Studies in Language Testing 4

The Development of IELTS

A study of the effect of background knowledge on reading comprehension

Caroline Clapham

Series Editor
Michael Milanovic

UNIVERSITY *of* CAMBRIDGE
ESOL Examinations

CAMBRIDGE
UNIVERSITY PRESS

The development of IELTS: A study of the effect of background knowledge on reading comprehension

STUDIES IN LANGUAGE TESTING...4
Series editor: Michael Milanovic

Also in this series:

An investigation into the comparability of two tests of English as a foreign language: The Cambridge-TOEFL comparability study
Lyle F. Bachman, F. Davidson, K. Ryan, I-C Choi

Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC) Cambridge and Arnhem, 1993

Test taker characteristics and performance: A structural modeling approach
Anthony John Kunnan

The development of IELTS: A study of the effect of background knowledge on reading comprehension

Caroline Clapham

Lancaster University, UK

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo, Delhi

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521567084

First published 1996
Re-issued in this digitally printed version 2009

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-56199-0 hardback
ISBN 978-0-521-56708-4 paperback

In memory of DCT

Contents

Series Editor's note

Background

This volume of Studies in Language Testing, the fourth in the series, is based on doctoral work carried out by the author, Caroline Clapham, over a number of years and supported by UCLES. Her particular interest lay in the testing of reading, which is one of the four skills-based modules that make up IELTS. She investigated the ESP beliefs underlying the design of the reading components of ELTS and an early version of IELTS. Her work shows how difficult it is both to classify students according to their background knowledge, and to select reading passages which are genuinely specific for people in any one subject area. She suggests, therefore, that it is fairer for students if they all take a single academic reading module. The conclusions are of general importance to the designers of EAP proficiency tests. The investigating methods used in the study for assessing the appropriacy of the reading passages, and the resulting demonstration that text selectors are often unable to estimate the specificity of the texts they choose, will be of interest to all those who have to select reading texts for testing purposes or for research into reading in a second language.

As a point of information, the International English Language Testing System (IELTS) as it now stands provides an assessment of whether candidates are ready to study or train through the medium of English. It is recognised widely as a language requirement for entry to courses in further and higher education, and is readily available, being offered 'on demand' at test centres around the world. IELTS is jointly managed by The University of Cambridge Local Examinations Syndicate (UCLES), the British Council and IDP Education Australia Limited.

IELTS can be taken at around 200 approved test centres in well over 100 different countries. The test is administered centrally by UCLES but the approved centres, most of which are British Council or IDP Offices, supervise the local administration of the test and ensure the provision of qualified and trained examiners. IELTS is not held on set dates during the year but is conducted according to demand. Most centres conduct a testing session at least once a month and more often at peak times.

On a historical note, the original English Language Testing Service (ELTS) was developed in the late seventies by British Council staff, headed by Brendan Carroll, and became operational in 1980. It was one of the first language tests to take into account the communicative movement and drew heavily on the work of Munby (1978). In 1986, a decision was made to revise this test and a team at Lancaster University, headed by Charles Alderson, took on the project. The team was joined in 1987 by David Ingram, who acted as a representative of the International Development Program of Australian Universities and Colleges (IDP). At this time the test was renamed the International English Language Testing System (IELTS). The revision team drew on a variety of sources to guide them. These included: the ELTS Validation Study, carried out by Alan Davies and Clive Criper, the views of ELTS users, EAP teachers, language testers and applied linguists.

The ELTS Validation Study reported strengths and weaknesses on a number of fronts. From the point of view of practicality, ELTS was felt to be rather cumbersome with its six academic modules. On the other hand, it was found to have high face validity for exactly the same reason. Flaws in the test design were attributed to weaknesses in the theory of English for Specific Purposes (ESP), which guided test development to some extent. Such findings along with extensive consultation guided the revision team. IELTS was released in late 1989. It had a focus on the four skills and made use of three academic reading modules and a single general training reading module. The Writing module was also subject specific and linked to the reading modules. The speaking and listening modules were general and taken by all candidates.

Consistent with UCLES policy of on-going validation, IELTS was carefully monitored in the early 1990s and by 1992, it was decided to modify the test. The monitoring suggested that the three academic modules should be reduced to one. Caroline Clapham's work reported in this volume informed this decision to some extent. The revised version of IELTS was introduced in April 1995. Materials for IELTS are now written by teams of item writers in the UK and Australia. All materials are pretested and calibrated to the IELTS scale. The test itself provides a profile of ability to use English. A score in each of the four modules or skills, and an overall score, are recorded as levels of ability, called Bands.

Assessment of performance in IELTS does not depend on reaching a fixed pass mark. It depends on how the candidate's ability in English relates to the language demands of courses of study or training. The appropriate level required for a given course of study or training is ultimately something which institutions must decide in the light of knowledge of their own course and their experience of overseas students taking them. There are six modules in IELTS. All candidates must take four modules, one in each of the four skills. All candidates take the same Listening and Speaking Modules, whereas there is a choice of Reading and Writing Modules with either a general training or academic focus.

Since its 1995 revision IELTS has adhered to a clearly stated code of practice. This has required the implementation and maintenance of systems and procedures designed specifically to validate the test, evaluate the impact of the test and provide relevant information to test users.

It is recognised that as a test provider, UCLES has an impact on educational processes and on society in general. This impact operates on at least two levels in terms of:

i education and society in general
ii people who are directly affected by tests and their results

We believe it to be important to be able to investigate the educational impact IELTS has within the context that it is used. As a point of principle, test developers must operate with the aim that their tests will not have a negative impact and, as far as possible, strive to make it positive. In general terms, this can be achieved through the development and presentation of test specifications and detailed syllabus designs, and provision of professional support programmes for institutions and individual teachers/students.

It is anticipated that positive educational impact in the case of IELTS can be achieved through the following practices:

the identification of suitable experts within any given field to work on all aspects of test development;
the training and employment of suitable experts to act as question/item writers in test production;
the training and employment of suitable experts to act as examiners.

Procedures are required to collect information which allows impact to be estimated and attention is being focused on the following areas through routine data collection or further research:

who is taