder-Richardson 20 formula, was 0.94. The test is particularly sensitive to the lower ability range and discriminates well between students in Levels A, B and C. It is timed to take the

full forty minutes for these students. Because students at a higher level can finish the test in under forty minutes, an additional written element is included. Students are asked to write briefly on two topics:

i. where they use English most;
ii. why they want to improve their English.

It was intended that these pieces of extended writing should provide helpful information for placement purposes later in the process when the student was meeting with a counsellor (a full-time teacher at the British Council).

5.2. Stage 2

The students complete a self-assessment based on six simple band descriptors roughly matching those used by the oral assessors (see Appendix 4). It was felt that the views of the students regarding their own level of proficiency were of importance for two reasons. Firstly, the student is the one with most first hand experience of his own level of proficiency yet his opinions are rarely sought or taken into consideration. In order to help humanize the placement testing

situation, as high a degree of student involvement as possible was desirable. Secondly, although the student's assessment was subjective, it provided a further piece of information to help make the placement as accurate as possible.

5.3. Stage 3

This stage involves the students in an oral interview situation. Each student meets with an assessor for five minutes on average. The oral assessment, while carefully structured, attempts to allow for a communicative interaction of some sort to take place between student and assessor. In the first part of the assessment, the assessor attempts to put the student at ease by adopting a friendly, welcoming attitude - shaking hands, introducing himself and so on.

The student is then asked to read aloud. In most oral tests reading aloud usually means reading a dialogue or passage of some sort. In general these are not the sorts of texts that people read aloud in any spontaneous situation. In order to overcome this drawback, most oral tests (Hong Kong School

Certificate, Cambridge First Certificate) allow students a few minutes to prepare their reading of the text in question. This type of preparation time was quite impractical in the situation being described here. In addition grave doubts were felt as to the validity of the exercise. However, reading aloud does provide a fully controlled example of the students' ease with the language. It also provides the assessor with an opportunity to listen without having to think of something to say, and by occurring at the beginning of the interview helps to dissipate some of the nervousness that the student will inevitably feel. It is interesting to note, that there was a visible expression of relief on the part of most students when they finished reading. They perceived this as the most daunting part of the interview.

As mentioned above, grave doubts were felt regarding the validity of the traditionally used text type in the reading aloud part of the test. Consequently, it was decided to use texts that were likely to be read aloud by native speakers without any preparation. Such texts include short, interesting newspaper stories, instructions for a game and either recipes or instructions about how to cook something. In this way it was felt that the advantages of a reading aloud task

in the oral test could be retained while at the same time there would be an emphasis on the face validity of the task.

The reading aloud formed part of the overall assessment of oral proficiency although it was not marked separately. Assessors were advised that reading aloud would probably seem weak compared to free speech and that this fact should be borne in mind. They were instructed to consider features such as:

accuracy of phoneme production;
intonation pattern;
amount of apparent comprehension;
ability to handle unfamiliar words.

The final part of the oral assessment attempted to create a situation where the student could reasonably believe that the assessor did not already know the answers to the questions that he was asking. One of the main problems in any oral assessment is likely to be the fact that there is little motivation for real communication to take place. With a picture description for example, such as that used in several well known

language tests (Cambridge First Certificate, Ilyin Oral Interview), the assessor already knows the answer to any question he asks and the student knows that he knows. The type of interaction is therefore quite untypical of most real life situations on one hand and somewhat demotivating for both assessor and student on the other.

An attempt is made in the final part of the oral interview to create situations that are structured, in order to allow for comparability of communicative ability, yet flexible enough to create the impression, at least, of a real communicative event. Two basic strategies were adopted to satisfy these requirements. The first strategy involved the use of a map of Hong Kong with several famous beaches marked as a stimulus. The assessor claimed to be a new arrival in Hong Kong and sought advice from the student as to the most suitable location to take some visitors. The student would often feel the need to make further enquiries of the assessor in order to provide the most appropriate advice. The second strategy involved the use of a map of China with several important cultural/tourist locations marked. In this case, the assessor asked the student for advice as to where to visit and so on. These types of situations are fairly representative of

the type of communication likely to occur between a Hong Kong Chinese resident and newly arrived expatriate. With a reasonable degree of commitment on the part of the assessor, it is quite possible to stimulate a fairly authentic and meaningful communicative event despite the inherent constraints of the testing situation.

Assessment was made on a six point scale based on band descriptors written after careful observation of videos of students at different levels of competence (see Appendix 4). The assessors were asked not to look at the self assessment or the written test score when coming to a decision as to the oral ability of the students that they spoke with. Every attempt was made to ensure that the oral assessment was not influenced by factors other than oral ability.

Due to the enormous number of students being tested, it was not possible to use language teachers as assessors for the most part. As a result a group of eighty individuals were recruited from the community. The main qualifications for the assessors were that they should have native speaker competence in English and that they should be intelligent, sensitive and competent enough

to carry out the task successfully. The assessors, both male and female, came from a range of backgrounds from professionals involved in the business community to housewives. They were given an initial full day training, sensitization and standardization session which was supplemented by a 3 hour refresher session prior to each placement testing period three times a year. Performance was monitored regularly and suspect assessors were either retrained or dismissed. In this way, standards of oral assessment remained fairly high.

5.4. Stage 4

The final stage of the assessment was referred to as the 'counselling session'. At this point the student would meet with a full time teacher of the institute. Each student would arrive with three to four pieces of information regarding his level of proficiency (i.e. the written test score, the self assessment, the oral assessment, and possibly the short piece of extended writing) as well as a registration form that required him to provide certain background information such as occupation, age, and education.

The information regarding level of proficiency includes an objectively marked written test score, a self assessment, an oral assessment and a short piece of free prose (assuming the student's level is high enough) on the subjects mentioned above. It is then up to the counsellor to place the student in the appropriate class to suit both language level and needs. Counsellors spend two minutes on average with each student. Where the levels of assessments match, placement can be carried out fairly quickly. Where there is an apparent discrepancy, the counsellor is obliged to take longer to satisfy himself that the student goes to the right level. The discrepancy could be due to a number of reasons. Some students, for example a taxi drivers, were in a situation where they used English orally fairly often, and might therefore have been able to speak quite well. On the other hand, they had left school after primary or Form III level and so perhaps had problems with grammar, which may have been revealed by the written test. Alternatively, quite a number of students were recent arrivals from mainland China. They may have done fairly well on the written test yet had serious difficulties with oral communication. It was also possible that the oral examiner and student had taken either an instant dislike or liking to each other, thus making the scoring less reliable. In about 60% of the cases final

placement is straight forward - that is to say the three scores match closely. In the other 40% there is enough of a mismatch to warrant closer examination. It was considered essential that the full time teachers of the institute had the final responsibility for placement. They had to feel that the students in the different courses were there because they, the teachers, decided that they should be there and not because some anonymous test or testing specialist had decided.

The whole procedure takes a student about two hours from start to finish. It attempts to assess efficiently and accurately in the shortest possible time. It also tries to combine information on the students' communicative ability, as measured by the oral interview and self assessment, with language accuracy as measured by the objectively marked multiple-choice cloze. The communicative emphasis, while time consuming, is proven necessary by the 40% discrepancy mentioned above and by the fact that the emphasis in the courses is primarily communicative as opposed to structurally oriented.

6. Considerations in the Design of the Progress Testing Instruments

After completing the placement procedure students register for the course best suited to their needs. The course syllabuses were a compromise solution in that while a basic needs analysis and specification of performance descriptions/objectives had been carried out there was nevertheless a series of standard textbooks with a fairly detailed course outline based on both of these sources (see Appendix 6). An attempt was being made on the part of the institute to modify the approach to language teaching internally: to move away from a total dependence on commercially available textbooks designed primarily for the European youth market towards a more sensitive integration of published teaching materials and the specific needs of the students in Hong Kong.

Literature on the problems of innovation in education abounds (Clarke, 1987; House, 1974; Pratt, 1980; Hurst, 1983). In order to achieve any degree of success in this project it was seen as necessary to implement change gradually and with the support and understanding of the teachers and students.

Course books continued to be used for two main reasons. Firstly, they were perceived as important by the students who felt a need for the support that a course book provides. Education in Hong Kong follows a traditional path with most learners absolutely convinced that successful study means learning the textbook from cover to cover. In a sense they are justified to some extent in their views since all public examinations are conducted in English. Proficiency in that language at secondary level is not always very high so that one way of achieving a degree of success is to learn model answers by heart and simply parrot them in the examination. This strategy is not at odds with traditional Chinese attitudes to education. Indeed, learning to read and write Chinese involves rote learning on a massive scale. Students would find the idea of studying without a textbook quite unnerving. Secondly, many teachers also felt more comfortable with a book to use. They were happy to supplement it but not very keen to replace it entirely with either in-house materials or simply their own work, partly through lack of self confidence, and partly because of the considerable amount of extra work involved. Supplementary materials were used and they came either from commercially available sources, or

from materials teachers designed themselves, both housed in the more than adequate resource centre at the institute.

While broad guidelines were laid down, in the form of objectives and a course outline, teachers were not obliged to adhere strictly to any fixed pattern as long as the important features of the course were covered. In addition it was expected that certain teachers would wish to experiment extensively with a more ESP oriented approach and enter into some form of negotiation with their students in order to achieve the most suitable mix of commercially available and tailored materials for their particular classes.

Since the courses were fairly dynamic and since the students came from a range of backgrounds, the problem of writing tests was a difficult one. It is all too often the case in educational settings that the test constructor is involved only at the very end of the course design and materials writing process, if at all. He is supposed to work from tight specifications drawn up after careful analysis of the syllabus and objectives. In a true ESP context this may be possible, in most General English contexts it is not. The result

is that in the majority of situations where General English courses are concerned, there are either no tests at all, ones that do not reflect what has been taught, or completely unreliable instruments that do more harm than good in the educational context.

In this project the test designer was involved in the initial needs analysis, and the specification of objectives from the start. The tests were to serve as one of the means by which teaching and learning effectiveness was to be measured. In addition they were seen as a means by which the process of curriculum reform could be accelerated. It is a well recognized fact that examinations and tests have a major influence on what happens in the classroom. One would expect that this influence would be greatest in a conventional school setting where pupils are preparing for public examinations. However, it is also a powerful force in the less formal setting of a British Council institute. As soon as the students know that they will be sitting a formal test they want to know what will be tested and how they should prepare for it. They exert a degree of pressure on the teacher, who whether he likes it or not, has to take the examination into consideration when preparing teaching materials. The test designer needs to be fully aware of the consequences of his

actions. To this end the tests he produces should at least:

i. reflect meaningful activities;
ii. sample effectively from the domains of use most pertinent to the students;
iii. measure performance in a reliable and consistent manner.

As long as they satisfy these conditions, they can be introduced with the confidence that they should enhance the effectiveness of any given course. It was quickly realized that the tests discussed in this thesis would play a major role in the changing syllabus and approach to the teaching of English in the institute. It was imperative therefore that the influence they exerted be a positive one.

7. What to Test and How to Test it

In considering what to test and how to test it the concept of construct validity, the extent to which any test reflects a principled view of language proficiency, is an important one. The view of language implicit in the development of the courses and tests in

the institute, and in the commercially available teaching materials that were used owed much, in the first instance, to the functional/notional approach (Wilkins, 1976; Van Ek, 1975; Van Ek & Alexander, 1977), which greatly influenced all language teaching in Britain in the seventies and eighties. In addition, the view that language functioned at the level of discourse (Sinclair & Coultard, 1975) as opposed to simply at sentence level was also an important influence. Furthermore, it was decided that the tests would be based on the premise that communicative competence was divisible rather than unitary (Oller and Hinofotis, 1980; Canale and Swain, 1980). In consequence, the tests were divided into four basic parts, in an attempt to isolate and test several distinct areas. The final division owed most to the model of communicative competence presented by Canale and Swain (1980). The four basic parts were as follows:

i. Listening (Discourse);
ii. Grammar (Grammatical);
iii. Appropriateness (Sociolinguistic);
iv. Reading and Writing (Discourse).

The information transfer principle (Johnson, 1982) which states that an important characteristic of

communicative language teaching in that it focuses attention on the ability to understand and convey information content was important in the formulation of many of the test items. Although Johnson primarily intended it for the production of teaching materials this principle is equally applicable in the testing context. A task can be said to be communicative to the extent that the student is being asked:

> "... not to comment on any point of grammatical structure or lexical meaning, but to extract certain pieces of information and to transfer them ...' (ibid, p. 164).

Morrow (1979) also laid down a series of features that describe communicative events and that he recommended should be taken into consideration when preparing test items. He argues that items should be performance-based, authentic, purposeful and interaction-based. Such considerations were taken into account where possible in the writing of items.

However, in an institutional context, the appearance of a test is influenced by a number of factors. While it is important for the test designer to bear in mind that he must hold a theoretical position, it is equally important for him to remain aware that tests exist within an educational context. They are real and

meaningful activities that matter to the students who take them. Practical testing situations, while well suited for research or innovative approaches, are also under the constraint of accountability. The well-being of the students and the good will of the teachers must never be sacrificed at the altar of research or innovation.

7.1. If the Test Comes at the End of a Course of Study, What has been covered in the Course?

A test should reflect course content. This is not to say that each item in the course needs to be tested. Unfortunately, in the minds of many teachers and students a test needs to cover all aspects of a course to be valid or fair. If the test is a discrete-point grammar test, testing a discrete-point grammar course then this may be possible if not desirable (Carroll, 1962). In almost any other context it is simply not possible to test all that has been taught in the time available for testing. The following points provided guiding principles in deciding what to test:

i. A representative sample of areas covered in the course need to appear in the test.

ii. Enough variety needs to be present to satisfy teachers and students that no one is being discriminated against or favoured in any way.

iii. The item types that appear in a test must be familiar to both teachers and students.

iv. The test content must not appear to be trivial.

v. There must not be an undue emphasis on areas of minor importance.

vi. The use of item formats suited primarily to testing purposes e.g. discrete-point multiple-choice, should be avoided as far as possible if they conflict with sound teaching principles.

It will be obvious from earlier discussion that the nature of the course was flexible. No clear definition of what was actually taught was possible. At best, one was able to deduce what might have been taught.

Therefore, in order to satisfy the condition of familiarity mentioned above it was decided to draw upon the real-world language use of the students as a source for the format of items in conjunction with the teaching materials both in the textbook and for supplementary use.

7.2. What do the Teaching Materials Look Like?

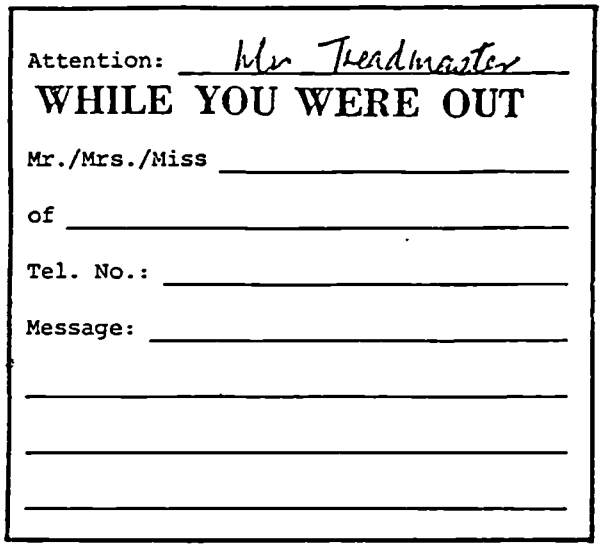
All too often tests do not in any way resemble the teaching materials in style and format. If teaching a language aims to prepare learners for real-world use of that language then it is reasonable to assume that certain tasks encountered in the classroom will, to some extent, reflect reality. Other tasks may be of a purely pedagogical nature. There must, for students and teachers, be either a pedagogical or real-world familiarity with items in a test - preferably both.

7.3. Are the Tests to be Performance/Task-based or Discrete-point-based?

This raises the question of whether the test items should be task/performance-based or discrete-point. As teaching becomes more whole-task-based it is inevitable that test items must follow. However, this causes two sets of problems from a testing point of view. Firstly, how is the tester to sample effectively from all the task-based activities and to what extent are the results obtained generalizable? This problem is discussed at length by Alderson (1981) and Weir (1981) who arrive at no satisfactory solution.

Secondly, in real life, a task is generally either successfully completed or not. In class, the teacher can focus on any aspect of the task in order to improve student performance. In the testing context, however, the task provides only one mark if treated as a unity: as long as an overall criterion for success can be defined. Such a task may take several minutes or longer to complete. If the test in which it resides is to be used for ranking or grading it can be extremely uneconomical to treat a task as a unity . An example of a task based item would be the telephone message form

illustrated below. Clearly, for the task to have been successfully completed all the relevant information needs to be present. Unfortunately this is rarely the case - mistakes are made, information is missing. It would be difficult to score such an item dichotomously and achieve a reasonable distribution of scores.


Attention: Mr Headmaster

WHILE YOU WERE OUT

Mr./Mrs./Miss _____

of _____

Tel. No.: _____

Message: _____

A compromise solution that satisfies the criterion of authentic appearance, allows the tester to allocate an appropriate number of points to the task to make it economical from a scoring point of view, and provides relevant data for validation, is to break a task down into discrete points for marking purposes. It is important the student does not perceive such a task as

a group of individual items but rather as a whole task. This was the approach adopted to the marking of performance-based items in the tests discussed in this thesis.

It is difficult to involve the students in test construction, but it is of great importance that their views are sought after pre-testing or test administration in order that objectionable items can at least be considered again. It is often enough for teachers to ask for informal feedback at the end of a test. With the battery of tests under discussion here, this was the approach adopted to capture the views of the students. Items which received consistently adverse criticisms were moderated again. If necessary they were rewritten or abandoned.

Equally important as the views of the students is that of the teachers. At best the concept of testing in the English Language Teaching context is unpopular and badly understood. For any approach to testing to succeed, therefore, three factors are of vital importance:

i. Teachers must gain some familiarity with the principles and practice of language testing. This is perhaps best achieved through some form of basic training course.

ii. Teachers must be involved in the process of test design, item format selection, and the writing of test items.

iii. Teachers must be familiar with the life cycle of a test and aware of the fact that good test construction cannot be haphazard.

It is unfortunately very difficult to achieve any of the three aims above in a short period of time with an entire teaching body of any size. In the case of the British Council institute in Hong Kong, there were more than one hundred teachers employed at any one time and so, training and involvement had to take place by degree. However, it was anticipated that the credibility of the tests and the process of consultation would be better accepted when those who were actually involved in working on the tests mixed with teachers who were not involved. The more teachers that could be made to feel a personal commitment to the tests, the more people there were who would be

available to explain and defend them as necessary. The image of the test constructor in the ivory tower having no contact with the teaching body had to be dispelled as fully as possible. Thus it was that there were generally between four and six teachers involved in test construction in any one term.

8. The Preparation of the Tests

Items to be included in the tests were selected on the basis of their relevance and familiarity and the extent to which they were, when incorporated into a test, reflective of the course students had followed and the ways in which they put language to use. Ideas for items were generated by the textbooks, supplementary materials and performance descriptions discussed earlier.

The first version of a test to be produced (the A3 Progress Test) was initially written by the testing specialist alone in order to trial the item types and approach. Subsequent versions of this test, and the other tests were a joint effort between the testing specialist and groups of teachers in the institute.

The A3 Progress test, like all the others, is divided into four basic parts. As mentioned earlier, the A3 level students are the least competent in their command of English therefore the test tasks that they have to perform are of the most basic kind. Every attempt was made, however to keep these tasks realistic and relevant.

8.1. **The Listening Test** comprises three item types. The first simulates a typical telephone situation that the students are likely to encounter, the second a face to face exchange at a hotel reception desk, and the third a face to face exchange between a travel agency clerk and a tourist booking a day tour.

Taking telephone messages

This involves:

- writing down spelling of names;
- writing down telephone numbers;
- writing down short messages (instructions, places, times);

Writing down information about a customer

This involves:

- writing down spelling of last name;
- writing down first name when not spelt;
- writing down 'Tokyo' (not spelt);
- writing down spelling of address;
- writing down name of local airline (not spelt);

Writing down information for customers at a travel desk

This involves:

- writing down spelling of name;
- writing down room number;
- writing down number of people going on trip;
- writing down times of day;
- writing down price.

In the real world, skills frequently tend to integrate. This feature of language use was accepted as fundamental to item design. However, it should be noted that reading and writing are kept to a minimum in the Listening test. It was felt that it would be unfair to include a significant element of either of these two skills, since the students' competence in both was likely to affect performance in listening. Enough reading and writing was retained to ensure the reality of the tasks while not hindering students in their completion of these tasks. The tape recordings were

made in studio conditions and various sound effects incorporated to make them more realistic.

8.2. **The Grammar Test** caused some concern. It was decided that the tests should include a section on grammar, or perhaps more appropriately, accuracy. The communicative approach has been much criticized by teachers and students for its perceived lack of concern for the formal features of language. In the Hong Kong context, it was very important to the students that there should be something called grammar in the test. From the theoretical point of view, it was also felt that emphasis should be placed on more formal features of language. How they should be tested was the difficult question. If standard discrete-point multiple-choice items were used, the washback effect on the classroom would have been negative in the sense that the multiple-choice approach to grammar teaching was not a feature of the teaching method in the British Council. It was also thought better to use an item type which was text-based as opposed to sentence-based. To this end a variation on the cloze procedure was developed for use in the A-level progress tests. It was given the name 'banked cloze' because above each text there was a bank of words, normally two or three more than there were spaces in the text. Students chose a

word from the bank to match one of the spaces. Each text was based on some authentic text-type relevant to and within the experience of the students. These are listed below:

An article from Student News.
A newspaper article.
A description of an office layout.
A letter to a friend.

It should be pointed out that the same format was not used at higher levels. A method of rational deletion (Alderson, 1983) was used instead. It was accepted that there were many potential hazards in the use of the cloze. However, it satisfied the washback requirements better than any other item-type that the writer was familiar with at the time.

8.3. <u>The Appropriacy Test</u> was based on the now common teaching technique, the half-and-half dialogue. Situations relevant to and within the experience of the students were selected. One person's part of the dialogue was left blank and it was up to the student to complete it as best he could. Clearly, writing down what would be said in a conversational context suffers

from the point of view that it is not very realistic. However, it was a teaching device commonly used in the institute, and thus familiar to the students. Furthermore, it focused attention on the sociolinguistic aspects of language and allowed for a degree of controlled creativity on the part of the student. The marking was carried out on two levels. If the response was inappropriate it received no marks, regardless of accuracy. If it was appropriate, then the marks were scaled according to accuracy. Only a response that was both appropriate and wholly accurate could receive full marks.

The types of functional responses that the students were expected to make are listed below:

- giving directions.
- asking about well being;
- offering a drink;
- asking for preference.
- asking about type of work/job;
- asking about starting time;
- asking about finishing time;
- giving information about own job;
- giving information about week-end activities.

8.4. **Reading and Writing** were the final two skills areas in this test. An attempt was made here to integrate the activity as much as possible, and to base the task on realistic texts. Students were asked to fill in a visa application form using a letter and passport as sources of information. The passport was authentic reading material, while the letter was specially written for the test. The form was a slightly modified version of a real visa application form. The introduction of authentic materials into the test as opposed to contrived teaching materials, and a focus on a situation that any of the students may need to deal with was an important statement. The test was attempting to do something that most of the teachers were not, that is, using authentic materials with low proficiency students. The teachers soon saw that the nature of the task was as important as the material. They were also able to see that students almost enjoyed this sort of activity, and immediately understood its relevance to their day-to-day lives. Informal feedback from teachers, after the introduction of the test, indicated that it had encouraged a greater focus on the use of authentic materials and realistic tasks in the classroom. Thus, one of the objectives of the testing programme, which was to change the focus of the teaching, was being achieved.

9. The Six Stages of Test Preparation

Little guidance has appeared on how to actually develop a communicative test battery or integrate it into the workings of a school environment. Carroll (1978; 1980) gives the matter of test development some coverage but he does not consider, in any depth, the consequences or role of testing in an educational context. With regard to involving teachers and integrating testing into the school environment, there is also very little guidance available. Alderson and Walters (1981) discuss the question of training teachers in testing techniques on a postgraduate course. The process of training and sensitization in-service is not considered.

Inextricably linked to the process of test development, as described in this document, was the need to actively involve and train teachers in the institute in test design and implementation. Each test underwent very similar treatment before it was finally implemented. It was through involving teachers in the stages of this treatment that some degree of training and

sensitization was achieved. Listed below are the six stages of test preparation.

9.1. Stage 1

At the beginning of term, the testing specialist met with a group of teachers specializing in writing items for a given level. In this case 'specializing' means teachers who had worked with students at that level and were preferably teaching them in that term. The purpose of the meeting was to discuss any ideas that the teachers may have, to take into account any feedback regarding the tests already operating and decide on a topic area that each teacher could focus on in order to prepare an item for the next meeting. Teachers were briefed on some of the difficulties they were likely to encounter and how they might cope with them.

9.2. Stage 2

The teachers write first draft items based on the discussion in the first meeting, their experience of

the materials and students, the course outlines and performance objectives .

9.3. Stage 3

A series of meetings is held when the items prepared by individual teachers are subjected to group moderation. The items are discussed in terms of their relevance, testing points, importance, and suitability for the students in question. It is important that any idiosyncrasies are removed at this stage.

Group moderation is a vital phase in the preparation of items for several reasons. Firstly, in test construction, where great precision and clarity are required, several people working on an item inevitably produce better results than just one person working alone. Secondly, a group product is generally better balanced and more widely applicable if worked on by teachers all actively engaged in teaching the course. Thirdly, the teachers in the test construction team are well prepared for many of the questions that might later arise from the use of a particular item and are able to justify its inclusion in a test.

Teachers are often found to rush moderation at first because they may be worried about offending their colleagues or unable to focus precisely enough on the likely problems or difficulties an item may pose, such as markability, reasonable restriction of possible answers and so forth. It is important to insist on thorough moderation at this stage since without it the product will probably be of inferior quality and may need complete re-writing and pretesting before it is of any use.

9.4. Stage 4

Completed items are then informally trialled with participating teachers' classes in order to uncover any glaring difficulties that the moderation team had not been able to predict. This helps to greatly increase the sensitivity of teachers engaged in item writing. It is all too commonly believed by teachers and administrators alike that test construction can be accomplished quickly and that the product will still be quite acceptable. Unfortunately, due to a number of factors such as the unpredictability of the students,

the short-sightedness of the test writer, the lack of clarity in instructions, this is rarely the case. Initial moderation helps to make teachers aware of some of the difficulties; trialling informally with their own classes is an invaluable addition to this sensitization process. Moreover, teachers have the opportunity of observing the reactions of students to the items and the way in which they attempt to do them. Both of these factors are of great consequence in the construction of task-based tests that attempt to have a positive washback effect on the classroom.

In the British Council institute in Hong Kong, the average term lasts twelve weeks. It was found that stages 1-4 above would take a whole term.

9.5. Stage 5

After initial trialling, the moderation team meets again, and in light of the experience gained so far prepares a pre-test version of a test or part of a test. The pre-test is then administered to a representative sample of the population and the results analyzed. It is generally necessary to pre-test twice

as many items as will eventually be required to achieve the appropriate quality.

9.6. Stage 6

The moderation team meets to discuss the results of the pretest and decide on the final form of the test items.

Any test item generally takes six months from inception to completion in the context under discussion here. Teachers are involved in the process from start to finish. Those teachers involved realize that the process of test construction, while lengthy and time consuming, is carried out with the greatest of care because the test results have a very real influence on the students in question. They are able to bear witness to the fact that no test is produced without due care and attention. To begin with most of them believe the approach to be unnecessarily long drawn out and tedious, but as they work on items and become fully aware of the fallibility of tests and test constructors, their attitudes change.

10. The Implementation of Tests

On completion of all the stages outlined above, the final versions of tests were produced. This does not mean to say that once produced the tests could not be changed. Further analysis was generally carried out on the tests when in use and items that were thought to be suspect at this stage were excluded from subsequent administrations. In addition, teachers comments were sought and taken into consideration in later test versions if relevant. The final versions presented in this thesis could be further improved and will be. However, they are all of acceptable quality.

Familiarization and standardization sessions were conducted before new tests were administered for the first time. This ensured that any inadequacies in the instructions to teachers and test keys were likely to be uncovered in advance.

Each test was administered in a single one and a half hour lesson two weeks before the end of the term. The listening test was always administered first, and generally lasted about twenty minutes. All answers were

written in the test booklet. At the end of the test, the scripts were collected by the teachers and marked within five working days. All scripts were returned to the testing specialist. Random samples were analyzed again, and spot checks were made on standards of marking. It was also anticipated that teachers would use information gained from marking the tests to embark on remedial teaching action as necessary in the last week of term. It was considered important that the test could play a formative role in addition to their obvious summative function.

One of the main uses to which the test results were put, was the identification of the weakest students in a level so that appropriate remedial action could be taken regarding these students. A cut-off point was assigned to each test, and teachers were advised that students falling below this cut-off point should be looked at carefully. The cut-off point was decided not on the basis of success or failure in any particular task or group of tasks since this type of decision was beyond the scope of the project. Instead, a very simple method was used. The mean and standard deviation on the whole test for the population was calculated, and any student with a score of more than one standard deviation below the mean was considered below the cut-

off. Statistically this accounted for the bottom sixteen percent of the population. Equally, students gaining a score of more than one standard deviation above the mean, that is the top sixteen percent, were considered as possible candidates for rapid promotion. Teachers were not informed as to how these cut-off scores were arrived at, unless they specifically asked, since although the methodology is extremely simple, it requires some basic understanding of statistics which most of the teachers did not have. The test scores were weighted informally against teacher assessments, outlined below, before any final decision about a student's academic future was made.

11. Teacher Assessments

In addition to marking the tests, teachers were also required to give a subjective assessment of each student. These subjective assessments focused on speaking, writing in class, progress and effort. It will have been noted that there was no oral element in the tests outlined above. In the institutional context, where there is a lot of talking in the classroom on a regular basis, it was thought better to grade oral

competence separately. Furthermore, formal oral tests are generally very time consuming and notoriously unreliable. Although writing was tested in the formal tests, it was decided that it should also be assessed separately by teachers. Each student therefore received a combination of marks at the end of the term. One set was based on the formal test, the other on the teachers' assessments and effectively based on the teachers' perception of each students' competence in the classroom context. The teachers' assessment was also regarded as a possible means of validating the objective test scores at a later date.

Teachers were also asked to make subjective assessments of the students' progress and effort. It was felt that affective evaluation of this kind would add a dimension to the overall assessment of each student. However, in this area difficulties were encountered from some members of staff that invalidated the results to an extent. Certain teachers believed that it was immoral to make assessments of this kind and either refused to do so or awarded each class member the same grade. It had been the intention to use this information for further validation of objective test scores. It was relevant, for example to investigate the extent to which performance matched teachers' expectations - a

question which could be answered to some extent using the information gained by the progress and effort scores. While this question is investigated later, the results are not as reliable as they might have been with the full support of all the teachers.

As with any subjective assessment there was some difficulty in standardizing teachers' perceptions. The oral and written grades mentioned above were marked on a fifteen point scale, while the progress and effort grades were marked on a five point scale. In essence the grading was very similar. Certain grades (7, 8, & 9 in the case of the oral and written assessment, and 3 in the case of the progress and effort) were categorized as 'average'. That is to say, if the teacher felt that a student was coping adequately with the materials he should be awarded one of the average grades. If the student was obviously making a greater effort than his peers or making more apparent progress, then he should be given an above average mark. Teachers were advised that most students were expected to be average, with a minority gaining higher or lower grades. In a sense this was creating a sort of normal distribution curve, the assumption being that most classes would be fairly similar in their characteristics. Unfortunately, this is slightly

counter intuitive, and many teachers had difficulties fully understanding the principle. However, when the mean scores for these four categories were subsequently calculated it was noted that the desired distribution had been achieved. With the written and oral assessments the mean score by level was consistently between 8.00 and 8.50 with a standard deviation between 2.00 and 3.00. Teachers were grading slightly leniently overall but within the average band. With the progress and effort scores the mean was consistently between 3.00 and 4.00. Here teachers obviously felt that their own students were trying harder than average, and making more progress.

12. Conclusion

In this chapter I have provided a comprehensive review of the context within which the seven tests, that are later to be statistically validated, were developed and implemented. I have paid particular attention to two important principles of test design within an institutional context that often appear to be ignored. Firstly, the test designer needs to be fully involved and familiar with the background of the students (including their attitudes to learning and teaching),

course design and pedagogic rationale before creating tests. Secondly, no single person should create tests in an ivory tower situation. It is important, considering the inevitable teacher resistance to testing within almost any educational institution, to make sure that teachers are familiar with the issues and procedures involved so that they can decide on a point of view from a position of knowledge rather than ignorance. Good testing and evaluation procedures tend to be welcomed when the benefits that they can bring are understood. Involving teachers in the specification of test content and the subsequent writing and design of items is therefore essential to the effective introduction of a formal battery of internal tests. Tests must be seen as part of the educational whole as opposed to an annoying and basically unnecessary addition.

[1] The Hong Kong School Certificate (HKSC) is officially linked to the University of London Examination Board GCE for overseas candidates. A grade A, B, and C pass in the HKSC is equivalent to a pass in the London Board GCE O-level for overseas candidates.

Chapter V

1. Introduction

In Chapter IV the background to the construction and design of the tests in this battery was considered in some detail. Chapter V is the first of two chapters focusing on results of data analysis. It discusses matters arising from the item analyses and a variety of correlational analyses carried out on the tests and related data. A detailed discussion of the item analysis results can be found in Appendix 9. Demographic information concerning the test populations is available in Appendix 7. A brief review is conducted of the features of a specially designed suite of computer programmes relevant to the data analysis. (A more detailed review is available in Appendix 8.) This review is followed by information and discussion on the overall statistical features of the tests. Full test statistics are presented in Appendix 2. It was noted that the performance of items when analyzed as part of the whole test was different from their performance when analyzed only as part of their respective subtests. This finding supported Hypothesis Two

relating to the divisibility of communicative competence as measured by these tests.

Each test was subjected to correlational analysis. Firstly, the four main subtests were correlated in order to establish the degree to which it could be said that they were measuring the same trait. While the correlations were always significant, which is to be expected, they were rarely high enough to claim that the amount of shared variance made any of the tests redundant. This further supported **Hypothesis Two**, where it was claimed that different traits, or aspects of communicative competence, as measured by the tests, were distinguishable. Several other variables were also compared, such as placement test scores, teachers' subjective assessments and attendance at courses.

A second correlation study was then carried out. Student performance on each task in the tests was isolated, and the tasks were treated as tests in their own right. There were between twelve and eighteen clearly distinguishable tasks in each test. According to the literature on convergent and discriminant validity in a multitrait-multimethod matrix (Campbell and Fiske, 1959; Stevenson, 1981; Clifford, 1981;

Bachman and Palmer, 1983) subtests measuring the same trait should be more closely related than subtests measuring different traits. When the tasks in this battery were correlated, there was a tendency for tasks testing the same trait to be more highly related than tasks testing different traits. This supported **Hypothesis Two** which claimed that there were different and clearly distinguishable components to language proficiency as measured by these tests. The hypothesis was further supported by the fact that the results were fairly consistent across all seven tests.

2. The Item Analysis Programme

In order to carry out the item analysis a suite of programmes was written in Basic to run on a microcomputer. It permitted a comprehensive and versatile preliminary analysis of the test data and allowed for thorough and principled moderation of the tests. The formulae and procedures employed by the programme were selected after consulting a range of texts on Educational Measurement including Hatch and Farhady (1982), Ebel (1972), Tuckman (1972), Nunnally (1967), Robson, (1979), and Guildford (1982) amongst others.

Very few programmes specifically for test analysis are available. Certain institutions, mostly examination boards, have analysis programmes, but they are not accessible for general use. Many universities and polytechnics utilize commercially available statistical packages such as the Statistical Package for the Social Sciences (SPSS) or the Statistical Analysis System (SAS) but they require careful and extensive programming before they can perform comprehensive item analysis. As a result many researchers use only the most basic functions that these packages offer to obtain overall test statistics as opposed to item specific ones.

An account of the features of the item analysis programme directly relevant to this thesis is given in Appendix 8. For a fuller account of the overall capacity of the programme, which allows for over ten different ways of viewing data, the manual should be consulted (Milanovic, 1988).

3. Method of Administration

Each test was administered in a standard ninety minute lesson by the teachers responsible for teaching the classes. The Listening Test was administered first, and lasted about twenty minutes in most cases. Students then completed the other sections of the test. In earlier administrations a time guide for each of the sections had been included so that students would get some idea of how long they should spend on each section and still complete the test in time. This was later removed following feedback from teachers.

On completion, the tests were collected and marked by the teachers. A carefully prepared 'key' (see Appendix 1) was supplied and training sessions were conducted where necessary. It was considered important that marking standards should be as consistent as possible. Inevitably there was some variation but in general it is safe to say that it was minimal - due in part to the guidelines supplied to the teachers, and in part to the objective nature of much of the marking. Specially designed grids were printed on the test papers (see Appendix 1) onto which teachers marked the score for

each item. These grids allowed for the reliable and rapid transfer of results for computer analysis.

When the marking was completed, all test scripts were returned to the testing specialist. Samples of between 250 and 300 were then selected from each of the batches. Each sample represented between 30% and 60% of the population and was taken at random from each whole batch.

A sample of scripts was checked to ensure that the marking had been properly carried out, and then scripts were keyed into the computer. The keying in was supervised by the testing specialist and checks made to ensure that there was minimal error. It was expected that there would be some error but since it was unlikely to be systematic it was not anticipated that it would have a significant effect on the overall reliability of the results. However, in order to check, the data was keyed in a second time. The results were found to be almost identical.

In addition to the test results, computer files were prepared of all the further information on each

student. This was based on the questionnaire described in Chapter 4 (see Appendix 4) and also administrative records kept by the British Council institute. There are between 300 and 700 records in each file. It was this file which allowed for the comparison of teacher assessments, attendance and placement test scores with progress test scores.

4. Overall characteristics of each test

Each test was analyzed in two ways. Firstly, it was treated as a unity, in the sense that none of the sections were analyzed separately. This means that the mean, standard deviation, reliability and standard error of measurement were established for the whole test. Then each section was treated as a separate test. This meant that there were four separate analyses of Listening, Grammar, Appropriacy, and Reading and Writing.

It was considered vital that the tests should perform well statistically for two reasons. Firstly, these tests were actually being used within the institute to make important decisions about the students concerned.

We had to feel confident that they were performing reliably and that any decisions that were to be based on the results could be justified to both students and teachers. Secondly, a firm statistical base was required in order to support any claims that may be made at a later date. The sample size was more than adequate for any of the subsequent analyses that were carried out and each analysis was based on the knowledge that the subtests conformed to high standards of reliability.

Table 5.1. below illustrates overall test and subtest statistics:

Table 5.1.

	WT	LIS	GRM	APP	RD/WT
A3 Test					
X	63%	55%	60%	81%	69%
SD	19%	24%	22%	24%	28%
KR20	0.95	0.92	0.88	0.84	0.92
NQ	89	28	29	10	22
NS	264	264	264	264	264
B1 Test					
X	54%	42%	52%	77%	53%
SD	16%	20%	21%	18%	28%
KR20	0.93	0.87	0.83	0.78	0.89
NQ	96	33	24	19	20
NS	305	305	305	305	305

B2 Test

X	58%	42%	57%	74%	65%
SD	14%	18%	18%	15%	19%
KR20	0.91	0.82	0.80	0.68	0.85
NQ	99	29	24	20	26
NS	259	259	259	259	259

B3 Test

X	56%	58%	48%	58%	63%
SD	17%	21%	16%	20%	26%
KR20	0.94	0.87	0.80	0.78	0.89
NQ	106	32	32	20	22
NS	201	201	201	201	201

C1 Test

X	57%	55%	46%	80%	64%
SD	16%	20%	19%	23%	24%
KR20	0.94	0.88	0.86	0.84	0.91
NQ	112	34	35	12	31
NS	250	250	250	250	250

C2 Test

X	58%	57%	49%	79%	62%
SD	18%	20%	21%	22%	27%
KR20	0.95	0.86	0.87	0.74	0.91
NQ	98	31	31	09	25
NS	242	242	242	242	242

C3 Test

X	59%	55%	40%	84%	82%	68%
SD	15%	18%	20%	21%	20%	30%
KR20	0.94	0.85	0.87	0.81	0.88	0.84
NQ	111	32	38	12	20	09
NS	221	221	221	221	221	221

1. All scores are expressed in percentages to allow for ease of comparability. The raw scores are all available on the printouts in Appendix 2.

2. *KEY*

WT = Whole Test
LIS = Listening
GRM = Grammar
APP = Appropriacy
RD/WT= Reading and writing
X = mean score;
SD = standard deviation;
KR20 = Kuder-Richardson 20 reliability quotient;
NQ = number of items in the test or subtest;
NS = number of students in the sample.

An overall mean score of 60% was the target for each of the tests. However, this was not achieved by any of them since all but one fell below the target mean by from one to six percent. This did not cause any undue concern and could be easily corrected in future administrations. 60% was selected as the target mean score because it was a familiar and acceptable standard of difficulty for the students. Although they were not told what the mean test score was, students 'know' if something is too difficult or too easy. It was hard to predict the exact level of difficulty for each test and equally challenging to try and maintain a level of difficulty throughout the battery. Considering this, the mean scores fall within acceptable limits. In Hong Kong schools, test mean scores are often lower than 60% but they are rarely much higher. To allow for too high a mean score in these tests might have reduced their value

and the value of the courses to some extent in the eyes of the students.

It was noted above that there was a tendency for the whole tests to be slightly more difficult than intended. This feature is an important one to be aware of when teachers are engaged in the construction of test batteries. In certain areas there seems to have been a fairly consistent tendency on the part of the test constructors to slightly overestimate the level of ability of the students they had been teaching.

The most difficult part of any test was generally the Grammar component. This may be due to several reasons:

i. teachers think that students know more than they actually do;

ii. accuracy tasks, such as those in the Grammar section, are actually more difficult than other tasks in the test;

iii. students are weakest in accuracy tasks;

iv. accuracy tasks are not given as much attention in the courses as other more realistic types of language use tasks, and consequently performance is weaker;

v. because this type of task was not as commonly used in teaching as the other tasks in the test the test constructors lacked the appropriate experience to gauge level of difficulty as accurately as with other tasks.

The extent to which any one of these reasons is responsible for the consistently more difficult status of the Grammar section is difficult to say. In all of the tests there were at least two and sometimes three or four cloze-type tasks. The difficulty level of the section as a whole could often be attributed largely to one of these passages rather than all of them.

The component in all of the tests, with the exception of the B3 Test, that achieved the highest mean score was the Appropriacy section. This is due in part to the criterion for correctness which was

adopted for the item analysis. It was explained in Chapter IV that this section was marked on two levels. If the response was appropriate, regardless of accuracy, then it was awarded a mark. If, on the other hand, it was inappropriate then it was awarded no marks, regardless of accuracy. Only once appropriacy had been established was accuracy taken into account. For computer analysis the degree of accuracy was ignored. The high mean scores in the Appropriacy section reflect this approach. The scores that teachers reported to students for this section were different from those used for the analysis in the sense that accuracy was a consideration.

In order to help teachers with marking and to achieve some degree of standardization, large numbers of scripts were taken and actual responses noted down and graded by a committee. The notes were distributed to teachers. A copy of these notes can be found in Appendix 11.

Overall the test statistics presented in Table 5.1. above are very good. The reliability of every whole test is above 0.9. In all but four of the twenty

eight subtests it is above 0.8. Considering that the number of items in the subtests was often fairly small the level of reliability permitted a high degree of confidence regarding the quality of the subtests.

The four subtests with a reliability of less than 0.8 were all appropriateness tests. Three of these in the B levels included ten multiple-choice questions out of a subtest total of nineteen or twenty questions. In general, a multiple-choice test requires more questions to achieve the same level of reliability as an open-ended test due to the unavoidable guessing factor that is far less significant with open-ended questions. The fourth Appropriacy subtest with a reliability of less than 0.8 was in the C2 test. In fact, there were only nine questions in this subtest so that the reliability of 0.74 can be considered rather high.

The standard deviation of the whole tests and subtests is always fairly high. One might expect it to be slightly lower considering the fact that the students in any level have been placed there on the basis of a complex and reliable placement procedure

which should have ensured a high degree of intra-level homogeneity. It should be noted however, that although the normal entry point for new students was at A3, B1 and C1, commercial pressure dictated that new students be allowed to enter at all levels in certain circumstances. Studies carried out indicated that the progress test performance of new students was often better than that of students who had come up through the system. This unavoidable state of affairs meant that the degree of homogeneity within a level was reduced.

While a higher standard deviation possibly indicates a wider spread of ability, it also indicates that the tests are succeeding in spreading the candidates effectively which is very important in grading and the distribution of the scores. With the whole tests and most of the subtests the level of the standard deviation is not a cause for concern; on the contrary, it makes certain administrative decisions, such as grading, somewhat easier to deal with.

In the case of the Reading and Writing section of all but the B2 test, the standard deviation is

particularly high when compared to the other sections. This is due in part to the fact that more students failed to complete this section than the others, since it is the last section of the test. Hence there were more lower scores than in the other sections. This could be interpreted as a design weakness of the tests, indicating that not enough time was allowed for the tasks to be properly completed. In fact,it was always a minority of students who did not complete any test, as is apparent from a study of the full test statistics available in Appendix 2. In addition, those who did not complete were generally the weakest in terms of performance on the test as a whole as indicated by the consistently low mean scores that they achieved as a group. Clearly the low mean scores were in part aggravated by students not finishing the test. However, even if this was the case, it is fairly safe to say that the students who did not finish the test were generally the weakest. This is why they did not finish. They were not the weakest because they did not finish.

The overall test results generated by the Item Analysis Programme demonstrate clearly that the items in the various progress tests perform, for the

most part, very well from a classical test analysis point of view. It can be said without doubt that Hypothesis One, which claims that performance-based communicative tests can be reliable measuring instruments, is strongly supported by these results and they provide the necessary foundation for further analyses.

Since it has been shown in this research that performance-based language tests can produce items, subtests and tests that perform to a very high standard as measured by classical test statistics there is no reason to adhere to traditional testing techniques, such as multiple-choice, on the grounds that they produce more reliable results. The tasks in this battery have high face and content validity, both of which are of great importance when tests form an integral part of a teaching/testing scenario. However, they can also be demonstrated to be highly reliable. This is a sound combination which can be achieved by careful preparation, pretesting and moderation with the cooperation of teachers - in short by following the guidelines laid down in Chapter IV. The much discussed conflict between testing formats and teaching materials is therefore quite avoidable. Judging by the results

achieved in this battery there is no reason for the test constructor to shy away from using items that closely reflect either teaching materials or real life activities on the grounds that they may not produce reliable results, since they clearly do. Indeed, it could even be argued that the results achieved by the tasks and items in this battery out-perform many more traditional formats, and do so more economically.

5. Issues Arising out of the Item Analysis

5.1. The Moderation of Tests

Each test in the battery had been moderated carefully on the basis of the item analysis, and changes made where necessary. It is very important that the moderation team is aware that it is fairly easy to get distracted by item statistics and so lose track of the original aim of a particular item or task in an attempt to make it statistically more acceptable. This can lead to a situation where purely internal test characteristics dominate the external role that an item was originally supposed

to play. In other words, unless care is taken at the moderation phase, there may well be a tendency for moderators to fall into the reliability/validity trap where the presence, format or content of an item or task in a test is justified primarily by its statistical characteristics internal to the test itself and not related to the initial aims of the test which are based on external criteria.

The more times a test is pretested, the more likely it is that the original aims of a task or item will be lost if it does not appear to function very well statistically. This being the case, it is often safest to discard an item or task that does not work well at an early stage, unless the reasons for its failure can be easily remedied. For example, it may be obvious that the main reason for an item's failure is due to unclear rubric. This can be readily corrected. Alternatively, it may be that candidates did not have enough time to complete the item, in which case the time allocation can be increased. However, in other cases there may be no obvious reason for an item's failure. In a task-based approach, such as the one adopted with this test battery, a task is made up of a number of individual items. For instance, taking down a

telephone message may involve three or four different skills represented by the same number of items. While they may be independent to an extent they may also be inter-related in some way if only in the sense that they are contextually bound together. In any case, the relationship between them is often fragile and unpredictable. Every task has an internal chemistry, for lack of a better term, and changes in aspects of a task can affect the internal chemistry unpredictably. Constant vigilance is required of moderation teams.

5.2. Subtest Analysis in Addition to Whole Test Analysis

The facility value of an item will not change whether it is analyzed as part of a whole test or as part of a subtest. The same number of students get an item right or wrong, regardless of how a test is subsequently split up. On the other hand, the relative ability of students in different subtests will not necessarily be the same. This difference will manifest itself in the magnitude of the point-biserial correlation of given items. The greater the degree of homogeneity of the skill or trait being

tested, the larger the point-biserial will tend to be, given that a particular item is functioning well. The greater the diversity, assuming it is based on different abilities, the lower the point-biserial correlation will tend to be.

While there were advantages when using the Item Analysis Programme to keying in whole tests, in that this allowed for greater ease in subsequent analysis, the programme allowed for the analysis of subtests as tests in their own right. This facility is of enormous value. In virtually all cases, when the point-biserial correlations of items were compared, they were higher when part of the subtest analysis than they were when part of the whole test analysis. This finding supported **Hypothesis Two** which claimed that students' ability in different subtests was not equivalent.

The point made above is also a significant consideration for moderators to bear in mind. It is not unusual in a teaching context for tests to focus on more than one trait. It is also likely that, for reasons of economy, a single item analysis is conducted and that subsequent moderation is based on

A wide range of tasks are tested in the Listening, Reading and Writing sections of the seven progress tests. Tasks were selected on the basis of their relevance to the course of instruction and the extent to which they reflected the sorts of activities that the students were likely to engage in their lives outside the classroom. While the skills being tested had certainly received coverage in the course, the tasks themselves may not have been covered. This was not seen as flying in the face of the principle of content validity. In the course students may have worked on a series of listening exercises, using the textbook, which involved noting down specific information about a Greek teenager living in London. In the test, they may have used the same, or very similar skills in the role of a secretary, to note down information about a Hong Kong businessman travelling to Singapore. While the teaching materials were often based on the textbook with questionable direct relevance to the student, the testing materials were always geared towards situations that students could perceive as relevant to their present or future needs. By adopting this focus, they emphasized the point that students were learning a language that they would actually be using, and encouraged

teachers to seek materials that were of direct relevance.

A major advantage of the performance-based approach is that it creates a positive washback effect on the both teaching and learning. However, care must be taken to ensure that the activities students engage in are indeed reflective, to some extent, of reality and not merely a product of the test constructors imagination.

An important observation concerning the language use of the students taking the tests is that the nature of the activities at different competence levels was not necessarily very different. For example, they all needed to be able to function adequately on the telephone, very often in a message taking capacity. Similarly, they were often asked to note down instructions or messages verbatim - in other words they had to take dictation. With the listening tasks, the differences between the levels lay in the amount of language that they could process at any one time, the speed at which they could process it, and the complexity of the input that they could deal with. Similar criteria are suggested by Morrow

(1979). With the writing tasks, the differences between the levels lay in the sophistication of the output, both from the point of view of accuracy and communicative effectiveness, and the nature of the stimulus texts used to make the tasks integrated in some way. It was these variables that were manipulated for the most part to make a given task appropriate to a particular level. When items were found to be too difficult, in a Listening test for example, the moderation team may have felt that the speed of delivery was responsible and so it was slowed down. We used our intuition about how a native speaker of English would speak to a minimally competent English user in Hong Kong.

This approach can and does often lead to criticism. By simply looking at the Listening sections of the seven tests under discussion in this thesis, a naive observer may comment that they all look the same and therefore by inference that the students are learning nothing different as they progress through the battery. This could reduce its face validity. In reality this is of course not the case. They are learning to process language more effectively as the courses get more advanced and to deal with more complex and less clear-cut realizations of language.

Several important considerations about item construction and performance were revealed by the item analysis. These are discussed below.

5.3.1. The Relationship Between Difficulty and Discrimination

Although this section makes an obvious point, it has been included because teachers were frequently confused by the difficulty/discrimination issue. Many teachers believed that there would be a direct relationship between the difficulty of an item and its discriminating power. Quite naturally, they want to feel that a test will provide a challenge for the more able students. At the same time, they do not want the less able students to score unnecessarily low marks.

Teachers can predict fairly accurately how difficult an item will be. Unless otherwise informed, they will assume that difficult items by default challenge the more able students. Unfortunately this

is not the case as the item analysis revealed in many instances. Clearly, when a test is being used as a grading or sorting instrument, a function which is required of most tests, each item needs to discriminate well. Moderating teachers were surprised to see that a difficult item did not necessarily discriminate well. They quickly realized that if this were the case then the item was redundant - it had no positive role to play in the test as a whole.

While these concepts are obvious when they are pointed out and demonstrated, they are not immediately apparent to the naive test constructor. The moderators realized that although they had to rely on their intuitions to some extent, regarding the relative difficulty of items and tasks, they needed the results of the item analysis in order to make final decisions as to test content. It became apparent to them that they were dealing with a complex and unpredictable area of knowledge and ability in the construction of test items. It also became apparent to them that this complexity applied equally to their teaching materials, that intuition alone was not always an adequate starting and finishing point. In this sense, teachers who work on

test construction in a principled way not only become better test constructors, but also more sensitive materials writers.

5.3.2. The Problem of Establishing Progress and Relative Task Difficulty

Whenever a battery of progress tests is produced one of the problems that arises is related to ensuring that tasks and items are actually performed better as the level of the students taking the tests increases. In many cases this is not a question that needs to be addressed since the nature of the tasks taught and tested is different at different levels and the relative difficulty of given tasks is of little importance. With a performance-based battery on the other hand the superficial nature of tasks at different levels may often be very similar. The problem of demonstrating progress and relative difficulty can be dealt with in several ways.

Firstly, the complexity of the input can be controlled. It is unfortunately not always clear what constitutes complexity. With the Listening

tasks it can be related to the amount of redundancy present, the speed of delivery and the amount of interference, such as white noise. With Reading tasks, complexity of input is related to the linguistic nature of the text, its content matter, and the amount of background knowledge that the students have.

Secondly, the nature of the output required of the students can be controlled. With Listening this may entail taking longer telephone messages, noting down names that are not spelt and so on. With Reading tasks, the degree of comprehension required is controlled. With Writing tasks, the subtlety, accuracy, relevance, logical organization and appropriacy of the message to be produced can be evaluated.

Both of the approaches outlined above, the manipulation of input and output, will generally be based largely on the intuition of the test constructor and moderation team. Unfortunately, while tasks may be intended to be more difficult or demanding, there is no guarantee that they will be. Intuitive methods were obviously used in the

construction of the battery under discussion in this thesis. However, it was felt that there should be some way of retrospectively demonstrating that these intuitions were accurate. Two methods were considered. The first involved students at different levels taking the same tests. In order to do this some teachers were asked to administer a range of progress tests to their classes. This method did not work because it intruded too much into class time and led to dissatisfaction both on the part of the teachers involved as well as the students. The alternative approach is to use certain items in more than one test and then compare performance across levels. Anchor items were used in three sections of the tests, Listening, Grammar and Appropriacy. Through these items we were able to show clearly that students were getting better as their level increased thus demonstrating that progress was taking place.

5.3.3. The Complexity of Task Format

One of the major causes of tasks failing to meet adequate measurement criteria was their complexity. In real life complex tasks are a problem at the best

of times. All of us have struggled, often unsuccessfully, with complicated forms, for example. Our failure is generally not due to an inadequate command of the language but rather to a lack of appropriate knowledge of a particular domain of use, or simply an inability to deal with forms as a medium of communication. The same applies to test items. The failure of a student to complete a particular task or item successfully may not be a consequence of a lack of language ability but may well be the result of unfamiliarity with a task, or an inability to unravel the complex nature of a task in the time allotted. Using conventional test items this problem of complexity is minimized because the item types used are of a very limited scope. As soon as we move into the realm of performance/task-based testing, on the other hand, we have to accept that the range of item and task types is virtually infinite. We are no longer simply in the world of the language test but rather in the world of real or simulated communicative events. All the problems that individuals may have with successfully doing the 'real thing' are magnified in the test situation because of the inevitability of test pressure. It is often unclear whether success or failure on a task or item is due to inadequate control of language or something quite different. The test constructor is

working in the dark to a large extent and has to rely on intuition, on the one hand, and the item analysis on the other.

We found that the difficulty level of a task often increased correspondingly with its complexity. In addition, the discrimination of the items in such a task was also adversely affected. A good example of this is task 3 in the C2 Listening Test. This task requires the student to make changes to an appointments' diary. The task itself is realistic yet it is also rather complex in nature. In order to make it work, we were obliged to ensure that the things the student had to cope with, in addition to carrying out the instructions that formed the basis of the question, were as straightforward as possible. We also had to feel confident that the average student would not be initially deterred by the complexity of the task format. This was achieved through pretesting and moderation. The importance to the quality of the final product of these two phases cannot be emphasized too strongly

Because Listening tests take place in real time the issue of complexity interference is of special

importance. One way of reducing the problem is to allow students sufficient time to study task format before actually doing the task itself. Unfortunately, it is only the test-wise, or possibly life-wise, student who takes advantage of this facility. Many students, according to observations carried out during test administration, were busy looking back over a previous question, or simply inactive in the time allowed for familiarization.

In producing task-based items it is important that the complexity of the task is not in itself an insurmountable obstacle to the student completing it. At the same time there must be sufficient reality in the task format and the demands of the task to make it a meaningful and realistic activity. This balance is often difficult to achieve and the results of an item analysis provide vital information for the test constructor.

5.3.4. The Problem of Memory

The extent to which short-term memory interferes with task completion is a problematic issue in the

construction of task-based listening tests in particular. The relationship between memory and listening comprehension, for example, is unclear. We did not want to get into a situation where the successful completion of a task hinged primarily on a student's ability to remember a particular thing rather than on his ability to understand it, where another student may also have understood but, because his short-term memory was less efficient, could not complete the task. In addition, with a task-based format, students are often required to negotiate their way around a table, chart or some other, possibly complex, grid format. At a very basic level the progression through the grid may be ordered in the same way as on the grid. At a more advanced level, however, it may be more realistic to move unpredictably around the grid. In such cases, it is often useful if the student can remember where particular types of information are to be found. Again, short-term memory can play an important role.

In order to reduce the impact of these problems students were advised to familiarize themselves with task format or questions before actually doing the task. However, as with the issue of complexity discussed above, many students did not take

advantage of this opportunity. The extent to which this had an effect on performance is not clear. From experience in pretesting, we know that in general terms Listening tasks requiring a lot of short-term memory recall, or involving complex formats never seemed to provide us with adequate results on the item analysis. The question of whether conventional item analysis is the most appropriate way of judging these types of tasks is an obvious one to ask at this point. Unfortunately, it is the only readily available methodology that has been developed to deal with the problem. Its strength is that it provides the test constructor with data that can be interpreted. Its weakness is that it is based exclusively on the internal context in which the task appears and cannot take into account external factors. Such external factors are inevitably left to the intuition of the test constructor. It is only possible to investigate the impact of external factors by pretesting items in different contexts in order to establish whether they perform in the same way. This is perhaps the catch-22 of conventional item analysis. It is context-bound, and there is no guarantee that items will perform in the same way in different contexts. This problem is widely recognized with regard to the unidimesionality of trait and was dealt with in this investigation by a

flexible approach to item analysis made possible by the specially developed software. However, it is potentially an equally serious problem when it comes to different task types. It can only be dealt with by placing tasks in a number of different contexts and analyzing the results. This would be time consuming and difficult, and was beyond the scope of this study.

5.3.5. The Problem of Spatial Awareness

In a performance-based approach it is often necessary to try and replicate reality through the use of visual aids. Certain types of task, particularly in the Listening section of the tests, proved very difficult to implement successfully. For example, one of the skills that students were taught at several of the levels, was the giving and receiving of directions. An obvious way of testing mastery of this skill was to ask students to follow directions on a map. It was found that this type of item, which is common in teaching materials, did not work at all well in the testing context. It seems that, negotiating around a map is the sort of thing that some people can cope with and others cannot.

The ability to do so does not appear to be related simply to language competence. We identified the problem as being related in some way to the spatial awareness of an individual.

A similar difficulty arose with task one in the B2 Listening test which involves identifying the location of common objects in an office. The objects were listed on the test paper in an attempt to facilitate the task, yet even so, three of the six items had a facility value of less than 50%, and the point-biserial correlation was greater than 0.4 with only two of the items. This was after considerable efforts to make the task manageable. We were obliged to accept that this type of task involved some sort of mental processing not directly related to language ability and despite our efforts, were not able to make it work well. Our solution was to avoid the use of this item type in the Listening sections of the battery.

5.3.6. The Use of Cloze

The cloze technique in its pure form requires the use of authentic text and an nth word deletion rate. Far reaching claims as to the intrinsic reliability and validity of cloze have been made by Oller and others and later refuted by Alderson (1978, 1983) and Lee (1985). It is not clear exactly what is being tested by cloze. There is certainly a reading aspect. After all, the technique was originally devised (Taylor, 1953) in order to measure reading ability in native speakers. It has also been used to establish the readability of texts. At the same time, cloze may be said to measure knowledge of vocabulary and language systems.

When planning the test battery under discussion in this thesis, it was seen as necessary to include a component that focused on an ability to deal with grammar in context. The discrete-point multiple-choice format is commonly used to test this ability. There is no doubt, however, that this format does not have a positive washback effect on classroom practice because it encourages a discrete-point view of language in use. It was decided to use cloze as

an alternative in this battery. It had the advantage of presenting language in context and did not appear to have as negative an effect on classroom practice. However, the nth word deletion strategy appeared to be too random. When it was suggested, teachers were not convinced that it was a valid way of testing. It was therefore decided to adopt a rational deletion policy, since this allowed for the selection of testing points related to the areas taught in the course.

Unfortunately it was found that the easiest items to select and mark were closed sets, such as prepositions, pronouns etc. Moreover, this type of item also tended to discriminate the best. However, it did not seem appropriate to focus exclusively on this type of item. Setters and moderators were therefore encouraged to extend the range of areas that a cloze passage tested.

When subjected to initial item analysis, it was found that if 50% of the items performed adequately, this could be considered a successful pretest. However this often meant that a the passage was no longer an economical testing instrument in that it

involved too much work on the part of the student for too little reward in terms of marks in the test. The strategy used to overcome this difficulty was to pretest the same passage twice but to delete a different set of words. In moderation, the two passages were combined, and the best items from each retained. In this way, we were generally able to maintain a reasonable deletion rate.

5.3.7. The Use of Dictation

Dictation as a testing instrument was popular many years ago. It was then rejected and discouraged until very recently. Dictation for dictation's sake is not easily justified in a teaching context. It has been claimed by Oller and others that it, like cloze, is a good measure of underlying competence. Unfortunately, whether this is true or not, it does not convince teachers that it should be used in teaching or testing. And yet, there are many instances in real life when we have to take dictation even though we do not perceive it as such.

It was decided to experiment with dictation as a task-type in the battery, but to make sure that the contexts in which it occurred were as realistic as possible. The question of context is very important in the whole approach advocated here, and no less so with regard to dictation. It may be that if no attention had been paid to context, the statistical characteristics of the items would have been the same. No experimental work was carried out to check this. However, from the point of view of the interaction of the tests with teaching and the real world activities of the students, it would have been unthinkable. The amount of time required to contextualize a short dictation, and after all we rarely engage in anything other than short dictations, is minimal. The washback effect on the other hand is enormous.

We found that the dictations were a robust and effective testing instrument. However, the scoring was not based on the word but rather on the meaning unit. This meant that moderation of unsuccessful items was rather difficult and time consuming. In order to find out exactly why a particular unit had failed scripts had to be carefully examined. It may be that at the pretest phase it would be advisable

to treat each word as an item in its own right. Although this is more demanding initially, it might pay dividends in the long run.

5.3.8. Testing Apporopriacy

Canale and Swain (1981) suggest a sociolinguistic dimension to communicative competence, and an attempt was made in this battery to test the students' ability to deal appropriately with situations they were likely to encounter in their day-to-day lives. The major focus on sociolinguistic competence in this battery was in the appropriacy section of the tests, and was based on face-to-face interaction, although appropriacy of response was also a marking criterion in the Writing sections. It was accepted that the tasks in the appropriacy sections were of an indirect nature. Two main formats were used. The first was an open-ended half-and-half dialogue, and the second multiple-choice.

The half and half dialogues worked well statistically and they were in line with teaching methodology. It is not an uncommon teaching strategy

to have students work individually or in small groups on an incomplete dialogue as a prelude to oral practice. Thus the task-type was familiar to both teachers and students. The open-ended nature of the tasks helped to make them discriminate well. We were fortunate in that all the markers were native speakers of English and hence were able to judge appropriacy of response reliably. We were able to establish this through the inspection of scripts. There was some inconsistency when it came to grading for accuracy, but even here we were satisfied that this was not a major problem.

The multiple-choice tasks employed at the B-level performed fairly well. The problem with these tasks, however, was making the options right or wrong a consequence of appropriacy as opposed to grammatical accuracy. This problem made the setting of items rather time consuming. In addition, we were not really happy with the use of multiple-choice in the battery because of its negative washback effect. It was generally felt that this type of task should be excluded from the battery in later administrations.

5.3.9. Testing Reading

Conventional approaches to the testing of reading involve multiple-choice and open-ended questions. The former was rejected as a viable option in the context of the teaching institute because of the negative washback. The latter was adopted with some modifications.

One of the problems with the testing of reading is the extent to which comprehension of the question is confused with comprehension of the text. Shohamy (1984) for example, was able to demonstrate that Israeli children performed better on reading comprehension tasks, whether open-ended or multiple-choice, when the questions were in Hebrew as opposed to English. The same problem arises with listening comprehension. We were able to overcome it there, to some extent, by making the questions as free of obvious verbal distractions as possible through the use of telephone message forms and a variety of other grid-type question formats. With the reading comprehension tasks this was not really possible except at the lowest levels. For example, in the A3 test the visa application form was used. Even here

it was found that students' confusion over the meaning of first name and surname on the form proved to be a problem. At higher levels we found it impossible to find suitably difficult tasks of a similar nature.

It was considered important that the texts used should be authentic where possible, or that they looked as if they were. Texts were selected on the basis of their availability to students and teachers outside the testing context, so that there was a ready source of relevant practice material. The reading tasks themselves were made as meaningful as possible. For example, in the B1 test a page from the local Yellow Pages Telephone Directory was used. The questions were based on extracting information that might realistically be extracted by someone wishing to get information from the directory. Similarly, at the C1 level, extracts were taken from the Oxford Advanced Learners Dictionary. The tasks were based on an investigation of how students actually used the dictionary. For example, one of the problems that clerical staff had was having to type up a hand-written text and not being able to read the hand writing. They could decipher the first letter or two but not the rest. This meant that they

had to use a dictionary to try and get the precise word. We felt that it was important to try and use authentic texts and authentic tasks in order to ensure that both students and teachers felt that the questions were relevant. In general the questions functioned well statistically and no adverse comments were received from teachers or students.

An attempt was also made to link the reading passage to the writing task in some of the tests. In real world communication situations tasks of this nature are often integrated. Integrating the tasks in the tests, where possible, proved to be a successful strategy that heightened the meaningful nature of the tests.

The reading texts and tasks were always of a very functional nature and it may have been a mistake to keep them so. We wanted to encourage teachers to use and students to read real English that was readily available. It did not seem realistic to focus on literature since the average Hong Kong adult does not really read for the sake of reading as is common in Europe. However, the tests may have been able to encourage a greater interest in extensive reading.

In fact, many teachers exhorted students to read graded readers in an attempt to broaden the scope of language input. At one time we considered the possibility of using graded readers as a source of test materials but rejected the idea because it seemed out of line with the approach adopted by the battery. This meant of course that the questions based on the texts actually used, tended to be of a very basic nature. Higher order reading skills were not really tested. However, higher order skills did not seem to be a feature of the sort of reading that most of the students engaged in.

5.3.10. Testing Writing

Testing writing involves the extraction of a sample of written text from the student. In tests this simply means setting a question which requires the student to write something. The task is often highly contrived. The general direction of the battery under discussion in this thesis was towards the creation of realistic and meaningful contexts for the tasks employed. To this end realistic writing situations as opposed to 'write an essay' types of questions were used. Most of the tasks were based on

writing short letters, notices or instructions of the sort that students could be expected to engage in in their day-to-day lives. In addition, the writing tasks were integrated with either listening or reading input where possible.

It is not normal to subject writing tasks to conventional item analysis. However, all of the writing tasks were subjected successfully to item analysis in this battery. This was made possible by breaking up the scoring procedures to include layout as well as content, accuracy and appropriacy. The Writing tasks proved to be highly effective testing instruments

6. Correlational Analysis of the Subtests and Other Related Factors.

The item analysis of the battery provided invaluable feedback for test moderation and also showed that the battery was highly reliable. Detailed item statistics indicated that many students performed differently in subtests, and that their ability in different traits may not be the same. To investigate

the issue of variable performance further, each test was subjected to correlational analysis in order to establish the degree to which subtests and tasks were related to each other, and to other measures. Correlational analysis was seen as the next logical step in the validation of the battery.

6.1. The Variables to be Correlated

It has already been established that the subtests as well as the written placement test were reliable measures in conventional testing terms. The item analysis results indicated that the subtests were probably measuring different aspects of communicative competence. Hence a very high correlation was not expected between the various subtests. Should a high correlation occur, then it could be argued that certain subtests may be redundant. It was also decided to compare the placement measures, teacher assessments and attendance at the courses with performance on the progress tests. Twelve variables were included in the study. These are:

i. **Part 1** (P1) of the progress test, the listening section;

ii. **Part 2** (P2) of the progress test, the grammar section;

iii. **Part 3** (P3) of the progress test, the appropriacy section;

iv. **Part 4** (P4) of the progress test, the reading and writing section;

v. The **Total** (T) score on the progress test;

vi. The **Written Placement** (PW) test score;

vii. The **Oral Placement** (PO) test score;

viii. The **Teachers' Written** (TW) subjective assessment;

ix. The **Teachers' Oral** (TO) subjective assessment;

x. The teachers' subjective assessment of **Progress** (PR);

xi. The teachers' subjective assessment of **Effort** (EF);

xii. The students' **Attendance** (AT) at courses for the term.

Results from three of the seven tests will be discussed in some detail in the following sections. A similar discussion of the four remaining tests can be found in Appendix 9.

6.2. The A3 Progress Test

The four subtests correlate fairly well, the most significant relationship being between Part 2 and Part 3 which is somewhat surprising since it was expected that Part 1 and Part 4 would be most closely related. The relationship between these two variables was fairly strong (0.478) and may have been better had all the students completed the test. As mentioned earlier, there were a significant number of omissions towards the end of the test.

Table 5.2.

	P1	P2	P3	P4	T	PW	PO	TW	TO	PR	EF	AT
P1	1.000	.381	.349	.478	.755	.440	.441	.290	.266	.305	.199	.062
P2	.381	1.000	.547	.394	.772	.353	.353	.357	.156	.269	.147	.218
P3	.349	.547	1.000	.389	.682	.404	.385	.162	.047	.323	.170	.239
P4	.478	.394	.389	1.000	.729	.316	.307	.330	.201	.243	.146	.094
T	.755	.772	.682	.729	1.000	.502	.496	.384	.213	.376	.220	.199
PW	.440	.353	.404	.316	.502	1.000	.876	.163	.113	.524	.425	.166
PO	.441	.353	.385	.307	.496	.876	1.000	.090	.066	.607	.496	.237
TW	.290	.357	.162	.330	.384	.163	.090	1.000	.240	.148	.079	.000
TO	.266	.156	.047	.201	.213	.113	.066	.240	1.000	.036	-.004	.004
PR	.305	.269	.323	.243	.376	.524	.607	.148	.036	1.000	.849	.384
EF	.199	.147	.170	.146	.220	.425	.496	.079	-.004	.849	1.000	.384
AT	.062	.218	.239	.094	.199	.166	.237	.000	.004	.384	.384	1.000

The correlations between the subtests and total progress test scores are all fairly high, as might be expected, though this is due, in part, to the fact that the subtests are being correlated with themselves to some extent. None of the subtests has a significantly stronger relationship with the total test score than any of the others. This is consistent throughout the battery and indicates that no subtest appears to be a consistently better predictor of overall competence.

Teachers' assessments of students' written work correlate best with the Grammar section of the progress test, while their assessment of oral work is most highly correlated with the Listening section. This seems to be a reasonable finding although it is not consistent throughout the battery. Teachers' assessments act as a measure of the concurrent validity of the tests in the battery. While they are at no point strikingly strong, they are generally high enough to indicate that there is some agreement between the two sets of scores. Of course it may be that the teachers were influenced in their subjective assessments by the progress test scores. The extent to which this is the case was not investigated.

Interestingly, attendance at courses correlates quite strongly with teachers' perceptions of progress and effort and not so well with the progress test scores. This is a fairly consistent trend throughout the battery. It may be that students who come to class more frequently make more progress than those that do not and it certainly seems likely that they make more effort or it may be that teachers simply perceive this to be the case. At the A3 level there is also a very significant correlation between progress and effort and the placement test scores. This result is not repeated in the rest of the battery.

The subtest was divided into eighteen variables that were roughly equivalent to the tasks in the test. The results of the correlation are presented below.

Table 5.3.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15	V16	V17	V18
V1	1.000	.470	.526	.557	.516	.477	.252	.205	.208	.254	.165	.187	.268	.110	.250	.296	.213	.156
V2	.470	1.000	.484	.548	.448	.486	.306	.345	.242	.271	.248	.221	.236	.125	.201	.294	.218	.170
V3	.526	.484	1.000	.492	.472	.446	.307	.324	.316	.285	.323	.248	.307	.233	.304	.324	.252	.182
V4	.557	.548	.492	1.000	.468	.485	.118	.180	.195	.138	.222	.233	.214	.160	.236	.249	.239	.138
V5	.516	.448	.472	.468	1.000	.502	.333	.255	.322	.198	.207	.103	.194	.203	.303	.356	.258	.203
V6	.477	.486	.446	.485	.502	1.000	.328	.239	.255	.246	.244	.169	.225	.194	.275	.372	.249	.230
V7	.252	.306	.307	.118	.333	.328	1.000	.567	.448	.583	.390	.245	.400	.180	.320	.483	.305	.249
V8	.205	.345	.324	.180	.255	.239	.567	1.000	.432	.584	.337	.246	.408	.254	.381	.426	.438	.266
V9	.208	.242	.316	.195	.322	.255	.448	.432	1.000	.462	.395	.261	.439	.295	.201	.365	.257	.180
V10	.254	.271	.285	.138	.198	.246	.583	.584	.462	1.000	.316	.344	.489	.187	.326	.427	.284	.122
V11	.165	.248	.323	.222	.207	.244	.390	.337	.395	.316	1.000	.376	.353	.223	.226	.322	.126	.108
V12	.187	.221	.248	.233	.103	.169	.245	.246	.261	.344	.376	1.000	.593	.205	.123	.211	.088	.134
V13	.268	.236	.307	.214	.194	.225	.400	.408	.439	.489	.353	.539	1.000	.232	.315	.422	.331	.174
V14	.110	.125	.233	.160	.203	.194	.180	.254	.295	.187	.223	.205	.232	1.000	.357	.336	.417	.286
V15	.250	.201	.304	.236	.303	.275	.320	.381	.201	.326	.226	.123	.315	.357	1.000	.491	.626	.427
V16	.296	.294	.324	.249	.356	.372	.483	.426	.365	.427	.322	.211	.422	.336	.491	1.000	.533	.407
V17	.213	.218	.252	.239	.258	.249	.305	.438	.257	.284	.126	.088	.331	.417	.626	.533	1.000	.548
V18	.156	.170	.182	.138	.203	.230	.249	.266	.180	.122	.108	.134	.174	.286	.427	.407	.548	1.000

Listening - V1 - V6 Grammar - V7 - V10

Appropriacy - V11 - V13 Read/Write - V14 - V18

Before any discussion of Table 5.3. begins, it should be noted that this analysis, and the ones that are to follow, are concerned with the convergent validity of the tasks as opposed to their divergent validity as defined by Clifford (1980) and Stevenson (1980) who were among the first researchers in recent years in Applied Linguistics to consider the question of using the multitrait-multimethod matrix with regard to the construct validation of language tests. This approach, first proposed by Campbell and Fiske (1959),

necessitates the measurement of at least two traits by two methods. It requires, according to Clifford (1980)

> "... (1) that separate methods measuring the same trait correlate more highly with one another than they do with other traits measured by different methods and (2) ideally, separate measures of the same trait correlate more highly with one another than with different traits measured by the same method."

The tests discussed here were not able to meet these conditions for the most part which meant that the multitrait-multimethod matrix could not be applied. On a more basic level however, tasks measuring the same trait were marked on the correlation matrices and the extent to which they correlated more highly with each other rather than with tasks measuring different traits was considered.

In addition, the mean correlation of each variable with other variables measuring the same trait was calculated and compared with its mean correlation with the variables intended to measure different traits. This allowed for a numerical value to be arrived at, which was able to show that in almost all cases the mean intra-trait correlations were higher than the mean inter-trait ones. The table below illustrates this:

Table 5.4.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
IAT	.52	.49	.48	.51	.48	.48	.53	.53	.45	.54
IRT	.22	.24	.28	.19	.24	.25	.30	.31	.28	.28

	V11	V12	V13	V14	V15	V16	V17	V18
IAT	.37	.46	.47	.35	.48	.44	.53	.42
IRT	.26	.20	.31	.20	.27	.35	.25	.18

IAT = Intra-trait correlation

IRT = Inter-trait correlation

This method of comparison clearly shows that the intra-trait relationships are more powerful than the inter-trait relationships in all of the above cases. One might expect this with the Grammar and Appropriacy tasks, since they are very similar. However, it was considered possible that the distinction may get blurred with the Listening and Reading/Writing sections because the task format in these sections is rather similar. However, task format did not appear to be having an effect.

The Listening tasks (V1 - V6) correlate most highly with themselves indicating that the listening skill as tested at this level is in some way distinct from the other skills. The same can be said of the four grammar tasks (V7 - V10). The appropriacy tasks are not as neatly differentiated as the other skills areas mentioned above. While V12 and V13 correlate highly, V11 does not. This may be due in part to the fact that V11 has only two marks.

The reading and writing task was divided into five parts because it represented so many points relative to the other variables. The divisions were as follows:

V14 - 68 - 70

V15 - 71 - 73

V16 - 74 - 76

V17 - 77 - 81

V18 - 82 - 89

These five variables correlate generally most highly with each other, indicating that they form some sort of group, which indeed they do. V14 is the weakest of the five, due probably to the confusion caused by first and last names. This point was mentioned in the discussion of the item analysis results.

It seems clear therefore, that the correlational analysis supports **Hypothesis Two** which claims that the different subtests are testing different skills, and that the tasks are grouped in the way that it was anticipated they would be. This issue is further investigated when the tasks are subjected to factor analysis.

6.3. The B1 Progress Test

At this level a striking feature of the correlation table is the relatively high correlation between the teachers' subjective assessments of written and oral ability with Part 1 of the progress test (0.456, 0.448). It is much higher than in the A3 test (0.290, 0.266). The teachers' assessment of written ability correlated highest with Part 4 of the progress test which is an encouraging result. Interestingly, at the other levels, the teachers' assessments did not necessarily correlate most highly with Part 1. With the B3 test for example, the teachers' assessment was most highly correlated with Part 2, the grammar component. There does not appear to be a predictable relationship

between the progress test subtests and teacher assessments. This may be due to differences in the tests, or to the inherent unreliability of teacher assessments. Although great efforts were made to standardize teachers' views on what constituted competence when it came to their subjective assessments, it is quite possible that their views differed nonetheless.

Table 5.5.

	P1	P2	P3	P4	T	PW	PO	TW	TO	PR	PO	AT
P1	1.000	.328	.411	.440	.491	.285	.272	.456	.448	.368	.273	.081
P2	.328	1.000	.482	.415	.399	.318	.218	.323	.290	.292	.204	.045
P3	.411	.482	1.000	.358	.366	.248	.171	.415	.375	.323	.287	.088
P4	.440	.415	.358	1.000	.516	.489	.318	.457	.391	.235	.124	.114
T	.491	.399	.366	.516	1.000	.330	.289	.365	.299	.213	.127	.036
PW	.285	.318	.248	.489	.330	1.000	.428	.275	.266	.220	.075	.000
PO	.272	.218	.171	.318	.289	.428	1.000	.201	.259	.272	.098	.026
TW	.456	.323	.415	.457	.365	.275	.201	1.000	.830	.533	.429	.135
TO	.448	.290	.375	.391	.299	.266	.259	.830	1.000	.616	.464	.197
PR	.368	.292	.323	.235	.213	.220	.272	.533	.616	1.000	.720	.260
EF	.273	.204	.287	.124	.127	.075	.098	.429	.464	.720	1.000	.321
AT	.081	.045	.088	.114	.036	1.000	.026	.135	.197	.260	.321	1.000

The correlations between the subtests are fairly low. Part 4 correlates highest with the total progress test score, and with the written placement test. As with the A3 test, Part 2 and Part 3 correlate most highly with each other. Parts 1 and 4 correlate slightly less well. It was anticipated that the latter relationship would

be closer than the former since the task types seem to be more closely related. Perhaps the lower correlation is due to the fact that a certain number of students did not complete the final section of the test.

Teachers' assessments of progress and effort again correlate most highly with attendance indicating that teachers' perception of these two features may be linked to the frequency with which they see students. This seems to be a logical relationship and the trend was repeated at all the other levels.

The B1 test was broken down into fifteen tasks. V1 - V7 were the listening tasks, V8 - V9 the grammar tasks, V10 - V12 the appropriacy tasks, V13 - V14 the reading, and V15 the writing. The listening tasks were generally most highly correlated with each other. The two dictation tasks, V6 and V7 were most closely related, indicating that the similar nature of the task may be playing a role. This could be a method effect.

Table 5.6.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1	1.000	.364	.225	.243	.264	.145	.106	.192	.074	.039	.157	.153	.138	.028	.091
V2	.364	1.000	.428	.333	.460	.293	.192	.218	.079	.057	.231	.200	.271	.191	.207
V3	.225	.428	1.000	.389	.343	.518	.366	.265	.141	.084	.278	.231	.292	.200	.183
V4	.243	.333	.389	1.000	.428	.468	.385	.166	.184	.193	.273	.302	.313	.263	.161
V5	.264	.460	.343	.428	1.000	.355	.273	.166	.089	.057	.226	.171	.425	.313	.328
V6	.145	.293	.518	.468	.355	1.000	.568	.203	.060	.047	.406	.242	.287	.351	.139
V7	.106	.192	.366	.385	.273	.568	1.000	.048	.284	.319	.455	.306	.274	.369	.133
V8	.192	.218	.265	.166	.166	.203	.048	1.000	.123	-.038	.176	.206	.285	.101	.195
V9	.074	.079	.141	.184	.089	.060	.284	.123	1.000	.775	.177	.407	.234	.158	.177
V10	.039	.057	.084	.193	.057	.047	.319	-.038	.775	1.000	.136	.286	.234	.160	.082
V11	.157	.231	.278	.273	.226	.406	.455	.176	.177	.136	1.000	.475	.208	.281	.163
V12	.153	.200	.231	.302	.171	.242	.306	.206	.407	.286	.475	1.000	.193	.193	.176
V13	.138	.271	.292	.313	.425	.287	.274	.285	.234	.234	.208	.193	1.000	.433	.401
V14	.028	.191	.200	.263	.313	.351	.369	.101	.158	.160	.281	.193	.433	1.000	.490
V15	.091	.207	.183	.161	.328	.139	.133	.195	.177	.082	.163	.176	.401	.490	1.000

Of the seven listening tasks V5, which involved noting down specific information, was most closely related to the two reading tasks, V13 and V14. These two tasks also involved locating specific information. It is possible that the relatively strong relationship between these two task types is due to some underlying skill factor.

The two open-ended appropriacy tasks (V11 and V12) were correlated fairly highly with one another, but not as highly with the multiple-choice appropriacy task (V10). This variable was however, correlated very highly (0.7751) with the second grammar task, V9. It is interesting to note that the grammar passage, V9, is in

dialogue format, so that it might be argued that there is a textual similarity between it and the multiple-choice dialogue items in V10. The grammar passage was re-used in the B3 test as V8 correlating fairly well again (0.5271) with the multiple-choice appropriacy task, suggesting that the high correlation was not accidental. Clearly the nature of the two tasks is quite different, one being productive and the other only requiring recognition of the correct answer, yet the nature of the text types is similar. Text type, even at this very crude level appears to have an influence on test performance.

When the correlations were averaged, the following results emerged:

Table 5.7.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
IAT	.23	.34	.39	.38	.35	.39	.32	.12	.12	.22
IRT	.11	.18	.21	.23	.22	.22	.27	.17	.22	.17

	V11	V12	V13	V14	V15
IAT	.31	.39	.42	.46	.45
IRT	.25	.23	.26	.22	.17

The intra-task mean correlations are not as high in the first three sections of the test as they were at the A3 level, whereas, they are fairly strong in the final section. However, the inter-task correlation means are all significantly lower with the exception of the two grammar tasks. It will be remembered that they correlated very poorly with each other and it was suggested that this may have something to do with the nature of the task. It may also have been due to the fact that V8 was far easier overall than V9. This too would be bound to affect the correlation.

It is difficult to make any conclusive statements about the relationships between the variables in this or any other of the tests under discussion here. However, it seems that on some occasions, the relationship between the tasks can be attributed, in part at least, to an underlying skill factor, for example V1 - V7 where the skill factor is supposed to be listening, or V11 - V12, where it is appropriacy of response. On other occasions it seems that the nature of the task may be exerting an influence as with the V5 - V13 - V14 relationship. On yet other occasions there appears to be some sort of text effect as is demonstrated by the V9 - V10 relationship.

6.4. The C1 Progress Test

The correlations between the subtests at this level are not particularly high, all of them falling below 0.4. However, the extensive item analysis carried out on these subtests provides satisfactory evidence that they are consistent in their measurement of the various traits. The level of the correlations in this test, and indeed all the others, is not therefore due to weak initial measures. It seems more likely that it is due to the fact that students have different abilities in the different aspects of communicative competence measured by the subtests. Such a conclusion is supported by the item analysis results and by Hypothesis Two.

Table 5.8.

	P1	P2	P3	P4	T	PW	PO	TW	TO	PR	EF	AT
P1	1.000	.370	.272	.250	.699	.293	.207	.235	.305	.085	-.026	.055
P2	.370	1.000	.391	.323	.724	.539	.207	.349	.318	.152	.048	.085
P3	.272	.391	1.000	.130	.515	.256	.083	.233	.246	085	.058	.032
P4	.250	.323	.130	1.000	.653	.390	.093	.352	.269	.129	.056	.110
T	.6996	.724	.515	.653	1.000	.551	.209	.444	.411	.184	.065	.109
PW	.293	.539	.256	.390	.551	1.000	.291	.313	.213	.121	.040	.136
PO	.207	.207	.083	.093	.209	.291	1.000	.139	.095	-.035	-.021	-.060
TW	.235	.349	.233	.352	.444	.313	.139	1.000	.767	.342	.268	.075
TO	.305	.318	.246	.269	.411	.213	.095	.767	1.000	.322	.224	.064
PR	.085	.152	.085	.129	.184	.121	-.035	.342	.322	1.000	.796	.210
EF	-.026	.048	.058	.056	.065	.040	-.021	.268	.224	.796	1.000	.271
AT	.055	.085	.032	.110	.109	.136	-.060	.075	.064	.210	.271	1.000

The strongest inter-subtest relationship is between Parts 2 and 3. Part 2 was also correlated relatively highly with the written and oral placement test scores as compared to the other subtests. The students at the C1 level have, for the most part, taken the placement test more recently than students at the other C-levels. It seems that placement procedure best predicts performance on grammar related tasks.

Part 2 was also most highly correlated with the teachers' subjective assessments of written and oral ability as a unit, though Part 4 was slightly better correlated with the teachers' subjective assessment of written work. Total test score had the strongest relationship with the teachers' assessments at this level, which it did not necessarily have at other levels.

Consistent with the trends in most other tests, the progress and effort scores correlate best with attendance.

The C1 test was subdivided into fourteen tasks for the correlational analysis. V1 - V5 were listening tasks, V6 - V8 grammar, V9 - V10 appropriacy, V11 - V13 were reading tasks, and V14 was a writing task.

Table 5.9.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
V1	1.000	.579	.290	.517	.387	.311	.271	.248	.253	.160	.114	.124	.273	.188
V2	.579	1.000	.323	.494	.305	.377	.240	.290	.239	.176	.087	.165	.176	.208
V3	.290	.323	1.000	.392	.205	.360	.326	.324	.264	.253	.073	.147	.227	.233
V4	.517	.494	.392	1.000	.484	.425	.359	.334	.291	.355	.129	.337	.380	.299
V5	.387	.305	.205	.484	1.000	.351	.153	.143	.155	.275	.171	.165	.341	.152
V6	.311	.377	.360	.425	.351	1.000	.549	.543	.439	.345	.245	.329	.206	.218
V7	.271	.240	.326	.359	.153	.549	1.000	.612	.447	.312	.252	.287	.355	.304
V8	.248	.290	.324	.334	.143	.543	.612	1.000	.436	.296	.183	.280	.192	.168
V9	.253	.239	.264	.291	.155	.439	.447	.436	1.000	.435	.133	.231	.226	.264
V10	.160	.176	.253	.355	.275	.345	.312	.296	.435	1.000	.189	.164	.205	.221
V11	.114	.087	.073	.129	.171	.245	.252	.183	.133	.189	1.000	.405	.209	.128
V12	.124	.165	.147	.337	.165	.329	.287	.280	.231	.164	.405	1.000	.294	.280
V13	.273	.176	.227	.380	.341	.206	.355	.192	.226	.205	.209	.294	1.000	.540
V14	.188	.208	.233	.299	.152	.218	.304	.168	.264	.221	.128	.280	.540	1.000

The listening tasks were generally more closely related to one another than the tasks in the other skills areas. The two telephone messages were highly correlated with each other, suggesting a task effect. Both were also highly correlated with V4, another telephone task though of a different kind. V4 was also highly correlated with V6, the cloze letter of application.

The two appropriacy tasks, V9 and V10, were quite highly correlated, as were the two dictionary reading tasks, V11 and V12. The pattern of tasks within a given subtest correlating best with other tasks in that subtest seems to be a constant one. However, other relationships are suggested between tasks that are not apparently related to the underlying skill of listening, or reading for example.

The mean intra and inter-task correlations are presented in Table 5.10.

Table 5.10.

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
IAT	.45	.43	.30	.47	.35	.55	.58	.58	.44	.44
IRT	.21	.22	.24	.32	.21	.33	.30	.26	.28	.25

	V11	V12	V13	V14
IAT	.25	.33	.35	.32
IRT	.16	.22	.27	.22

Again the pattern in the previous tests is repeated. It is interesting to note that none of the tasks or task-types has a mean correlation that is significantly

higher than any of the others. One might have expected that the cloze passages would correlate best with the other tasks, in part because there are generally more items in those tasks than in the others. Although they do correlate slightly higher at this level the difference is not really significant.

6.5. Concluding Comments

The correlational analyses in this section were of two types. Firstly, the progress subtests were correlated with a variety of other measures. The subtests were always moderately correlated with each other although the extent of the correlations indicated that the subtests were testing different abilities. This position was supported by the findings of initial item analysis. Teachers' assessments of written and oral ability were generally related to the progress test scores to some extent. It was important that this should be the case since it would have been unacceptable for the teachers' assessments to have been seriously at variance with test scores. The extent to which the teachers' assessments correlated with any of the subtests varied, and no clear pattern emerged. The assessments of progress and effort were generally not

very well correlated with other measures. They tended to relate most closely to the students' attendance in fact. This seems to be a reasonable finding. Progress and effort are extremely difficult to define as well as to assess, and no very clear guidelines were supplied to teachers (see Appendix 10 for an example of the guidelines). In addition, some teachers expressed a moral objection to assessing these two qualities and even sabotaged their registers. The results that emerged therefore, are not very reliable.

The second set of analyses focused on the relationships between the tasks in the tests. It was generally the case that tasks measuring the same skill or trait were more highly correlated with each other than they were with tasks measuring other skills. It was to be expected that there would be some exceptions which might be due to a number of possible causes such as the nature of a task or the text type involved. In addition, there were very few points available for some of the tasks and so minor variations in score could have an important effect on the correlation. Considering these difficulties the results were very clear. They supported Hypothesis Two which claimed that the tests were measuring different components of communicative competence.

7. Conclusion

In this chapter it has been demonstrated that it is possible to produce a reliable battery of task-based performance tests that have face and content validity thus satisfying Hypothesis 1. The correlations with teachers' assessments indicate that the results produced by the tests are not completely out of line with the views that teachers had of students' language ability, thus indicating that the tests have some concurrent validity with teachers' subjective assessments.

In addition, it seems that communicative competence, as measured by the tests in this battery, is divisible into components. The inter-subtest correlations were never high enough to suggest that any two subtests were measuring the same trait. Furthermore, when the tests were divided into tasks it was clearly the case that the relationships between tasks measuring the same trait were generally stronger than with those measuring different traits. The first stage in the construct validation of the battery was thus completed. It

supported **Hypothesis Two** which claimed that communicative competence can be divided into components, and that these components can be demonstrated to be statistically as well as intuitively distinct.

Chapter VI

1. Introduction

In Chapter 5, the items in each test were subjected to detailed item analysis. The systematic variation in the point-biserial correlations at all levels, between the whole test and the subtests indicated clearly that students did not perform uniformly in tests of different skills. It seemed that the variation might be on the level of skills such as listening, reading etc., microskills, tasks, and text types. The correlational analysis of the subtests showed that while the relationship between them was significant, it did not account for enough of the variance at any of the levels to suggest that they were so closely related as to be measuring the same trait. The results generated by the extensive item analysis carried out earlier allow the reasonable assumption that the imperfect degree of fit between the various subtests was not due to the unreliability of the subtests since they were all highly reliable, and most individual items were proven to function well as measured by traditional item analysis techniques. When the subtests were broken down by task, and subjected to correlational analysis, it

was generally the case that tasks attempting to measure the same trait were more closely related to each other than they were with tasks measuring different traits. It appeared that they were convergent (Tapp and Berkley, 1974; Stevenson, 1981; Clifford, 1981). All of these findings supported the view that the students' proficiency in English as measured by the battery of progress tests under discussion in this thesis, was composed of a number of distinguishable proficiency areas which, while clearly related, did not overlap sufficiently for it to be claimed that any of the components was dominant or indeed that a single ability could be said to underlie all the measures thus making some of them redundant.

It was decided, in light of the evidence presented above, to analyze the battery further in order to investigate the relationships between the subtests, tasks and individual items more closely. The most commonly used statistical technique that allows this to be done is factor analysis.

2. The Factor Analytic Technique

2.1. The Purpose of Factor Analysis

The beginnings of factor analysis are attributed to Spearman (1904) although a considerable amount of work on the approach took place over the next twenty years, the principal contributors including such famous names as Pearson, Burt, Thomson and Holzinger.

Factor analysis comprises a number of statistical techniques whose common objective is to represent a set of factors in terms of a smaller number of hypothetical variables. It is based on the fundamental assumption that some underlying factors, which are smaller in number than the number of observed variables, are responsible for the covariation among the observed variables (Kim, 1982). The analysis itself is not of course capable of attributing any identity to the factors that it produces. It is up to the researcher, by considering the variables loading on any factor, to decide, if he can, what these factors might represent. There is therefore a degree of interpretation involved that may or may not be agreed on by different

researchers. Ideally, groupings should be clear enough to make personal interpretation irrelevant although unfortunately, this is not always the case.

Factor analysis can be exploratory, in which case the researcher has no clear idea of how many underlying dimensions there are in a given set of data. In such cases, factor analysis is used as way of finding out the minimal number of hypothetical factors that may account for the observed covariance, and as a means of investigating the data for possible data reduction. This procedure is referred to as exploratory factor analysis and it is the most common type to be used in research in the social sciences.

Factor analysis can be confirmatory (Bock and Bargmann, 1966; Mulaik, 1972), in which case it tests the specific hypotheses of the researcher. Should the researcher feel that there are a specific number of factors accounting for a particular situation, and that certain variables will belong to certain factors he may choose confirmatory factor analysis over exploratory. It is generally the case, however, that an exploratory factor analysis is carried out in the first instance.

Harman (1976, p. 6) explains the difference between the two approaches:

> "... "exploratory" factor analysis may be useful in formulating theories in the behavioral and social sciences, but the "analytic tools" (including factor analysis) should not be confused with the science. As an exploratory tool (among others), factor analysis can be used to verify or modify theories through new experiments and new data subjected to fresh analyses for the purposes of clarifying or polishing previous formulations. By contrast, "confirmatory" factor analysis may be used to check or test a preconceived or given hypothesis about the structure of empirical data."

The data investigated in this thesis was subjected to exploratory factor analysis despite strong arguments in favour of confirmatory factor analysis put forward by Bachman and Palmer (1983), Vollmer and Sang (1983), and Sang et al. (1986). It was felt that at this stage of the investigation there was not anything to confirm while there were many areas in need of exploration particularly as previous studies in the field had focused almost exclusively on the analysis of the relationships between fairly conventional tests. The battery under discussion in this thesis was different, in the sense that subtests were performance-based to a large extent, and the relationships between items themselves as well as those between tasks and subtests

were investigated. This represents what might be termed a 'hierarchical' approach to construct validation as opposed to the more conventional 'horizontal' one.

2.2. Differing Opinions on Factor Analysis

Factor analysis is a contentious area and it has critics as well as supporters. Lawley and Maxwell (1971, p.38) hold the view that factor analysis as a model is only useful as an approximation of reality and should not be taken too seriously. Hills (1977, p. 340) holds that factor analysis is not worth the time necessary to understand it and carry it out. And Chatfield and Collins (1980) claim that factor analysis may not be worth using in any but a few very specific applications.

On the other hand, Cattell (1971, p.24) is of the opinion that factor analysis has influenced theory in countless areas. And, in a similar vein, Royce (1973, p. 1) holds that there is especially great potential in the combination of multivariate research and theory construction.

The literature related to factor analysis is full of contradictory opinions concerning the technique itself and the details of its application. Despite these, it has been widely used in research in the social and behavioural sciences and it was felt that it was the most appropriate method of investigating the data discussed in this thesis. The results obtained are a strong confirmation of that view.

3. Initial Factor Extraction

Factor analysis involves two main stages. Firstly a correlation matrix is created. It contains all the variables to be analyzed. This matrix is the starting point for the factor analysis and the so-called initial factors are extracted at this stage. Secondly, the initial factor matrix is rotated in order to allow for an easier interpretation of the results.

Several pieces of factor analytic research in Language Testing used Principal Components Analysis (PCA) to extract the initial factors, and some did not rotate the initial factor solution. PCA is generally

considered cruder than Principal Axis Factoring (PAF) (Cattell, 1978; Rummel, 1979). This is largely because PCA sets the diagonal values in the correlation matrix at unity which means that all of the test variance used to obtain the correlations is entered into the analysis (Farhady, 1983). Using PCA, therefore:

> "... common and unique variance is mixed in an inextricable way that obscures the view of what the variables have in common with each other." (Comrey, 1973, p. 98)

PFA proceeds in much the same way as PCA except for one important difference. The diagonals on the correlation matrix are replaced by estimates of the communality of the variable. This means that only the proportion of the variance that is explained is used in factor extraction thus making the extraction more sensitive.

There are numerous methods of exploratory factor analysis including the Generalized Least Squares Method, the Maximum-Likelihood Method, and the Alpha Method. These will not be discussed in detail here. Suffice it to say, experts agree (Harman, 1976; Comrey, 1973; Farhady, 1983) that Principal Axis Factoring is the most applicable to the type of data analyzed in this thesis. Therefore, PAF was used for all initial factor extraction.

4. Some Considerations to Bear in Mind

Factor analysis is a volatile and complex tool that is sensitive to minor variations in the definitions of the parameters. Such variations are capable of having a major influence on the results obtained by the analysis. In the following sections some of these variations are discussed in detail.

4.1 Decision Making Criteria for Factor Analysis

One of the major uncertainties of factor analysis is deciding when one should stop extracting factors. Rummel (1979, p.350) divides methods for making this decision into three groups. The first depends on inferential criteria that require a variety of statistical tests. The second involves mathematical criteria such as Guttman's Bounds (Guttman, 1954) and Harris Scaling (Harris, 1962). The third method employs rules of thumb that have been developed on the basis of experience rather than mathematical theorums. They tend to yield results that are quite acceptable but much

easier to arrive at than by either inferential or mathematical methods

4.1.1. The Scree Test

One of the best known rules of thumb is the "scree test" which was first proposed by Cattell (1966). By this method the factors are plotted against the proportion of the variance that they extract. The resultant curve will have a negative slope. When what might be classified as random error or trivial factors begin to appear the curve levels off. All factors above the levelling off point may be considered as relevant, while those below it should be excluded from the analysis. Nowadays, the scree plot is a default setting in most sophisticated statistical packages. Unfortunately, it is often difficult to determine exactly where the curve levels off, particularly where there are a large number of initial factors, since there is a tendency for the levelling off to take place in stages. Therefore, a certain amount of trial and error may be required before a final decision can be made as to how many factors are actually significant.

4.1.2. Using Eigenvalues and Factor Loadings

Another commonly used method of deciding on which factors are relevant, is to refer to their eigenvalues. The eigenvalue of a factor is:

> "... the sum of the squared loadings of input measures explained by that factor. It is therefore an index of the relative importance of the factor in explaining the total variance of all the variables. If a given factor were a perfect explanatory variable, it would have an eigenvalue equal to the number of variables." (Farhady, 1983)

It is generally agreed that factors with eigenvalues greater than one are probably significant, although the exact value is dependant also on the loadings of the variables on given factors and the method of extraction used. If the variable loadings are at 0.3 or above the factor may also be significant. On the other hand, it is possible for an eigenvalue to be greater than one while all the loadings are below 0.3. In such cases the factor would probably not be considered significant and would be excluded from the analysis. In this thesis minimum variable loadings were never set lower than 0.3 though they were sometimes set higher.

4.1.3. The 1/P Method

Woods et al. (1986, p.285) suggest a method put forward by Eastman and Krzanowski (1982) for determining the number of relevant factors. By this method:

> "... if the original data has 'P' dimensions, assume that components which account for less than a fraction 1/P of the total variance should be discarded."

This method was applied as a check on the relevance of factors. It never suggested that there were less relevant factors than the number extracted by other methods.

4.1.4. Concluding Comments

In many cases it may be proper to allow the factor analysis to run without limiting the number of factors that are extracted. However, if there are a large number of variables it may be necessary to specify the number of factors to be extracted since many of them will be insignificant or impossible to interpret. In such cases the rules of thumb outlined above are the most straightforward and convenient way of doing this.

It should be noted that once a cut off point has been set, either by specifying a particular eigenvalue or predetermining the number of factors to be extracted, any factor with an eigenvalue less than the required amount will be excluded from the analysis because the extraction process will terminate at that point. For this reason Cattell (1952), amongst others, recommends overfactoring as opposed to underfactoring. However, this view is not shared by Kaiser (1963). It seems safest therefore to carry out some initial analyses and, on the basis of these, decide how many factors should ultimately be extracted. Factors where the variables have very low loadings, or are uninterpretable should then be excluded from further analysis. While the percentage of variance attributable to each factor is of interest, the factor structure can be considered to be of greater significance. In this thesis factors with an eigenvalue greater than one were generally considered significant and discussed. Those with an eigenvalue of less than one were excluded unless the scree plot indicated that they may be of interest.

4.2 What is the most appropriate method of rotation?

4.2.1. Why rotate?

The first step in factor analysis is to extract initial factors. This can be done using a number of techniques as mentioned in the previous section. While certain research studies can be designed so that the factors are interpretable without rotation (Gorsuch, 1974) it is generally the case that regardless of the initial technique used to establish factor structures, the second obligatory step in the process is to rotate these structures. Rotation is employed in order to make the interpretation of the factor analysis simpler and psychologically more meaningful. However, Kim et al. (1982) remind us that:

> "... no method of rotation improves the degree of fit between the data and the factor structure. Any rotated factor solution explains exactly as much covariation in the data as the initial solution. What is attempted through rotation is a possible "simplification". There exist different criteria of simplicity which lead to different methods of rotation." (p. 50)

By rotating the initial factor solution therefore, it is hoped that each variable will load primarily on one factor with each factor accounting for a maximum of

variance generated by the variables that load on it (Farhady, 1983).

4.2.2. Orthogonal or oblique rotation?

Rotational methods fall into two main categories - orthogonal and oblique. Orthogonal rotation begins with the assumption that the factors extracted are independent of each other or uncorrelated. Oblique rotation on the other hand does not make the assumption of independence, assuming rather that the factors are related in some way, that they are correlated in other words. Opinion as to which method of rotation to use differs widely with Cattell (1978) strongly in favour of the oblique method holding that if orthogonal axes are imposed on factors that are oblique, the meaning of the factors is distorted, and the result is a "mixture of true factors" (Cattell, 1978 p.128). Guildford (1973) takes an alternative stance to Cattell. He does not question the idea of correlations existing among factors but objects to the arbitrary nature of some of the methodology associated with oblique rotation. Guildford's view is supported by Eysenck (1977). Hinofotis (1983) also considers this question of

rotational method in some detail. She makes the following point:

> "... in dealing with natural language data with a focus on the communication process, it is not clear that orthogonal factors will best reflect the relationship among the variables. While it might be possible to separate the factors involved in communication on a conceptual level, in the actual communication process the factors are highly integrated. Thus it is feasible that correlated rather than uncorrelated factors will prove more meaningful when working with language data."

In practice there is often little to choose between these two basic methods of rotation. Hinofotis (1983) used both with the data she was examining and achieved similar results. Some of the data discussed in this thesis was also subjected to oblique as well as orthogonal rotation. The results were similar, though the orthogonal method was easier to interpret. Nunnally (1978, p. 376) supports the view that rotational method may not affect the results of a factor analysis when he writes:

> "...the two approaches lead essentially to the same conclusions about the number and kinds of factors inherent in a particular matrix of correlations."

Clearly experts' opinions differ on the choice of rotational method. Nunnally (1978), Guildford (1973) and Farhady (1983) for example support the orthogonal

approach. Cattell (1978) favours oblique rotation. Some studies have been carried out to compare the results when different approaches are used (Dielman et al., 1972; Hinofotis, 1983). They conclude that there is often little difference in end results whatever method of rotation is used. Orthogonal solutions are however, easier to interpret and this is perhaps the most significant distinguishing feature of the two methods for all practical purposes. When both methods were used with the data under discussion in this thesis the orthogonal solutions were clearer and easier to interpret than the oblique ones. This being the case, the orthogonal method was preferred.

4.2.3. Methods of orthogonal rotation

There are several methods of orthogonal rotation. Quartimax emphasizes simple interpretation of variables which means that the solution tends to minimize the number of factors needed to explain a variable. This method of rotation often leads to the appearance of a general factor of some sort. Varimax attempts to minimize the number of variables with high loadings on a given factor which should make it easier to interpret the factors. Equamax is a combination of varimax, which

simplifies the factors, and quartimax, which simplifies the variables. The most commonly used and most highly recommended of these methods is varimax (Gorsuch, 1970; Farhady, 1983). Varimax was therefore used to rotate the initial factor structure of the analyses discussed in this thesis.

4.3 How to deal with dichotomous variables

An assumption underlying factor analysis is that the variables will be on a continuous scale. Much of the analysis in this thesis involved continuous scales thus posing no methodological difficulties on this point. However, it was also decided to analyze tests and subtests in a slightly unusual way by treating individual items as unique variables in order to establish whether they would group according to skills, microskills or tasks. The variables then become dichotomous as opposed to continuous. Experts (Cattell, 1978; Harman, 1976; Gorsuch, 1974; Rummel, 1979) warn that this can lead to difficulty factors emerging. In other words there is the possibility that variables will cluster together on the basis of level alone regardless of their content.

The literature on how to deal with this situation is contradictory as it is with many other aspects of factor analysis. Cattell (1978) discusses the use of the "phi-over-phi-max" coefficient in preference to the Pearson Product-Moment to generate the correlation matrix. However, he points out that it tends to distort communalities and thus may not be suitable. Harman (1976) discusses the use of tetrachoric correlations recommended by Carroll (1961) but concludes that while it may be relatively easy to calculate them, there is no assurance that the matrix will be consistent. "In other words, a matrix of tetrachoric correlations may not be proper for factor analysis." (Harman, 1976, p.24). Comrey and Levonian (1958) argue that the phi coefficient is the most suitable method to use.

Holley and Guildford (1964) suggest the use of what they call the G-coefficient as being both easier and more appropriate for dealing with dichotomous data than any of the methods outlined above. By this method an extended score matrix is created. For each of the original subjects, a new subject is created whose scores are the exact opposite of the original subject. If Subject One has a score of +1 on a given variable, then his mirror image has a score of -1 for the same

variable. The effect of this is that each variable effectively has a 50/50 split. G is thus independent of the way in which the item is originally scored. In theory therefore the factor analysis of the resultant correlation matrix should not be affected by item difficulty.

So we see that there are at least four approaches to the problem of dealing with dichotomous data. Each has its adherents and critics. It was decided therefore to ignore the arguments in favour of any of these approaches in the first instance and see whether the resultant analyses appeared to be affected by difficulty factors. If they were, then it was decided to employ the G-coefficient in preference to any of the other approaches since it was the one that received least criticism.

5. Some Factor Analytic studies in EFL/ESL

Factor analysis became a popular tool in language testing research during the seventies due principally to the work of Oller in the first instance. He and various co-workers carried out a number of studies one

of the main aims of which was to try and establish whether it was possible to statistically isolate a general factor that could be said to underlie language competence.

Oller and Hinofotis (1980) discuss what they describe as two mutually exclusive hypotheses about second language ability. Hypothesis 1 claims:

> "... that language skill is separable into components related either to linguistically defined categories (e.g. phonology, syntax, and lexicon) or the traditionally recognized skills (i.e. listening, speaking, reading and writing."

On the other hand, Hypothesis 2 proposes that:

> "... second language ability may be a more unitary factor such that once the common variance on a variety of language tasks is explained, essentially no meaningful unique variance attributable to separate components will remain."

Oller and Hinofotis gave seven subtests to two groups of students. The first group were all Iranian while the second came from a variety of language backgrounds. In addition, the second group also took an oral test. While Oller and Hinofotis argue that the results from the first group supported Hypothesis 2 quite clearly, they concede that the results from the second did not.

There seems to be a clear indication from the factor analysis (a principal components analysis with a varimax rotation) that at least two major factors are present - the first related to oral proficiency, and the second to the other measures. Oller and Hinofotis' attempts to downplay the finding that there is more than one factor present are less than convincing (Vollmer and Sang, 1983).

A third analysis was carried out on a relatively small group of subjects (51). It supports still more clearly the existence of more than one underlying factor accounting for the variance in test performance. In this case three factors emerge. The first groups most of the non-oral subtests together. The second the oral subtests, and the third a mix of non-oral subtests. Again Oller and Hinofotis downplay the importance of this finding, though they do concede that there is:

> "... some evidence to suggest that (excluding the oral interview data) if the data represent the whole range of subject variability, the unitary competence hypothesis may be the best explanation, but if the variability is somewhat less, a moderate version of a separate skills hypothesis would be preferred."

It is clear from this research that the original claim that a single factor underlies language proficiency is no longer feasible.

This view is supported by the work of Farhady (1983), Vollmer (1981), and Abu-Sayf et al., (1979). Farhady criticizes early work by Oller and others for using Principal Components Analysis (PCA), and not rotating the factor matrix. He argues convincingly against the use of PCA and for Principal Axis Factoring (PAF) a view that is shared by most experts on the factor analysis technique (Harman, 1976; Comrey, 1973; Kim et al., 1981). Farhady also devotes some time to a useful discussion on the need for rotation of the initial factor matrix. He concludes:

> "If ... one uses incomplete methods, it will appear, in study after study, that the first factor, whatever it may be called, is the only factor underlying all the variables. Therefore, previous interpretations of unrotated factor matrices are called into question and further investigation is required to determine the actual composition of language proficiency."

Vollmer and Sang (1983) carry out an extensive review of factor analytic studies in the areas of language aptitude (Carroll, 1958; Pimsleur et al., 1962; Gardner and Lambert, 1965), and the nature of language proficiency (Lofgren, 1969; Carroll, 1975; Steltmann,

1979; Bonheim et al., 1979; Hosley and Meridith, 1979). They conclude that there is little evidence of a clear unifying purpose underlying these studies. Each one pursued very specific questions which meant that a vast range of variables were investigated, on a number of different populations. While most of these studies imply that language competence is divisible rather than unitary, Vollmer and Sang consider that the interpretations of the various factor analyses were often either too narrow and one-sided or made claims that were too far reaching for the results obtained. With regard to studies attempting to prove the psychological reality of a unitary competence hypothesis, Vollmer and Sang conclude, with strong evidence to support them, that the results obtained in many studies were a direct result of the statistical techniques used rather than a reflection of the nature language ability.

Bachman and Palmer (1983) moved in a different direction from earlier research. In their attempt to measure the construct validity of the FSI oral interview, they adopted the classic multitrait-multimethod matrix first used by Campbell and Fiske (1959) and then applied confirmatory factor analysis to the data. In this way they were able to identify and

quantify the effect that method and trait may be having on language proficiency. They were able to provide strong evidence to refute the unitary competence theory and the complete divisibility theory.

Bachman and Palmer conclude that method effect needs to be taken into account in validation studies and suggest that the multimethod-multitrait model be used. They also suggest that confirmatory factor analysis is more appropriate than exploratory factor analysis.

Sang et al. (1986) agree that confirmatory factor analysis is more appropriate than exploratory. In their study they attempt to confirm a model of language proficiency on three levels (elementary, complex and communicative). While they are able to confirm this multiple factor model Sang et al. express some caution as to the generalizability of their findings. In addition, they state that:

> "... the structure of L2 competence cannot be seen as independent, either of cognitive prerequisites on the side of the learner or of the teaching strategy adopted by a particular foreign language teacher."

The final study to be mentioned in this section is one carried out by Lee (1985). It differs from the other studies discussed above in the sense that it approaches the validation question from a micro level as opposed to a macro level. He subjects a number of cloze passages to principal components analysis treating each item as a subtest in its own right in order to ascertain whether the characteristics of individual items vary thus bringing into question the validity of global scoring methods. He found that the nature of items did have an effect. While it appeared that the passages measured some sort of overall language ability, it also seemed that:

> "... there are possibly two underlying language abilities being measured, corresponding to an "openness" versus "closedness" opposition. Indeed, it may be pointed out that the "openness" versus "closedness" contrast may be a behavioural manifestation of the general underlying opposition between the "paradigmatic" and the "syntagmatic" relation in general linguistic theory."

Despite important differences in methodology, most of the studies discussed above have in common the feature that the language tests used are rather traditional in nature. They are for the most part contrived, indirect measures of linguistic proficiency, rarely adopting a performance-based approach. Many of the measures come

from the TOEFL battery or something similar. In addition, no attempt is made, with the exception of Lee (1985) and Bachman (1982) to some extent, to validate tests at a micro level. While this is partly due to the nature of the measures used, it is nonetheless a weakness. The tests discussed in this thesis differ from those in earlier studies in the sense that they are performance-based to a large extent, and as such it is appropriate to analyze them not only on the macro level of subtests but also on the micro level of tasks and microskills. Since this had not been attempted before with communicative performance-based tests there were no findings upon which to base expectation. This was the main reason for adopting an exploratory approach as opposed to a confirmatory one.

6. A Factor Analysis of the Tasks in the Test Battery

The factor analysis of the battery was divided into two main parts. Each test was first split into tasks in the same way as it had been in Chapter 5 for the correlational analyses, with each task acting as a subtest in its own right. Factor analyses were carried out in order to establish whether the tasks would group together into the broad skills areas that the

tests were divided into - that is, listening, grammar, appropriacy, reading and writing. If they did then this could be considered confirmation of the fact that the tasks were linked together by an underlying trait, and that the division of the test into these skills areas could be justified on a statistical as well as intuitive level.

Each test was then subjected to further factor analysis but in this instance each item was treated as a subtest in its own right. If the items grouped according to skill, then it could be further confirmed that the areas being tested were indeed somewhat distinct. It was considered possible however, that items would group by task, or even microskill. In addition, an item-by-item factor analysis was also carried out on the Listening section of each test. The Listening section could be considered to be the most performance/skills-based of the four subtests which is why it was selected for further factor analysis. If factors related to microskills were to emerge, then Hypothesis Three, which claimed that they could be statistically as well as intuitively isolated, would be confirmed.

In the first instance the eigenvalues of the analyses were left at the default setting of one. If this value appeared not to do justice to the data, then it adjusted accordingly, either up or down. The variables were generally sorted so that they appeared in descending order of importance under their strongest factor and all loadings of less than 0.3 or 0.4 in some cases, were left blank. Therefore, if a variable had a loading of less than 0.3 for any factor there would simply be a blank space. This layout allowed for a much clearer and easier interpretation of the results. It was sometimes the case that a variable did not load at 0.3 on any of the factors. In such cases there was no entry at all for that variable. However, since it grouped with the variables that it was most closely associated with this was some indication of its relationship with other variables.

6.1. The A3 Progress Test

The factor analysis of the tasks in the A3 progress test clearly revealed four principal factors with an eigenvalue greater than one which accounted for 61% of the variance. The cut off point for significant loadings of variables on given factors was set at 0.4.

The table below illustrates the results of the analysis.

Table 6.1.

	F 1	F 2	F 3	F 4
V4	.74541			
V1	.70287			
V5	.64940			
V2	.64851			
V6	.63182			
V3	.62097			
V7		.75096		
V10		.69953		
V8		.63289		
V9		.51338		
V11				
V17			.86169	
V15			.65637	
V18			.59547	
V16		.41049	.52075	
V14			.44136	
V12				.75125
V13		.41395		.54720

V1 - V6 Listening V7 - V10 Grammar
V11 - V13 Appropriacy V14 - V18 Reading/Writing

It will be noted that two of the variables feature in more than one of the factors. However, it is often the case that the inter-factor boundaries are not absolutely clear cut. Overall, the result demonstrates beyond doubt that the variables group together in the way it was anticipated they should. This supported

Hypothesis Two which claimed that the skills as measured by these tasks are empirically distinguishable from each other.

6.2. The B1 Progress Test

The factor analysis of the tasks in the B1 progress test did not produce results that were as clear cut as those demonstrated by the A3 progress test. Six main factors emerged accounting for 72% of the variance and the minimum loading for any variable was set at 0.4. While there was a definite tendency towards grouping according to skills it also appeared that the nature of the tasks themselves may be playing some sort of role. For example, V8 and V9 are both grammar tasks, yet they do not feature as components of the same factor with V8 standing alone as factor 6. It should be noted that the task in V8 employs a bank of words above the passage which are used to fill the blanks, while in V9 the students are given no help in deciding which words to use. While both tasks are purportedly measuring the same trait - knowledge of language systems - they do so in a different way. It appears that this difference has a significant effect. In other words, the nature of the task is of some importance. On the other hand, Factor 2

groups V9 and V10 together. V10 is supposed to be testing appropriacy in a multiple-choice format. The format of V9 is cloze. Clearly, the superficial nature of the tasks is of quite a different sort and yet they are grouped together. In this case the nature of the task does not affect the way in which variables group. It may be that the multiple-choice appropriacy task is not really measuring anything very different from the cloze. This being the case, doubts are cast as to exactly what either of the tasks is measuring.

Table 6.2.

	F 1	F 2	F 3	F 4	F 5	F 6
V6	.81883					
V7	.62349					
V3	.46218			.40318		
V4	.41512			.40496		
V10		.89456				
V9		.85293				
V14			.70796			
V15			.66096			
V13			.53193			
V2				.66508		
V5				.56092		
V1				.53193		
V11					.66829	
V12					.56865	
V8						.64778

Listening	V1 - V7	Grammar	V8 - V9
Appropriacy	V10 - V12	Reading/Writing	V13 - V15

The listening tasks fall into two groups, though there is significant overlap. The dictation task, V7, which is wholly verbal and at the discourse level may require a more complex form of language processing. It does not overlap at all with the more numerical tasks represented by V1, V2 and V5. Conversely, V3 and V4 which require the students to write down weather words and dates feature in both factor 1 and 4. It seems likely that some sort of listening factor is linking these two categories of task yet it also appears that the nature of the language processing required to complete them affects the way that they group together. There seems to be a distinction between verbal and numerical processing skills.

6.3. The B2 Progress Test

When the eigenvalue for this analysis was at one, only two factors were extracted. However, on the evidence of the scree plot there should have been at least four significant factors. The minimum eigenvalue was therefore reset at 0.8 with the result that five factors emerged. Since they were all interpretable,

this analysis was retained. The loading of variables was not quite as high across the board as with the previous tests and the minimum level was set at 0.3.

Table 6.3.

	F 1	F 2	F 3	F 4	F 5
V3	.88683				
V4	.56091				
V1	.39922		.30301		
V2	.37420		.31119		
V5		.78267			
V6		.48526			
V7		.34203			
V10			.69087		
V11			.35379		
V9				.79673	
V8					
V12					.82156

Listening	V1 - V4	Grammar	V5 - V6
Appropriacy	V7 - V9	Reading/Writing	V10 - V12

It will be noted that the breakdown of the tasks fits into the five skills fairly neatly, with a writing skill emerging for the first time. The four listening tasks all group together but it is interesting to note that the two information retrieval items also load on factor 3. All of these tasks involve the extraction of specific information the difference between them being

that V1 and V2 are listening tasks and V10 and V11 are reading. It may be that there is an underlying skill that links these four variables and is more powerful than the obvious difference in medium. It may also be that the nature of the four tasks is similar in some way. This does not seem very likely however, since on a superficial level they are quite distinct.

V7, the multiple-choice appropriacy task again groups with the grammar factor as opposed to the other appropriacy tasks (V8 and V9). This confirms the suspicion regarding the nature of this task that arose in the B1 factor analysis. Factor 4 is an appropriacy factor, with V9 loading very heavily. V8 has a loading of less than 0.3 and so no entry is made. However, it is most closely allied to V9 according to the grouping. V12, the letter writing task is clearly a factor in its own right, a finding which seems quite reasonable.

6.4. The B3 Progress Test

The results of this factor analysis were slightly less clear cut than the previous ones. With the eigenvalue set at one, three factors appeared. However, the scree

plot indicated that there were at least four meaningful factors. The minimum eigenvalue was therefore reduced to 0.8 in order to take this into account. The four factors that then emerged accounted for 63% of the variance. The cut off point for variable loadings was set at 0.4.

Table 6.4.

	F 1	F 2	F 3	F 4
V1	.77608			
V2	.71176			
V4	.63479			
V3	.57885			
V12	.45537			
V15		.64029		
V6		.61941		
V14		.58603		
V7		.45412		
V10		.40283		
V5				
V11			.64024	
V8			.55245	
V9		.47730	.49503	
V13				.77282

V1 - V7 Listening V8 - V10 Grammar
V11 - V13 Appropriacy V14 - V15 Reading/Writing

Factor 1 is clearly a listening factor although, it is interesting that the second of the appropriacy tasks should group so strongly here. This appropriacy task is in fact a simulated telephone conversation, and it may

be that this similarity is the reason for V12's presence in factor 1. A familiarity with appropriate telephone technique may be an influential underlying skill. The other listening tasks group together under factor 2.

Factor 2 is a mixture of tasks coming from most sections of the test. It is the first indication that a general factor of any sort may underlie the skills tested in this battery. There appears to be no readily explainable cause for the grouping in factor 2.

Factor 3 groups two of the grammar tasks with the multiple-choice appropriacy task again. This happened in both of the earlier B-level tests. It is clearly very questionable that tasks like V11 test appropriacy. This being the case they should either be excluded from the battery in the future or renamed. Since the washback effect of multiple-choice tests is questionable perhaps the best course of action to take with this type of item is to exclude it from the tests in future.

Factor 4 contains only the third of the appropriacy tasks - V13. Overall, so far in the analyses, there seems to be adequate justification for the retention of this type of item in the test battery since it generally appears as a factor in its own right. However, it seems that the item type may be slightly volatile in the sense that it can be affected by a knowledge of the world as seems to be the case in this test where V12 groups with the telephone listening tasks as opposed to with V13, the other open-ended appropriacy task.

6.5. The C1 Progress Test

Four factors emerged when the eigenvalue was on the default setting of one. However, the scree plot indicated that there were five factors of significance, accounting for 68% of the variance. The minimum eigenvalue was therefore set at 0.9 to allow for the extraction of these five factors.

Table 6.5.

	F 1	F 2	F 3	F 4	F 5
V8	.73815				
V7	.72873				
V6	.58760				
V9	.51736				
V3					
V1		.72273			
V2		.67621			
V4		.64485			
V5		.51876			
V13			.79442		
V14			.58932		
V12				.59565	
V11				.57867	
V10					.76000

Listening V1 - V5 Grammar V6 - V8
Appropriacy V9 - V10 Reading/Writing V11 - V14

Factor loadings of less than 0.4 were omitted from the table above. However, it should be noted that there were four loadings of between 0.3 and 0.4. V6, the grammar task involving a letter of application, also loaded on factor 2, which was primarily a listening factor, and V3 a listening task relating to a job interview, loaded on both factors 1 and 2. V9, an appropriacy task also concerning a job interview loaded more heavily on factor 1 than it did on factor 5, which contained the second appropriacy task. Factor 1

therefore, which seems to be predominantly grammar related, also seemed to attract tasks where the topic was related to job application procedures in some way, either interviews or letters of application. It may be the case that there is such a thing as a topic effect related to the candidates' knowledge of the world.

Factor 4 groups together the two tasks related to the use of the dictionary. Although they are classified as reading tasks they do not feature in factor 3 which includes V13, a more traditional reading comprehension task. The skills required by V11 and V12 are on a much more basic level - checking spelling, deciphering illegible words in a text, and alphabetical ordering. Perhaps the complexity of the task is having an effect here.

Factor 3 groups together the reading and the writing tasks. This happens with a number of the tests and it is not clear exactly why tasks which are essentially opposite in nature in the sense that one is productive, and the other largely receptive, should be grouped together.

6.6. The C2 Progress Test

With the eigenvalue set at one, three factors emerged. However, the scree plot again indicated that there were in fact four significant factors and so the minimum eigenvalue was reduced to 0.8 to take this into account. The four factors accounted for 65% of the variance. In order to make the interpretation of the factors more straightforward, the minimum loading for the significance of a variable was set at 0.4.

Table 6.6.

	F 1	F 2	F 3	F 4
V11	.78938			
V12	.75462			
V6	.41102			
V13				
V1		.73871		
V2		.69158		
V3		.49112		
V8			.69926	
V7			.66455	
V4			.46186	.43373
V9				.64133
V10				.43037
V5				

V1 - V6 Listening V7 - V8 Grammar
V9 - V10 Appropriacy V11 - V13 Reading

The four factors represent fairly closely the four skills that, it is hypothesized, are being tested. Factor 1 is a reading factor although it is interesting that one of the listening variables (V6), concerned with understanding the news, loads most heavily on this factor. However, it also loads at above 0.3 on factors 2 and 3. The other news listening variable (V5) also loads at above 0.3 on three factors. However, these are factors 2, 3 and 4 as opposed to factors 1, 2 and 3. The spread of loadings on the news variables could be due to a number of reasons. Firstly, the nature of the task itself is fairly complex and requires not only specific understanding of detail but the ability to recall information with some precision and comprehend the questions. Secondly, listening to and understanding the news depends to some extent on one's knowledge of the world. Although an attempt was made to keep the topics to ones of local relevance it may well be that performance was affected by background knowledge. It seems likely that tasks which are broader in scope, those requiring some knowledge of the world as well as basic language processing skills, may be of a more general nature. The dictation task for example, also loaded at above 0.3 on three of the factors. This was a fairly complex activity that required the students to take down a job advertisement. While it may also be possible that knowledge of the real world is of

importance with this task, there is the further possibility that this type of task is one that requires a general language processing skill.

6.7. The C3 Progress Test

With the eigenvalue set at one three factors were extracted however, the scree plot indicated that additional factors may also be significant, and so the minimum eigenvalue was set at 0.8, which resulted in the extraction of five principal factors that accounted for 71% of the variance. The minimum factor loading was set at 0.3.

There seems to be a general trend that the overlap between factors is greater as the English proficiency level of the students increases. There was minimal overlap in the A3 test while in the C3 test there is a significant amount. Inspite of this at least three skills emerge, those of grammar, listening and reading/writing. Appropriacy, while loading on two factors, does not feature as a factor in its own right, as it had done in all of the earlier progress tests.

Table 6.7.

	F 1	F 2	F 3	F 4	F 5
V8	.75597				
V9	.69591				
V7	.49697			.45086	
V6	.34318		.31793		
V12		.66666			
V11		.56740		.39280	
V13		.55843			
V10		.37065	.33933		
V1			.75398		
V2			.64506		
V3			.32961	.79452	
V4					.78639
V5	.33828			.32058	.38894

Listening	V1 - V6	Grammar	V7 - V9
Appropriacy	V10	Reading/Writing	V11 - V13

The three grammar tasks group together to form the first factor although V6 the listening note-taking task appears in this factor as well. However, it also features in factor 3, which is a listening factor. Interestingly, the first grammar task, V7, also loads quite heavily on factor 4. The main variable in this factor is V3 a listening task that involves the editing and correction of a job advertisement. The grammar task, V7, is a cloze also based on a job advertisement.

It seems that the nature of the text or subject matter is the main feature that links these two variables.

Factor 2 includes the reading (V11-12) and writing (V13) tasks, along with appropriacy (V10). Reading and writing have been grouped together before. In some ways this is slightly strange, since one of them is a productive skill, and the other receptive.

Factor 3 seems to be essentially listening, with four of the six listening tasks loading significantly on this factor. V4 and V5 which involve listening to the news form factor 5 although V5 also loads on two other factors.

6.8. Concluding Comments

It seems reasonable that there should be a certain amount of overlap between factors, since the tests are all language based. The extent to which the tasks tend to group together according to general skills areas or traits is nonetheless striking. In only one of the progress tests (B3) does there appear to be any

significant grouping of tasks of a strikingly different nature under one factor.

When the tasks grouped together in unpredictable ways, explanations were generally possible for the groupings which, although guesswork to a large extent, seemed somewhat logical and provided useful insights into the nature of the tests. It appeared that the nature of the tasks themselves could have had some role to play in factor grouping or the lack of it. In the B1 progress test for example, the two grammar tasks although fairly similar in format, were not grouped together. Perhaps this was because the nature of the tasks was slightly different, or that one was less demanding and complex than the other. On the other hand, in a performance-based battery of this kind, it would not be true to say that task-type or method, was a major underlying cause of grouping as has been suggested in some studies involving more conventional test formats (Bachman and Palmer, 1983). The listening tasks in several of the progress tests grouped together in ways that could only be trait related as opposed to method dictated. In certain instances, particularly at the B-levels when multiple-choice appropriacy tasks grouped with cloze, neither method nor obvious trait effect appeared to be

responsible for the grouping, thus indicating that less readily explainable causes underlie some factors.

It also appeared that the complexity of a task in terms of language processing may be responsible for factor grouping in certain instances. Such a view is supported by the work of Sang et al. (1986). For example, in the B1 progress test there seemed to be a clear distinction between verbal or discourse processing and numerical processing. At the C1 level the fact that the reading tasks did not group together may also have been due, in part, to the complexity of the language processing skills required.

It is interesting that the writing skill rarely appeared as a separate factor. At the A3 and C2 levels this was to be expected since the tests did not include a clearly distinguishable writing task. At the other levels, while there was generally a degree of integration between reading and writing, the writing tended to group with reading. It was only at the B2 level that a clear separate writing factor emerged.

Another interesting finding was that similar microskills sometimes seemed to be responsible for factor grouping. For example, at the B2 level, the microskill involving the extraction of specific information appeared to represent the only obvious clue to the relationship between the reading and listening tasks that grouped together to form factor 3.

It also appeared that knowledge of the world, or familiarity with the subject of a task in real life, may have had an influence. At the B3 level it seemed that familiarity with telephone techniques may have contributed to V12 grouping with the other variables in factor 1. In a similar vein, at the C levels tasks which involved listening to the news were far more difficult to pin down than other listening tasks. The requirement they made for a level of general knowledge may have had some bearing on this.

No clear answers to the problems posed in the paragraphs above are available. It seems likely, however, that a number of different causes account for the relationships that emerged between tasks. For example, a knowledge of the world may sometimes be a factor in language processing. The nature of the task

and possibly text may also have an effect. The extent to which a task is cognitively demanding seems to have an influence in the way that it groups with other variables. However, the strongest influence for the most part appears to be the main skill or trait that is being tested.

7. The factor analysis of items

In the following sections the results of the factor analysis by item will be discussed. The tables for the Whole Test analysis along with some discussion are presented in Appendix 12 while the Listening Test factor analyses are discussed in detail here.

Before beginning the discussion however, several points need to be made.

When factor analysis is applied to a large number of variables it may be that the eigenvalue one criterion will not apply in the same way as it does with a smaller number of variables (Cureton, 1983). When the whole tests were analyzed, each containing between

eighty nine and one hundred and twelve variables as many as thirty five factors with an eigenvalue greater than one emerged. Since most of these factors were uninterpretable, all loadings falling below 0.3, it was decided to use the scree plot as the primary indicator of the number of factors that should be extracted. It was not always absolutely clear where the plot leveled out but it seemed to be at a point equivalent to between six and ten factors with all of the tests.

The Listening Tests, with only about thirty variables, did not produce an unacceptably large number of factors and so it was not necessary to modify eigenvalue minimums in order to make the results interpretable. With both sets of factor analyses, all loadings of a variable on any factor of less than 0.3 were omitted and considered as non-significant.

Factor analysis attempts to reveal the strongest underlying relationships amongst variables. The nature of these relationships are therefore clearly dependant on the variables subjected to analysis. This being the case, one would not expect that the same variables would necessarily group together in the same way in different contexts. Thus, when a whole test is

analyzed, it may well be that trait factors are the most powerful. On the other hand, when a single trait is analyzed one would expect that the nature of skills or tasks would be more responsible for factor structure, since the trait factor is a constant.

The possible problem of variables grouping according to difficulty was borne in mind throughout the interpretation of results. In virtually no instance however, did difficulty appear to be responsible for factor groupings. Item difficulty statistics, available in Appendix 2 and 9, can be consulted to confirm this. The apparent non-interference of difficulty may be due to the fact that there were relatively few items with extreme facility values, and that the trait, task or skill influences were stronger.

7.1. The A3 Listening Test

When the Listening Test was analyzed separately eight factors were extracted which corresponded very closely to the skills that, it was hypothesized, were being tested. These skills were:

a. writing down spelling of names and addresses (1, 12, 16, 18, 21)

b. writing down names and places that are not spelt (5, 9, 17, 19, 20)

c. writing down telephone numbers (2, 10)

d. writing down messages (instructions, places) (3, 6, 7, 8, 11, 13, 15)

e. writing down times (4, 14, 24, 25)

f. writing down simple numbers (22, 23)

g. simple comprehension (26, 27)

h. writing down prices (28)

The percentage of variance accounted for by the factors is shown in Table 6.8. below:

Table 6.8.

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	7.39715	26.4	26.4
2	1.72548	6.2	32.6
3	1.49755	5.3	37.9
4	1.39943	5.0	42.9
5	1.16087	4.1	47.1
6	1.09436	3.9	51.0
7	1.06595	3.8	54.8
8	1.02241	3.7	58.4

Factor one (please refer to Table 6.9.) groups together items that require the candidate to write down the spelling of names and addresses. In fact, there were two types of spelling represented by skills a and b above though the analysis did not differentiate between them. Factor two groups together items 24 and 25 both of which involve writing down times. Factor three also appears to be partly a time factor. It was anticipated that these four variables would group together although they did not. This may have something to do with the context in which they occurred. With items 24 and 25 there is a clear spot on the form indicating that a time is required whereas with items 4 and 14 it is up to the student to infer that the time is a necessary part of the message. The degree of guidance made available to the student seems to be having an effect here. Item 15 the third variable to group under factor 3 also requires a partly numerical response. It may be that the numerical nature of these items links them more than does their temporal nature, due perhaps to the relatively unrestricted format. However, factor four also predominantly groups together items that require the writing down of numbers - two phone numbers and a room number. This may be because it is a simpler process than that required for the previous factor. Factor five is the wholly verbal message of the second telephone task while factor six groups together simple

comprehension and the writing down of a price. Factors seven and eight include two variables each. It is not clear why items 12 and 13 both feature as part of factor seven since the skills required appear to be different. They are both part of the same task however, and this may be the reason for the grouping. Items 20 and 3 group together to form factor eight. Both items involve writing down words that have been said but not spelt, 3 being part of a message, and 20 the name of an airline.

Table 6.9.

	F1	F2	F3	F4	F5	F6	F7	F8
16	.640							
18	.498							
1	.493							
5	.353				.304			
9	.331							
19	.326							
24		.830						
25		.751						
15			.570					
4			.515					
14			.502			.425		
11								
10				.742				
21	.351			.473				
22				.442				
2				.402				
7					.652			
6					.469		.352	
8			.346		.426			
17					.306			
28						.513		
26						.433		
27						.368		
23						.346		
13							.566	
12	.303						.477	
20								.796
3			.313					.354

The results of this factor analysis confirm that there are grounds to believe that microskills involving different types of language processing skill exist, and that they can be statistically isolated. An oblique rotation produced a similar matrix although it was not as clear as the orthogonal. While item difficulty may

be playing some sort of role, it is not immediately obvious what this role is since the facility values of items in given factors are wide ranging.

7.2. The B1 Listening Test

Table 6.10. shows that ten factors with eigenvalues in excess of one were extracted accounting for 59% of the variance. The scree plot indicated that no more than five or six of these were of significance. Only eight of them were able to produce loadings of more than 0.3 on one or more of the variables. This being the case, eight factors are actually discussed.

Table 6.10.

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	6.66382	20.2	20.2
2	2.42710	7.4	27.5
3	1.67724	5.1	32.6
4	1.46944	4.5	37.1
5	1.36249	4.1	41.2
6	1.23134	3.7	44.9
7	1.20537	3.7	48.6
8	1.16863	3.5	52.1
9	1.08278	3.3	55.4
10	1.06224	3.2	58.6

It was hypothesized that the following skills were being tested:

a. Making the decision on whether something is true or false (1, 3)

b. Writing down temperatures in present and future (2, 3, 5, 6)

c. Writing down weather conditions (7, 8, 9)

d. Writing down dates (10, 11, 12)

e. Writing down stock numbers (14, 16, 18, 22)

f. Writing down quantity of goods remaining (13, 15, 17, 19, 20, 21, 23)

g. Taking down dictation -
   i. fire regulations (24 - 28)
   ii. greetings card (29 - 32)

The factor matrix in Table 6.11. groups all of the dictation items into factor 1. Dictation is a higher order language processing skill than the others in this test since it involves dealing with fairly large segments of text as opposed to small units which might require either simple numerical or verbal processing. It seems reasonable, therefore, that these items should group together.

Table 6.11.

	F1	F2	F3	F4	F5	F6	F7	F8
29	.683							
33	.605							
26	.572							
24	.547							
28	.530							
25	.501							
31	.486							
32	.386							
27	.370							
30	.344							
15		.760						
13		.600						
20		.403						
14		.374						
19		.319						
21								
8	.310		.813					
7			.784					
9			.444					
12				.598				
11				.570				
10	.336			.521				
3					.646			
2					.422			
18					.327			
6					.316			
5					.308			
23						.641		
22						.636		
1							.441	
4							.416	
17								.717
16								.440

Factor 2 corresponds quite closely to skill (f) which is a number processing skill. Factor 3 groups together items 7, 8 and 9 all of which require the student to

write down weather conditions (skill c). Factor 4 involves items that required the student to note down posting dates (skill d) whereas factor 5 groups together the items that require students to write down temperatures (skill b). These groups of items constitute separate tasks as well as skills, and it is not possible to be sure of the extent to which either the task type or the skill is responsible for the factor grouping. However, it is interesting to note that Factor 7 groups together items 1 and 4 that required Yes/No answers (skill a). If the task, or method effect were dominant, then these two items, which represent a different skill, would have been grouped with items 2, 3, 5 and 6. The fact that they were not, supports the view that the grouping is based, in part at least, on the nature of the skill being tested. With facility values of 75% and 47% respectively, items 1 and 4 are clearly not grouped on the basis of difficulty. Factors 6 and 8 are not easily interpreted and it may well be that they are grouped together because they are particularly difficult all of them with facility values of 12% or less.

This analysis again confirms **Hypothesis Three** which claimed that predefined skills can be isolated statistically thus providing further evidence to

support the validity of this performance-based battery of language tests.

7.3. The B2 Listening Test

The analysis extracted eight factors which accounted for 54% of the variance. The details are presented in Table 6.12.

Table 6.12

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	5.09489	17.6	17.6
2	2.04699	7.1	24.6
3	1.97377	6.8	31.4
4	1.59870	5.5	36.9
5	1.46227	5.0	42.0
6	1.24666	4.3	46.3
7	1.13398	3.9	50.2
8	1.04314	3.6	53.8

The following skills were being tested:

a understanding and differentiating between simple office items (1 - 6)

b understanding the location of items in an office (1 - 6)

c deciding whether items are present, not present or simply not mentioned (8 - 14)

d writing down prices (15)

e writing down dates (16)

f taking down dictated message about a series of tasks that have to be completed (17 - 21)

g taking down a dictated recipe (22 - 29)

Table 6.13. shows the rotated factor matrix. As with the B1 test, the dictation items group together to form Factor 1. Factors 2, 3 and 5 all form part of the task that involved getting information about a hotel in Singapore. The task was split into three skills areas when it was written (c, d, e), however the factor analysis divided it in a different way. Factor 2 groups together most of the items that required a Yes/No response. This type of grouping had also occurred in th B1 test. Factor 3 grouped together the two items that required a "Don't Know" response. This is a reasonable distinction in the sense that the latter responses require a degree of inferencing on the part of the student which the former do not. Factor 5 groups together items 7 and 16. Both of these require the student to actually write something down - the name of the hotel and some dates. It can be argued that this is a slightly higher order skill than the other two. It

appears therefore that the division of this task into three factors makes as much sense as the intuitive division made by the test constructors.

Table 6.13.

	F1	F2	F3	F4	F5	F6	F7	F8
23	.629							
22	.583							
27	.565							
18	.544				.422			
25	.463							
26	.360			.352				
24	.335							
19	.310							
8		.795						
10		.551						
12		.546						
14		.352						
13			.926					
9			.582		.308			
28				.593				
29	.311			.469				
21				.353				
15				.347				
16					.573			
7					.315			
11								
5								
4						.576		
6						.439		
1						.409		
17							.611	
20							-.560	
3								.743
2						.322		.392

It may be that factor 4, which groups together three of the dictation items and one of the items in V2, is the result of a difficulty factor more so than anything else. The first task in the test, related to an office inventory, divides into two factors, 6 and 8. It is not clear exactly why this should be the case.

The results of this analysis are not as convincing as the earlier ones. There nevertheless appears to be a clear skills distinction which supports the earlier findings.

7.4. The B3 Listening Test

A total of ten factors, presented in Table 6.14., had an eigenvalue greater than one and they accounted for 62% of the variance. However, only seven of them loaded on variables meaningfully.

Table 6.14.

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	6.69368	20.9	20.9
2	2.36644	7.4	28.3
3	1.79390	5.6	33.9
4	1.60906	5.0	38.9
5	1.40374	4.4	43.3
6	1.31351	4.1	47.4
7	1.21463	3.8	51.2
8	1.17131	3.7	54.9
9	1.09136	3.4	58.3
10	1.07331	3.4	61.7

The factor structure is again fairly close to the skills that it was intended should be tested. A list of these skills follows:

a writing down names:
- not-spelt (1, 2, 5, 6, 9, 10, 13)
- spelt (14)

b writing down messages (3, 4, 7, 8, 11, 12, 15, 16, 17)

c recognizing descriptions of people/things (18 - 21)

d writing down a dictated letter (22 - 28)

e writing down arrival/departure times (29 - 30)

f writing down a city destination (Tokyo) (31)

g writing down a flight number (32)

The rotated matrix for the seven factors is presented in Table 6.15. Factors 1, 4, and 5 were all basically

involved with the telephone messages tasks. Factor 1 predominantly grouped together items that required the writing down of a name. At this level of proficiency, the names were generally not spelt out as they had been at the A3 level. The item with the weakest loading on this factor was item 14, where the name was actually spelt out. Factors 4 and 5 on the other hand comprised items that involved writing down messages. It is not clear why these items should have split into two factors, nor why items 4 and 8 should have grouped with Factor 1 rather than one of the two message factors, however, it is interesting to note that items 3, 7 and 15, which form Factor 4, are all based on the first part of the message. Perhaps this has something to do with the nature of the grouping.

Table 6.15.

	F1	F2	F3	F4	F5	F6	F7
1	.828						
13	.769						
5	.682						
2	.515						
9	.471						
8	.401						
4	.392						
30		.802					
32		.647					
29		.622					
28			.595				
26			.499				
23			.482				
31			.381				
25			.372				
22			.319				
27							
10				.565			
3				.541			
7				.519			
15				.489			
12					.839		
11				.394	.489		
17					.336		
24						.720	
16						.375	
19							.552
20							.349
6							
14	.360						
21							.366
18							

A dictation factor (Factor 3) reemerged even though it had not appeared in the whole test analysis (see Appendix 12). Factor 2 grouped together three of the

four items in the airport information task. It is noteworthy that item 31, requiring the candidate to write down the word "Tokyo" as opposed to a time or flight number grouped with the dictation factor as opposed to the flight information task factor. Factor 7 appeared to be a multiple-choice task factor.

7.5. The C1 Listening Test

Ten factors emerged in the factor analysis of the listening test although only seven of these proved to be interpretable. Percentages of variance are listed in the table below:

Table 6.16.

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	7.28477	21.4	21.4
2	2.37058	7.0	28.4
3	1.72261	5.1	33.5
4	1.45355	4.3	37.7
5	1.38814	4.1	41.8
6	1.30648	3.8	45.7
7	1.23020	3.6	49.3
8	1.16535	3.4	52.7
9	1.14418	3.4	56.1
10	1.09211	3.2	59.3

The rotated factor matrix is presented in Table 6.17. It was intended that the following skills should be tested:

a writing down names (not spelt) (1, 4)

b writing down names (spelt) (2, 5)

c writing down short messages (3, 7)

d writing down telephone numbers (6)

e making simple decision about English level (8. 9)

f making simple decision about education (10, 11)

g noting down information about:

- job experience (12, 13)
- age (14, 15)
- personality (16, 17)

h Noting down information about:

- destinations (23)
- departure times/destinations (18, 19, 24, 30)
- activities (20, 21, 22, 25 -29)

i Taking a dictation about a day's activity (31 - 34)

Factor 1 is mainly comprised of items in the travel task (skill h) which required the writing of information onto a grid. Items 28 and 29 which required the candidates to infer that particular slots on the

timetable were free time did not load on factor 1. They grouped to form factor 4.

Table 6.17.

	F1	F2	F3	F4	F5	F6	F7
25	.583						
21	.568						
19	.544						
24	.504						
23	.491						
20	.486						
18	.375		.322				
27	.355						
30	.353						
22	.351						
26							
11		.725					
10		.694					
9		.590					
12		.469					
13							
32			.619				
31			.614				
33			.573				
34			.514				
29				.897			
28	.315			.765			
5					.538		
2					.451		
3	.385				.423		
6					.413		
7					.346		
16						.475	
1						.421	
17						.377	
4						.323	
8						.321	
15							.631
14							.617

The interview task, V3, was split amongst three factors, 2, 6, and 7. Factor 2 included most of the items that required a Yes/No response (skills e and f). Factor 7 involved the two items that required writing down the ages of candidates. The two items that involved writing down an impression of the personalities of the two candidates grouped with two of the telephone task items that required the writing down of names (skill a). Factor 3 was based on the dictation task, while factor 5 involved skills (b) and (c), taking messages and writing down names.

The breakdown of skills for this test were rather detailed, and the factor analysis was not able to differentiate them to the same extent. It does however, support the general trend for items to fall into categories that are more or less in line with intuition.

7.6. The C2 Listening Test

The nine factors extracted by the factor analysis accounted for 56% of the variance as shown in Table 6.18. below.

Table 6.18.

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	6.17793	19.9	19.9
2	2.06072	6.6	26.6
3	1.66496	5.4	31.9
4	1.57514	5.1	37.0
5	1.36023	4.4	41.4
6	1.19105	3.8	45.3
7	1.09172	3.5	48.8
8	1.07636	3.5	52.3
9	1.04151	3.4	55.6

The following skills were being tested:

a Writing down names:
- not spelt (1, 2, 6)
- spelt (7)

b Writing down address (8)

c Writing down time (5)

d Writing down short messages (3, 4, 9)

e Making changes to an appointments diary:
- recognizing which information to change
- crossing out old information
- writing in new information (10 - 16)

f Writing down a dictated job advert (17 - 24)

g Answering specific questions on two news broadcast extracts (25 - 31)

The rotated factor matrix (see Table 6.19.) produced eight interpretable factors. As with previous analyses, the dictation task formed a clear factor, in this case factor 1. The news items loaded for the most part on factor 2. Writing down names seemed to be the main feature explaining factor 3. The other items grouped under one or more of the remaining factors. Items 14 and 15, which involved editing the advertisement, loaded on factor 4, while items 12 and 13, which seemed to represent a similar skill loaded on factor 8.

Table 6.19.

	F1	F2	F3	F4	F5	F6	F7	F8
23	.589							
19	.584							
21	.519							
17	.516							
18	.502							
26	.449							
20	.416							
22	.382							
25	.344							
4	.332				.317			
31								
30		.705						
28		.579						
29		.485				.417		
10		.377						
6			.739					
8	.317		.542					
7			.470					
16			.350				-.326	
2			.338			.305	.312	
14				.888				
15				.560				
5					.544			
3								
27								
1						.526		
11							.452	
9							.310	
13								.551
12								.443
24					.355			

As the tests become more advanced, it is clear that the results are not as neat as with those of a lower level. This seems reasonable since the tasks are supposed to

be more integrated and more difficult as the level goes up. However, there still appears to be a relationship between hypothesized skills and the factors revealed by the analysis.

7.7. The C3 Listening Test

Nine factors were extracted by the analysis although only seven of them were interpretable. The nine factors accounted for 56% of the variance.

Table 6.20.

FACTOR	EIGENVALUE	PCT OF VAR	CUM PCT
1	6.16496	19.3	19.3
2	2.46065	7.7	27.0
3	1.81607	5.7	32.6
4	1.54176	4.8	37.4
5	1.51977	4.7	42.2
6	1.21489	3.8	46.0
7	1.09790	3.4	49.4
8	1.07193	3.3	52.8
9	1.04521	3.3	56.0

The following skills were being tested:

a writing down names (1, 2, 3, 6, 7)

b writing down short messages (4, 5, 8, 9)

Correcting an advertisement from instructions. This involves:

c - changing names (10)

d - changing numbers (12, 14, 16)

e - correcting spelling (11, 13)

f - adding information (15, 17)

g Answering specific questions on two news broadcast extracts (18 - 22, 23 - 27)

h Listening to instructions and noting down:

- duties (28)
- location (29)
- objects required (30 - 32)

The first factor (see Table 6.21.) seems to be made up of a combination of skill (a) and (b) items.

Table 6.21.

	F1	F2	F3	F4	F5	F6	F7	F8
7	.640							
6	.596							
3	.561							
5	.518							
2	.425							
4	.414							
8	.338							
12		.783						
13		.608						
15		.579						
23		.370		.362				
22			.636					
27			.525					
20			.400					
19			.397					
18								
24				.463				
25			.419	.422				
26				.383				
14		.367		.379				
31					.753			
9	.310				.379			
32					.326			
10						.604		
11		.409				.481		
16				.351		.382		
17								
1	.336						.710	
30							.347	
29								.651
21								
28								

All of the items are related to taking telephone messages. It may be that there is a task effect at play

here. Factors 2 and 6 involve items in the advertisement editing task. It is not clear why they are split up in this way. Similarly, the news items are split between Factors 3 and 4, although there is some overlap. These items were similarly split up in the C2 test. It is not clear why this is the case. If a task effect were dominant then it would seem logical that they should group together. It may be that there is something in the content of the news broadcasts that is having an effect. The difficulty of the items may also have something to do with the split. Factor 5 appears to be mostly related to skill (h), and one would have expected the other items in this task to group here too. They did not.

7.8. Concluding Comments

Several conclusions can be drawn based on the analyses carried out in this section. Firstly, it seems clear that items in all of the tests tend to group by trait (see Appendix 12). The groupings reflect closely those that were achieved in the previous section (factor analysis by task) which gives the results additional strength. It is clear that the groupings are not based on statistical peculiarities but are rather a

reflection of the divisible nature of language proficiency as measured by this battery. The fact that the results are fairly consistent across the whole battery, which includes seven tests, taken by different students, containing different items and marked by different teachers, demonstrates beyond doubt that the results are valid.

Secondly, it appears that the item-by-item factor analysis often produces results that are more sensitive than the task-by-task analysis. The performance-based approach is clearly powerful and discriminating in language test construction.

Thirdly, it is often the case that items within a particular trait split to form two or more factors (see Appendix 12). It is not always clear why this is the case, but it appears that there a number of influences interacting to affect test performance. The scope of this study did not permit a detailed investigation of these. However, the nature of tasks, their difficulty, text types, topics, and the complexity of processing required all seem to have an influence on performance.

Fourthly, the subtest analyses tended to group items according to the skills areas that it was hypothesized were being tested. Although this grouping was far from perfect, and often suggested that the skills were based on unpredicted features of proficiency, there was enough evidence in individual subtests, and across the battery that it is possible to isolate microskills statistically. There is strong evidence to suggest that language processing, as measured by the listening tests at least, involves different abilities. There appears to be such a thing as numerical processing as distinct from verbal processing. There also appears to be a difference in the ability to process short pieces of text as opposed to longer ones. Additionally, there appears to be a distinction in the ability to deal with complex tasks as opposed to simpler ones. These findings reflect intuitive feelings that both teachers and to some extent language testers have. Previously, however, there has been no attempt to isolate and statistically confirm the existence of the microskills that form the basis of many approaches to both teaching and testing.

It was clear from all the analyses presented in this chapter and suggested by the previous one that language proficiency as measured by this battery of tests is

divisible in nature, thus supporting **Hypothesis Two**. It may be that this divisibility is due in part to the nature of the tasks present in the tests although the range of tasks make this view somewhat untenable. It seems more likely that while there may be a general language processing factor of some sort, individuals have different experiences and capabilities that have a significant effect on their ability to deal with different types of language and language situations. This makes for variable performance in different tests. **Hypothesis Three**, which claimed that microskills could be statistically isolated, found a reasonable degree of support in the factor analyses by item of the listening section of each progress test.

8. Conclusion

Factor analysis is a volatile and complex tool that is susceptible to major variations in results with only minor modifications in procedure. Much early research into the nature of language competence using the technique has been criticized for precisely this reason, and many results attributed to the nature of the statistics used as opposed to the nature of language proficiency. Care needs to be taken to ensure

that the methodology used in any statistical analysis is appropriate and this can be done in several ways. A detailed study of the particular technique used needs to be carried out. Analysis should not be limited to a narrow sample of students or testing instruments. Where possible parallel analyses should be carried out in order to confirm results. All of the above strategies were employed with regard to the statistical analyses carried out in this chapter thus leading the writer to conclude that the results are not due to the procedures but related in some way to the nature of language proficiency.

Factor analysis suggests solutions to research questions, it does not provide them. The interpretation of results is the responsibility of the researcher. Due to the volatility of the technique different solutions can be generated from the same data set. It is important to be aware that any set of results is open to question and that their validity is dependant on the competence and honesty of the researcher.

Chapter VII

1. Introduction

This thesis has attempted to show how a performance-based battery of English language progress tests was developed and validated.

Canale (1985) amongst others has pointed out that there is often a mismatch between teaching/learning materials and those that appear in proficiency-oriented achievement tests. He attributes this mismatch to what he calls the 'image problem', which he breaks down into several categories. First he focuses on the role of the learner in testing and describes him as typically:

> "an obedient examinee, a disinterested consumer, a powerless patient or even an unwilling victim."

Canale also focuses on the type of situation that current achievement testing often represents:

> "... it is frequently a crude, contrived, confusing, threatening, and above all intrusive event that replaces what many learners (and teachers) find to be more rewarding and constructive opportunities for learning and use."

The problems that Canale outlines, which are also of concern to Swain (1985), were perceived as major difficulties in the acceptability of testing as an important and useful part of the educational process by the writer of this thesis. Several strategies were adopted to overcome such difficulties in the context of the British Council institute in Hong Kong.

Firstly, it was considered of vital importance that the testing programme be integrated into the life of the teaching institute. The testing specialist took an involvement in needs analysis and course development, always trying to ensure that the tests were not seen as simply an afterthought in curriculum design but an integrated part of the whole process.

Secondly, the materials used in the tests always attempted to reflect the types of activities that went on in the classroom and/or the lives of the students taking the tests. In this way it was anticipated that both students and teachers would clearly see their relevance. It was considered important that the tests should not fall out of line with the needs of the students and practice of the teachers. In fact, the

tests sometimes even preceded teaching practice to some extent, with the introduction of realistic tasks and authentic materials, particularly at the lower levels.

Canale (1985) and Swain (1985) argue that there is not enough student involvement in the testing process. It is true that students were not involved in deciding on testing activities in the context of the work described in this thesis. However, within the institute, through the development of the descriptions of student performance outside the classroom, teachers were encouraged to negotiate with their students as to the type of activities that were of most relevance to them and the tests always tried to focus on such activities. In this way, it was felt that students would realize that they were participating in deciding what they should learn and consequently, what should be tested.

Thirdly, teachers' sometimes inadequate understanding of testing purposes, procedures and principles were considered to be a major potential barrier in the successful integration of testing into the curriculum. In order to overcome it, teachers were actively encouraged to become involved in the writing of tests, and there was a heavy emphasis on the training of those

teachers who did become involved. This strategy not only improved the quality of the tests, in terms of reliability and validity, but also meant that an ever increasing number of teachers were becoming familiar with testing as a discipline. Thus the tests were a joint effort between testing specialist and teachers as opposed to simply the results of one person's work. This greatly increased their acceptability.

Three Hypotheses were formulated at the beginning of this project, and they are now discussed below, in light of the research findings.

2. Hypothesis 1

Hypothesis 1 stated that:

> A reliable and valid performance-based test battery could be constructed that would be of at least equivalent standard (as measured by classical test theory) to tests of a traditional format.

2.1. The Question of Reliability

Through the development of a comprehensive and versatile item analysis package for the microcomputer, previously not in existence, I was able to show that the seven tests under discussion in this thesis conformed to high standards of reliability. Comparative data, based on a more conventional approach to testing, was not gathered even though it is common practice in research contexts of this sort to establish the accuracy of a hypothesis through comparison. Due to the context in which this project took place, it was not really feasible or valid to devise a conventional test battery to run in parallel with the performance-based one. This being the case, any comparative claims made about the reliability of the battery had to be based on the strength of its results alone. By traditional testing standards, the reliability of the subtests and items was extremely high. It may be that a more conventional approach would have yielded equivalent results but highly unlikely that they would have been better. On the other hand, a more conventional approach would not have been able to boast the same degree of relevance to real-world activities. Nor would it have been able to claim the same degree of positive washback effect.

A problem that is immediately apparent, however, is related to the question of context in which items appear in relation to test reliability. In Chapter 4,

I stated that it was important for there to be adequate sampling procedures. I did not lay down any detailed guidelines as to how they were to be achieved over and above ensuring that thorough descriptions of courses and student behaviour outside the classroom should be available to the test constructors. Had it been possible to base the tests purely on a detailed linguistic description, for example, then the problem of generalizing from the sample, while still a difficulty, may not have appeared very important. However, these tests also served as a means of modifying the approach to teaching. They drew on real-world performance, as well as course content, as a source of items. The students came from a range of backgrounds, with varying experience and exposure to English. The context in which an item might be analyzed could therefore appear to be radically different. The reliability statistics are based on the immediate context of items in a test. Had the context been different then items may have displayed different characteristics in terms of reliability statistics. What is not clear is the extent to which this

difference in performance of items may be significant. In reality, of course, this problem concerns any collection of items in a test. Item Response Theory, using either two or three parameter models (Tung, 1986), claims to be able to overcome the problem of context to some extent. However, unidimensionality of trait and very large sample sizes are required when the three parameter model is applied, and to date, it has been used largely with areas of competence such as grammar and vocabulary, with items falling into the multiple-choice category for the most part. It is true that Pollit and Hutchinson (1987) have applied Rasch partial credit analysis to the performance of writing, but this work is very recent. It is accepted that traditional test analysis, which also requires unidimensionality of trait, may not be the best way of establishing the reliability of items beyond the context in which they occur, even though it has been used for decades. It will be necessary in future to experiment with other approaches, such as those mentioned above, in order to try and take into account the difficulty of context, and provide some empirical data to help compensate for the difficulties of sampling and in turn extrapolation from test results.

2.2. The Question of Validity

These tests satisfy several conditions which make them valid measures, at least in terms of content and face validity. Through informal feedback we were able to confirm that both teachers and students agreed the tests seemed to be testing relevant features of the courses and aspects of students' immediate and future requirements of the English language. An inspection of the descriptions of real-world performance and course outlines (Appendix 5 and 6) shows that the battery is content valid. The comparisons of teachers' subjective assessments of students' performance in the courses with their performance on the progress testing battery indicated that there was a significant correlation between the two. While this measure was not of central concern, it did at least indicate that the battery could claim a degree of concurrent validity. Through informal feedback, we were also able confirm that teachers rarely felt that the tests disagreed significantly with their own view of students' competence at the extreme ranges of ability. That is to say, students teachers regarded as weak overall generally scored low on the test, while very good students scored high. We were not able to say anything definite about the middle range, but in the context of

the teaching institute this was not of such great concern. At some point in the future it would be desirable to conduct a more in depth investigation of the relationship between teachers' subjective gradings and the results generated by the tests. It must be pointed out, however, that such an investigation is by no means a simple matter. Concurrent validation, particularly with the subjective gradings of teachers, begs many questions as to its own validity. These would need to be addressed in depth before any meaningful investigation could be conducted.

It would also be worthwhile attempting to establish the predictive validity of the battery since one of the underlying premises of a performance-based approach is that test results should predict the students' ability to actually use language to carry out real-world tasks. However, an equally important function of performance-based tests, within the context of a teaching institute, is the powerful positive effect that they can exert on the content of courses, the practice of teachers and the attitudes and motivation of students (Wesche, 1987). Within the context under discussion here, the positive washback effect was of greater concern, in the first instance, than the predictive validity of the battery. In the future, however, it

will probably be necessary to carry out some studies that attempt to establish predictive validity. While great efforts were made to integrate testing successfully into the life of the institute it must not be forgotten that the test results will also be used by students outside the institutional context to make claims to their employers about their ability to use the English language. Should it be the case that the tests are not adequately predictive of real-world performance then the institute will encounter problems of credibility. This issue is largely ignored in most educational contexts. For example, many people have been claiming for years that A-levels are not a very good predictor of university performance yet nothing ever seems to happen to change the status quo in any significant way. On the other hand, a commercial operation, such as the British Council in Hong Kong, tends to find itself more accountable than an Examining Board. An employer would, for example, accept that a certain grade in the School Certificate examination may not be very reliable in determining the real-world performance of a potential or current employee. He would be far less forgiving of a British Council certificate which claimed that a student could use the telephone adequately only to find out that this was not the case. Some discussion of the issues concerned in predictive validation was conducted in Chapter 3. From

these it was evident that the area is fraught with problems. In the context of the British Council in Hong Kong, establishing some degree of predictive validity would be a major and complex undertaking. It was quite beyond the scope of this investigation.

3. Hypothesis Two

Hypothesis Two stated that:

> It can be demonstrated statistically that students' ability in different language skills and areas of communicative competence are not equivalent.

The reasons for the investigation of this hypothesis are elaborated in Chapter 1. A number of previous studies have claimed that some sort of unitary underlying competence accounts for superficial variation in test performance. Such a claim has been supported by evidence based on correlations and factor analyses. At the time of test construction, evidence was already mounting that the strong form of the unitary competence hypothesis was no longer viable and

that while it could not be claimed that competence was divisible in the extreme, it was at least partially so.

Evidence to support **Hypothesis Two** came from three different sources. Firstly, the conventional item analysis revealed that the point-biserial correlations of items were significantly higher when they were analyzed as part of a subtest as opposed to part of a whole test. This indicated that performance on the subtests was different and that if they were analyzed together these differences would act against each other. While there was no doubt that some sort of overall ability existed, as it would if any group of tests were put together, there was also clearly a distinct set of abilities related to the subtests that would be submerged if the differences were not taken into account.

This condition has important implications for test moderation. Items which do not perform well as part of the whole test may well be more than adequate in the context of their own subtest. It would not therefore be appropriate to moderate them on the basis of whole test statistics. Unfortunately, this finding returns us to the problem of the context in which items occur, which

was discussed. Conventional item analysis is context-bound and yet the dimensions of language ability are by no means clearly defined. As yet there has been no conclusive research in this area. While I was aware of these difficulties, I decided to stop modifying the context for item analysis at the level of the subtest. However, the item analysis programme was designed to allow for the selection of groups of items as the basis for analysis. Any group of items from any part of a test could be isolated and treated as a test in their own right. This facility was not used in this thesis since it would have magnified the scope of the investigation beyond reasonable limits. However, it is an area that can be explored at some time in the future from the data already gathered.

The second piece of evidence to support **Hypothesis Two** came from the correlational analyses discussed in Chapter 5. The subtests were always significantly correlated among themselves, and correlated at between 0.5 and 0.8. with the total test score. The subtests did not overlap sufficiently for it to be claimed that any of them are redundant, or that performance on any one can accurately predict performance on the others.

In addition, the tasks were also correlated. It was generally the case that tasks purportedly measuring the same trait were more closely related to each other than they were to tasks measuring different traits. It would have been better if I had been able to apply the multimethod-multitrait approach in order to establish the extent to which this convergence of tasks was a result of method as opposed to trait. Unfortunately, the constraints of designing a test battery that was actually to be used made it very difficult to do this. I was able to show, however, that the mean correlations of intra-trait tasks were significantly higher than those of inter-trait tasks. It would be worthwhile conducting a true multitrait-multimethod study at some time in the future, particularly as most investigations into trait and method effect have not focused on performance-based items.

Hypothesis Two was also supported by the factor analyses discussed in Chapter 6. The factor analysis of tasks clearly showed that there was a relatively strong primary factor, but that there were also a number of other significant factors. The analyses tended to group tasks according to trait. Furthermore, when the items themselves were subjected to factor analysis, they too

tended to group according to trait, although there also appeared a number of other reasons for groupings.

3.1. The Relationship Between Proficiency Level and the Structure of Communicative Competence

For the purposes of the test battery discussed in this thesis it was hypothesized that communicative competence was at least partially divisible and to that end the tests were broken into four - five sections. The correlational analyses did not indicate that the relationships between these sections varied systematically at different levels of proficiency. However, when the tests were subjected to factor analysis by task, and then by item it appeared to be the case that the dividing lines between tasks and items became rather more blurred as the level of proficiency increased. With the A3 test, for example, both of the factor analyses produced very neat cuts with a minimal amount of overlap and most of the items loaded at above 0.3 in the factor analysis by item. There appeared to be a clear structure at this level. On the other hand, at the C3 level, while a structure was still in evidence, it was no longer as clear and neat. This breakdown appeared to be progressive as the

level of proficiency increased and the skills seemed to overlap more as the students got better at English.

This type of blurring may be due to several reasons. Firstly, it could be that the tests at the lower levels were better written than those at higher levels, or perhaps that the students at the lower levels formed a more homogeneous group vis a vis language proficiency. Secondly, and this point is related to the first, we, as test constructors, certainly found it easier to isolate skills at the lower levels than we did at higher levels. Thirdly, assuming that there was no difference in the quality of the tests, and the item analysis does not indicate that the higher level tests were any worse than the lower level ones, it might be claimed that communicative competence is more complex phenomenon at higher levels of proficiency. Whereas it appeared that trait was the main reason for factor groupings at lower levels, a task effect seemed in evidence at the higher levels. The research discussed here only made indications that the nature of communicative competence may differ at different levels of proficiency. How and why this might be the case is not clear. However, it is an area that is worthy of further investigation.

While there was always a principal factor emerging in all of the analyses, this principal factor was never of a magnitude to compare with most previous studies. This may have been due to several reasons. Firstly, as the number of variables included in a factor analysis increases it is likely that the magnitude of the principal factor will decrease (Cureton, 1982). In this study the item-by-item factor analyses included as many as 112 variables while the task-based analyses included an average of about 15 variables. This is a larger number than many previous studies. Secondly, the nature of the variables analyzed differed from previous studies in the sense that they were performance-task based or item-based, as opposed to subtest-based, hence the degree variability was greater. It may be that one of the reasons for such powerful principal factors in previous studies is based as much on the type of data subjected to factor analysis as on the nature of communicative competence. The same argument may be used to criticize the data investigated in this thesis, but the fact that very powerful principal factors do not emerge casts doubt on the validity of much previous research.

4. Hypothesis Three

Hypothesis Three stated that:

> It can be demonstrated statistically that performance in communicative tasks is, at least partially, divisible into micro-skills.

The methodology used for the factor analysis of the Listening subtests was discussed in Chapter 6. There was clear evidence to support the hypothesis that microskills could be isolated statistically, and that items tended to group in a way which had been predicted. As with the other factor analyses, the groupings tended to be more clear-cut at lower proficiency levels. This is possibly the case because the tests were better constructed at lower levels, or because the microskills, as measured by the tests in the battery at lower levels, are simpler and easier to isolate. Furthermore, most of the microskills were of a fairly low order throughout the battery. It would be appropriate to adopt the same methodology with higher order skills in order to establish whether the same results would be achieved.

5. The Importance of Sound Test Construction

The quality of the results generated in this thesis is high and it is important to ask why this was the case, since most test batteries do not perform as well. It seems likely that the high quality is due to the time and effort taken in test construction. Each test was extensively moderated and pretested prior to the state it is now in. It is extremely important in test construction to devote adequate time and effort to the moderation and pretesting phases. The procedures adopted with regard to these two phases are clearly outlined in Chapter 4.

In addition, it is my belief that the process of test construction, within the context of a teaching institute, needs to be a cooperative effort. That is to say, it is imperative that teachers, who are actually working with the students need to share in the test writing process. They can bring to bear their valuable experience as to the appropriateness of tasks and the extent to which they are relevant and of the right level.

It is also important that teachers are involved in test construction because their awareness of the procedures and problems can be greatly increased in this way. The involvement of teachers can compensate for some extent for the training deficiencies mentioned in Chapter 2. It was certainly the case in the British Council institute in Hong Kong that a strong emphasis on teacher involvement in test construction helped to integrate testing into the life of the institute, equip teachers better for test construction in the future and improve the quality, reliability and validity of the battery.

6. The Use of Factor Analysis in Language Testing

Factor analysis has been used frequently in recent Language Testing research. It has been used by Oller and others as a means of exploring underlying features of language processing. It has also been used by researchers like Bachman, Palmer, Vollmer and Sang as a means of theory confirmation. The two approaches require different types of factor analysis, the former being exploratory, and the latter confirmatory. The study discussed here falls somewhere between the two

approaches mentioned above, and it uses factor analysis in a way that the others do not.

Previous research, be it confirmatory or exploratory in nature, has tended to restrict itself to the level of test or subtest analysis. This is reasonable given the types of tests it appears the researchers were working with. It is not apparent that any of the earlier research using factor analysis attempted to approach language from a performance-based point of view. Researchers might of course claim that their tasks, cloze for example, are performance-based because they are integrative activities, requiring the processing of real language. They do not, on the other hand, represent activities that are performance-based in the sense that real users of the language actually have to do them. Unfortunately, language tests developed in Britain over the last ten years, such as the R.S.A. Examination in the Communicative Use of English, or the English Language Testing Service Examination, while adopting a more performance-based approach, in the sense mentioned above, have not been construct validated using factor analysis. Some attempts have been made at establishing predictive validity and others have focused on a priori validation (Weir, 1983). However, a posteriori construct validation has

been lacking. The study discussed here bridges this gap to some extent.

Firstly, the items in the tests are generally performance-based in that they attempt to test things that students actually have to do using English. Secondly, the factor analysis does not focus on the subtest level as previous studies have done, but rather on the level of task and item. In performance-based testing this is certainly a more pertinent focus. It is surprising that it has not been done before. Thirdly, the tasks included in the analyses are decided on the basis of research and pedagogic requirements, and thus seem to be in themselves a more valid selection of tasks than many previous studies have been able to claim. It would be a good idea in future for studies of this sort to devote more attention to the types of tasks included in an analysis. Fourthly, the tests cover a range of proficiency levels and a large number of task types. Most previous studies have limited themselves to a single proficiency level and a rather unimaginative selection of task types. In a similar vein, however, the study discussed in this thesis may be criticized because of the monolingual nature of the population and the fact that the subjects are enrolled in a teaching institute. It may be that pedagogic

influences, as suggested by Sang et al. (1986), are influential in the generation of results.

Factor analysis, like any other statistical tool, is dependant on input in order to generate meaningful output. Most previous studies have satisfied themselves with viewing input in one way only - on the level of the subtest. Even confirmatory studies like that of Bachman and Palmer (1983), for example, which claim numerous attempts at matching a theoretical position with the data available, do not modify the way in which the factor analysis actually looks at a set of data. In the study discussed in this thesis, the hierarchical approach to the exploration of the data is an important feature. It has the potential to reveal a greater number of relationships on both a macro and micro level than the more conventional horizontal approach to factor analysis. It seems to me that construct validation needs to move further in the direction of vertical, or hierarchical analysis if it is to continue to make meaningful revelations about the nature of communicative competence.

References

Abbs, B. and I. Freebairn, 1977, Starting Strategies (1), Longman, London.

Abbs, B. and I. Freebairn, 1979, Building Strategies (2), Longman, London.

Abbs, B. and I. Freebairn, 1980, Developing Strategies (3), Longman, London.

Abbs, B. and I. Freebairn, 1981, Studying Strategies (4), Longman, London.

Abu-Sayf, F.K., J.B. Herbolich and S. Spurling, 1979, The identification of major components for testing English as a foreign language, TESOL Quarterly, Vol. 13, 117-120.

Acheson, P., 1977, English for speakers of other languages: a survey of teacher preparation programs in American and British colleges and universities, In Fanslow, J.F., and R.L. Light, (eds.) 1977, 69-81.

Alderson, C.J., 1978, A study of cloze procedure with native and non-native speakers of English, Unpublished Phd. Dissertation. University of Edinburgh.

Alderson, C.J., 1981, Reaction to the Morrow paper, In Alderson, C.J. and A. Hughes, (eds.), 1981.

Alderson, C.J., 1983, The cloze procedure and proficiency in English as a foreign language, In Oller, J.W. (ed), 1983.

Alderson, J.C. & Hughes, A., (eds.), 1981, Issues in Language Testing, ELT Documents 111, The British Council, London.

Alderson, J.C. and A. Waters, 1983, A course in testing and evaluation for ESP teachers or 'How bad were my tests?' In A. Waters, (ed.) 1983.

Allen, D., 1982, The Oxford Placement Test, Oxford University Press, London.

Allen, H.B. and R.N. Campbell (eds), 1972, Teaching English as a Second Language: A Book of Readings, McGraw-Hill, New York.

Allen, J.P.B. and S.P. Corder, (eds.), 1974, The Edinburgh Course in Applied Linguistics: Techniques in

Applied Linguistics - Volume 4, Oxford University Press, London.

Allen, J.P.B. and A. Davies (eds.), 1977, The Edinburgh Course in Applied Linguistics: Testing and Experimental Methods, Oxford University Press, London, Volume 4.

Alpert, R. and R.N. Haber, 1960, Anxiety in academic achievement situations, Journal of Abnormal Psychology, Volume 61, 207-15.

American Psychological Association, 1974, Standards for Educational and Psychological Tests and Manuals, American Psychological Association, Washington D.C..

Bachman, L.F., 1982, The trait structure of cloze test scores, TESOL Quarterly, Vol. 16, 61-70.

Bachman, L.F., 1981, Formative evaluation in programme development, In Mackay R. and J.D. Palmer (eds.), 106-116.

Bachman, L.F. & Palmer, A.S., 1981, A multitrait-multimethod investigation into the construct validity of six tests of speaking and reading, TESOL Quarterly.

Bachman, L.F. and A.S. Palmer, 1983, The construct validity of the FSI oral interview, In Oller, J.W. (ed), 1983.

Beebe, L. and J. Zerengler, 1983, Accomodation theory: an explanation of style shifting in second language dialects, In Wolfson, N. and E. Judd, (eds.), 1983.

Bird, E. and M. Dennison, 1987, Teaching GCSE: Modern Languages, Hodder and Stoughton, London.

Bock, R.D. and R.E. Bargmann, 1966, Analysis of covariance structures, Psychometrica, Vol. 31, 507-534.

Broadfoot, P., 1979, Assessment, Schools and Society, Methuen, London.

Broadfoot, P. (ed.), 1986, Profiles and Records of Achievement - A Review of Issues and Practice, Holt Educational, London.

Brown, F.G., 1981, Measuring Classroom Achievement, Holt, Rinehart and Winston, New York.

Brown, J.D., 1983, A closer look at cloze: Validity and Reliability, In Oller, J.W. (ed), 1983.

Brown, S. & Munn, P. (eds.), 1985, The Changing Face of Education 14 to 16: Curriculum and Assessment, NFER - Nelson, Windsor.

Brumfit, C.J., 1984.

Brumfit, C.J. and K. Johnson (eds), 1979, The Communicative Approach to Language Teaching, Oxford University Press, London.

Campbell, D.T. and D.W. Fiske, 1959, Convergent and discriminant validation by the multitrait-multimethod matrix, Psychological Bulletin, Vol. 56, 81-105.

Canale, M. & Swain, M., 1980, Theoretical bases of communicative approaches to second language teaching and testing, Applied Linguistics, Vol. 1, No. 1.

Canale, M. & Swain, M., 1981, A theoretical framework for communicative competence, In Palmer et al., (eds.), 31-36.

Carroll, B.J., 1978, Guidelines for the Development of Communicative Tests, Royal Society of Arts, London.

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

Carroll, B.J. and P.J. Hall, 1985, Make your own Language Tests: A practical Guide to Writing Language Performance Tests, Pergamon Press, Oxford.

Carroll, J.B., 1958, A factor analysis of two foreign language aptitude batteries, Journal of General Psychology, Vol. 58, 3-19.

Carroll, J.B., 1961, The nature of the data, or how to choose a correlation coefficient, Psychometrica, Vol. 26, 347-372.

Carroll, J.B., 1968, Psychology of language tests, In Davies, A. (ed.) 1968.

Carroll, J.B., 1972, Fundemental considerations in testing for English language proficiency in foreign language students, In Allen H.B. and R.N. Campbell (eds), 1972.

Carroll, J.B., 1975, The Teaching of French as a Foreign Language in Eight Countries, Halsted, New York.

Carroll, J.B., 1983, Psychometric theory and language testing, In Oller, J.W. (ed), 1983.

Canale, M., 1985, Proficiency-oriented achievement testing, Paper presented in the Master Lecture Series of the American Council on the Teaching of Foreign Languages.

Carroll, J.B. and S.M. Sappon, 1959, The Modern Language Aptitude Test, Harcourt Brace Jovanovich/The Psychological Corporation, New York.

Cattell, R.B., 1971, Abilities - their Structure, Growth and Action, Houghton Mifflin, Boston.

Cattell, R.B., 1952, Factor Analysis: An Introduction and Manual for the Psychologist and Social Scientist, Harper and Row, New York.

Cattell, R.B., 1966, The Scree Test for the number of factors, Multivariate Behavioural Research, Vol. 1, 245-276.

Cattell, R.B., 1978, The Scientific Use of Factor Analysis in Behavioural and Life Sciences, Plenum Press, New York.

Chaplen. E.F., 1970, The Identification of Non-Native Speakers of English Likely to Under Achieve in University Courses through Inadequate Command of the Language, Unpublished Ph.D. Thesis, University of Manchester.

Chater, P., 1984, Marking and Assessment in English, Methuen, London.

Chomsky, N., 1965, Aspects of the Theory of Syntax, M.I.T. Press, Massachusetts.

Clark, J.L., 1987, Curriculum Renewal in School Foreign Language Learning, Oxford University Press, Oxford.

Clark, J.L.D., 1972, Foreign Language Testing: Theory and Practice, The Center for Curriculum Development, Philadelphia.

Clarke, J.L.D., 1978, Psychometric considerations in language testing, In Spolsky, B. (ed), 1978.

Clifford, R.T., 1981, Convergent and discriminant validation in integrated and unitary language skills: the need for a research model, In Palmer A.S. et al. (eds) 1981.

Cohen, A.D., 1980, Testing Language Ability in the Classroom, Newbury House, Rowley Massachusetts.

Comrey, A.L., 1973, A First Course in Factor Analysis, Academic Press, New York.

Comrey, A.L. and E.A. Levonian, 1958, A comparison of three point coefficients in factor analysis of MMPI items, Educational and Psychological Measurement, Vol. 18, 739-755.

Cronbach, L.J., 1951, Coefficient Alpha and the internal structure of tests, Psychometrica, 297-334.

Cronbach, L.J., 1970, Essentials of Pshychological Testing (3rd Edition), Harper and Row, New York.

Culhane, T., C. Klein-Braley & D. Stevenson (eds.), 1982, Practice and Problems in Language Testing: Procedings of the Fourth International Language Testing Symposium of the Interuniversitare Sprachtestgruppe, University of Essex, Colchester.

Cummins, J., 1979, Cognitive/academic language proficiency, linguistic interdependence , the optimum age question and some other matters, Working Papers in Bilingualism, Vol. 19, 197-205.

Cummins, J., 1983, Language proficiency and academic achievement, In Oller, J.W. (ed), 1983.

Cummins, J. and M. Swain, 1983, Analysis-by-rhetoric: reading the text or the reader's own projections? - A reply to Edelsky et al., Applied Linguistics, Vol. 4, No. 1, 23-41.

Cureton, E.E. and R.B. D'Agostino, 1982, Factor Analysis: An Applied Approach, Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Cziko, G.A., 1983, Some problems with empirically-based models of communicative competence, Applied Linguistics, Vol. 5, No. 1.

Cziko, G.A., 1983, Psychometric and edumetric approaches to language testing, In Oller, J.W. (ed), 1983.

Davies, A., 1985, Follow my Leader: Is That What Language Tests Do?, In Lee, Y.P. et al. (eds.), 1985.

Davies, A., 1978, Language testing, Language Teaching and Linguistics Abstracts, Vol. 11, 145-159.

Davies, A., 1979, Language testing, Language Teaching and Linguistics Abstracts, Vol. 11, 215-231.

Davies, A., 1977, The construction of language tests, In Allen, J.P.B. and A. Davies (eds), 1977.

Davies, A. (ed), 1968, Language Testing Symposium: A Psycholinguistic Approach, Longman, London.

Deale, R.N., 1975, Assessment and Testing in Secondary School, Evans/Methuen, London.

Dielman, T.E., R.B. Cattell, and A. Wagner, 1972, Evidence on the simple structure and factor invariance achieved by five rotational methods on four types of data, Multivariate Behavioural Research, Vol. 7, 223-232.

Eastman, H.T., and W.J. Krzanowski, 1982, Cross-validatory choice of the number of components from a principal components analysis, Technometrics, Vol. 24, 73-77.

Ebel, R.L., 1972, Essentials of Educational Measurement, Prentice-Hall, Englewood Cliffs, N.J..

Ebel, R.L., 1983, The practical validation of tests of ability, Educational Measurement: Issues and Practice, Volume 2: No. 2, 7 - 10.

Edelsky, C., et al., 1983, Semilingualism and language deficit, Applied Linguistics, Vol. 4, No. 1, 1-22.

Eisenck, H.J., 1977, Personality and factor analysis: a reply to Guildford, Psychological Bulletin, Vol. 84, 405-411.

Ek, J.A. van, 1975, The Threshold Level, The Council of Europe.

Ek, J.A. van and L.G. Alexander,

Falvey, M. and M. Milanovic, 1984, The design and implementation of performance objectives for communicative language teaching and their implications for testing, Paper presented at the 19th Annual SEAMEO Regional Seminar.

Fanslow, J.F., and R.L. Light, (eds.), 1977, Bilingual, ESOL and Foreign Language Teacher Preparation: Models, Practices, Issues, TESOL, Washington D.C..

Farhady, H., 1981, Testing Funcional ESL in ESP Contexts, Unpublished PhD. Thesis, UCLA.

Farhady, H., 1982, Measures of language proficiency from the learner's perspective, TESOL Quarterly, Vol. 16, No. 1.

Farhady, H., 1983, On the plausibility of the unitary language proficiency factor, In Oller, J.W. (ed), 1983.

Farhady, H., 1983, New directions in ESL proficiency testing, In Oller, J.W. (ed), 1983.

Farhady, H., 1983, The disjunctive fallacy between discrete-point and integrative tests, In Oller, J.W. (ed), 1983.

Finnochiaro, M. and S. Sato, 1983, Foreign Language Testing: A Practical Approach, Regents Publishing Company, New York.

Gardner, R.C., and W.E. Lambert, 1965, Language aptitude, intelligence and second language achievement, Journal of Educational Psychology, Vol. 56, 191-199.

Giles, H. and P.F. Powesland, 1975, Speech Style and Social Evaluation, Academic Press, London.

Glaser, R., 1963, Instructional technology and the measurement of learning outcomes, American Psychologist, Volume 18, 519-521.

Glaser, R. and A.K. Nitko, 1971, Measurement in learning and instruction, In Thorndike, R.L. (ed.), 1971.

Gorsuch, R.L., 1974, Factor Analysis, Saunders, Philadelphia.

Gorsuch, R.L., 1970, A comparison of biquartimin, maxplane, promax and varimax, Educational and Psychological Measurement, Vol. 30, 861-.

Guildford, J.P., 1973, Theoretical issues and operational-informational psychology, In Royce, J.R. (ed), 1973.

Guildford, J.P. and B. Fruchter, 1973, Fundemental Statistics in Psychology and Education, McGraw-Hill, New York.

Gumperz, J.J. and D. Hymes, (eds.), 1970, Directions in Sociolinguistics, Holt Rinehart and Winston, New York.

Guttman, L., 1954, Some necessary conditions for common factor analysis, Psychometrica, Vol. 19 (2), 149-.

Harman, H.H., 1976, Modern Factor Analysis, University of Chicago Press, Chicago.

Harris, C.W., 1962, Some Rao-Guttman relationships, Psychometrica, Vol. 27, 247-.

Harris, C.W.,(ed), 1963, Problems in Measuring Change, University of Wisconsin Press, Madison.

Harris, D.P., 1969, Testing English as a Second Language, McGraw-Hill, New York.

Harrison, A., 1983, A Language Testing Handbook, McMillan, London.

Hatch, E. and H. Farhady, 1982, Research Design and Statistics for Applied Linguistics, Newbury House, Rowley, Massachusetts.

Hawkey, R., 1982, An Investigation of Inter-relationships Between Cognitive/Affective and Social Factors and Language Learning, Unpublished PhD Thesis, University of London.

Heaton, J.B., 1975, Writing English Language Tests, Longman, London.

Heaton, J.B. (ed.), 1982, Language Testing, Modern English Publications, Hayes, Middlesex.

Hill, K.T., 1983, Interfering effects of test anxiety on test performance: a growing educational problem and solutions to it, Illinois School Research Development, Volume 20, 8-19.

Hills, M., 1977, Review of factor analysis, Applied Statistics, Vol. 26, 329-340.

Hinofotis, F.B., 1983, The structure of oral communication in an educational envirionment: a comparison of factor analytic rotational procedures, In Oller, J.W. (ed), 1983.

Holley, J.W. and J.P. Guildford, 1964, A note on the G-index of agreement, Educational and Psychological Measurement, Vol. 24, 749-753

Holt, J., 1970, The Underachieving School, Pitman.

Holt, M., 1981, Evaluating the Evaluators, Hodder and Stoughton, London.

Hutchinson, T. and A. Waters, 1987, English for Specific Purposes: A Learning-centred approach, Cambridge University Press, Cambridge.

Hopkins, C.D. and R.L. Antes, 1985, **Classroom Measurement and Evaluation** (2nd Edition), F.E. Peacock Publishers, Itasca, Illinois.

Howatt, A.P.R., 1984, **A History of English Language Teaching**, Oxford University Press, Oxford.

Hosely, D. and K. Meridith, 1979, **Inter- and intra-test correlates of the TOEFL**, TESOL Quarterly, Vol. 13, 209-217.

Hymes, D., 1970, **On Communicative Competence**, In Gumperz, J.J. and D. Hymes, (eds.), 1970.

Ingram, E., 1974, **Language Testing**, In Allen, J.P.B. and S.P. Corder, (eds.), 1974.

Ingram, E., 1968, In Davies, A. (ed.), 1968.

Johnson, K., 1982, **Communicative Syllabus Design and Methodology**, Pergamon, Oxford.

Kaiser, H.F., 1963, **Image analysis**, In Harris, C.W. (ed), 1963.

Kim, J. and C.W. Mueller, 1982, **Introduction to Factor Analysis: What it is and how to do it**, Sage Publications, California.

Kim, J. and C.W.Mueller, 1981, **Factor Analysis: Statistical Methods and Practical Issues**, Sage Publications, California.

Klein-Braley, C., 1983, **A cloze is a cloze is a question**, In Oller, J.W. (ed), 1983.

Klein-Braley, C., 1985, **A cloze-up on the C-Test: a study in the construct validation of authentic tests**, Language Testing, Vol. 2, No. 1.

Klein-Braley, C. and D.K. Stevenson, (eds.), 1981, **Practice and Problems in Language Testing 1: Proceedings of the First International Language Testing Symposium of the Interuniversitare Sprachtesstgruppe**

Lado, R., 1961, **Language Testing**, Longman, London.

Lawley, D.N. and A.E. Maxwell, 1971, **Factor Analysis as a Statistical Method**, Butterworth, London.

Lee, Y.P., 1985, **Investigating the Validity of the Cloze Score**, In Lee, Y.P. et al. (eds.), 1985.

Lee, Y.P. et al. (eds), 1985, **New Directions in Language Testing**, Pergamon, Oxford.

Littlewood, W., 1981, Communicative Language Teaching, Cambridge University Press, Cambridge.

Lofgren, H., 1969, Measuring proficiency in the German language: a study of pupils in grade 7, Didakometry No. 25, Malmo, Sweden: School of Ed..

Lord, R. and H.N.L. Cheng, (eds.), 1987, Language Education in Hong Kong, Chinese University Press, Hong Kong.

Low, G.D. and Y.P. Lee, 1985, How Shall a Test be Referenced, In Lee, Y.P. et al. (eds), 1985.

Mackay, R. and J.D. Palmer (eds.), 1981, Languages for Specific Purposes: Program Design and Evaluation, Newbury House, Rowley, Massachusetts.

Madsen, H.S., 1982, Determining the debilitative impact of test anxiety, Language Learning, Volume 32, 133-43.

Maxwell, A.F., 1977, Multivariate Analysis in Behavioural Research, Chapman and Hall, London.

McCormick, R. & James, M., 1983, Curriculum Evaluation in Schools, Croom Helm, London.

Mehrens, W.A. and I.J. Lehmann, 1984, Measurement and Evaluation in Education and Psychology (3rd Edition), Holt, New York.

Milanovic, M., 1987, Large-scale language testing, In Lord, R. and H.N.L. Cheng, (eds.), 1987.

Milanovic, M., 1985, Communicative language testing: profiles and certification, Paper presented at the 19th Annual TESOL Convention, New York.

Moller, A., 1983, A Study in the Validation of Proficiency Tests of English as a Foreign Language, Unpublished Ph.D. dissertation, University of Edinburgh.

Moller, A., 1983, Reaction to the Morrow paper (2), In Alderson, C.J. and A. Hughes, (eds.), 1983, 38-44.

Morrow, K., 1979, Communicative language testing: revolution or evolution, In Brumfit, C.J. and K. Johnson (eds.), 1979.

Mulaik, S.A., 1972, The Foundations of Factor Analysis, McGraw-Hill, New York.

Munby, J., 1978, Communicative Syllabus Design, Cambridge University Press, Cambridge.

Nunnally, J.C., 1978, Psychometric Theory, McGraw-Hill, New York.

Oller, J.W., 1973, Discrete-point tests versus tests of integretive skills, In Oller, J.W. and J.C. Richards (eds.), 1973.

Oller, J.W. (ed.), 1983, Issues in Language Testing Research, Newbury House, Rowley, Massachusetts.

Oller J.W. and F.B. Hinofotis, 1980, Two mutually exclusive hypotheses about second language ability: in divisible or partially divisible competence, In Oller, J.W. and K. Perkins (eds), 1980.

Oller, J.W. and J.C. Richards, (eds.), 1973, Focus on the Learner: Pragmatic Perspectives for the Language Teacher, Newbury House, Rowley, Massachusetts.

Oller, J.W. and K. Perkins (eds), 1980, Research in Language Testing, Newbury House, Rowley, Massachusetts.

Pimsleur, P., 1966, Language Aptitude Battery, Harcourt Brace Jovanovich, New York.

Pimsleur, P., R.P. Stockwell, and A.L. Comrey, 1962, Foreign language learning ability, Journal of Educational Psychology, Vol. 53, 15-26.

Pollitt, A. and C. Hutchinson, 1987, Calibrating graded assessments: Rasch partial credit analysis of performance in writing, Language Testing, Vol. No.

Popham, W.J., 1981, Modern Educational Measurement, Prentice-Hall, Englewood Cliffs, New Jersey.

Pratt, D., 1980, Curriculum: Design and Development, Harcourt Brace Jovanovich, New York.

Priestly, M., 1982, Performance Assessment in Education and Training: Alternative Techniques, Educational Technology Publications, Englewood Cliffs, New Jersey.

Rea, P.M., 1985, Language Testing and the Communicative Language Teaching Curriculum, In Lee, Y.P. et al. (eds.), 1985.

Rivers, W.M., 1968, Teaching Foreign Language Skills, University of Chicago Press, Chicago.

Robson, C., 1979, Experiment, Design and Statistics in Psychology, Penguin, Suffolk.

Royce, J.R., 1973, Multivariate Analysis and Psychological Theory, Academic Press, London.

Rummel, R.J., 1979, Applied Factor Analysis, Northwestern University Press, Evanston.

Rutter, M. et al., 1979, Fifteen Thousand Hours: Secondary Schools and their Effects on Children.

Sang, F. et al., 1986, Models of second language competence: a structural equation approach, Language Testing, Vol. 3, 54-79.

Schmidt, R.W., 1983, Interaction, acculturation, and the acquisition of communicative competence: a case study of an adult, In Wolfson, N. and E. Judd, (eds.), 1983.

Sinclair, J. and M. Coulthard, 1975, Towards an Analysis of Discourse, Oxford University Press, Oxford.

Shohamy, E., 1983, The stability of the oral proficiency assesment on the oral interview testing procedure, Language Learning, Vol. 33, 527-40.

Shohamy, E., 1984, Does the testing method make a difference? The case of reading comprehension, Language Testing, Vol.

Spearman, C., 1904, General intelligence, objectively determined and measured, American Journal of Psychology, 15:201-293.

Spielberger, C.D., 1966, Theory and research on anxiety, In Spielberger, C.D. (ed.), 1966.

Spielberger, C.D. (ed.), 1966, Anxiety and Behavior, Academic Press, New York.

Spolsky, B., 1975, Language testing: Art or science, Paper presented at the fourth AILA International Congress, Stuttgart.

Spolsky, B., 1978, Advances in Language Testing Series 2: Approaches to Language Testing, Center for Applied Linguistics, Arlington, Virginia.

Stanley, J.C. and Hopkins, K.D., 1972, Educational and Psychological Measurement and Evaluation, Prentice-Hall, Englewood Cliffs, New Jersey.

Stansfield, & Hanson,, 1982, The relationship of field dependent-independent cognitive styles to foreign

language achievement, Language Learning, Vol. 31, No. 2.

Stansfield, & Hanson,, Field dependence-independence as a variable in second language cloze test performance, TESOL Quarterly.

Steltmann, K., 1979, Faktoren der Fremdsprachenleistung, Kastellaun, Bonn.

Stern, H.H., 1983, Fundemental Concepts of Language Teaching, Oxford University Press, Oxford.

Stevenson, D.K., 1981, Beyond faith and face validity: the multitrait-multimethod matrix and the convergent and discriminant validity of oral proficiency tests, In Palmer. A.S. et al. (eds.), 1981.

Stevenson, D.K., 1985, Pop Validity and Performance Testing, In Lee, Y.P. et al. (eds), 1985.

Stevenson, D.K., and U. Riewe, 1982, Teachers attitudes towards language tests and testing, In Culhane, T. et al. (eds.), 1982, 146-155.

Swain, M., 1985, Large-scale Communicative Language Testing: A Case Study, In Lee, Y.P. et al. (eds), 1985.

Tapp G.S. and J.R. Barclay, 1974, Convergent and discriminant validity of the Barclay Classroom Climate Inventory, Educational and Psychological Measurement, Vol. 34, 2, 439-447.

Taylor, W.L., 1953, Cloze procedure: A new tool for measuring readability, Journalism Quarterly, Vol. 30, 415-433.

The Royal Society of Arts, Examination in the Communicative Use of English, Royal Society of Arts.

Thorndike, R.L. and E.P. Hagen, 1977, Measurement and Evaluation in Psychology and Education (4th Edition), John Wiley and Sons, New York.

Thorndikek, R.L. (ed.), 1971, Educational Measurement (Second Edition), American Council on Education, Washington D.C..

Tuckman, R., 1971, Conducting Educational Research, Harcourt Brace Jovanovich, New York.

Tung, P., 1985,

Underhill, N., 1982, The great reliability/validity trade off: problems in assessing the productive skills, In Heaton J.B. (ed.), 1982.

Valette, R., 1967, Modern Language Testing - A Handbook (1st edition), Harcourt Brace Jovanovic, New York.

Valette, R., 1977, Modern Language Testing - A Handbook (2nd edition), Harcourt Brace Jovanovic, New York.

Valette, R. and R.S. Disick, 1972, Modern Language Performance objectives and Individualization: A Handbook, Harcourt Brace Jovanovic, New York.

Van Ek, J.A., 1975, Threshold level English, Pergamon, Oxford.

Vollmer, H.J., 1981, Why are we interested in general language proficiency? In Klein-Braley, C. and D.K. Stevenson, (eds.), 1981.

Vollmer, H.J. and F. Sang, 1983, Competing hypotheses about second language ability: a plea for caution, In Oller, J.W. (ed), 1983.

Walsh, B.A. and N.E. Betz, 1985, Tests and Assessment, Prentice-Hall, Englewood Cliffs, New Jersey.

Waters, A., (ed.) 1983, Lancaster Papers in English Language Education, Vol. 5, Pergamon, Oxford.

Weir, C.J., 1983, Identifying the Language Problems of Overseas Students in Tertiary Education in the United Kingdom, Unpublished PhD Thesis, University of London.

Weir, C.J., 1981, Reaction to the Morrow paper (1), In Alderson, C.J. and A. Hughes, (eds.), 1981, 26-37.

Wesche, M.B., 1987, Second language performance testing: the Ontario test of ESL as an example, Language Testing, Vol. 4, No. 1, 28-47.

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

Wilkins, D.A., 1976, Notional Syllabuses, Oxford University Press, Oxford.

Wolfson, N. and E. Judd, (eds.), 1983, Sociolinguistics and Language Acquisition, Newbury House, Rowley, Massachusetts.

Woods, A., P. Fletcher and A. Hughes, 1986, **Statistics in Language Studies**, Cambridge University Press, Cambridge.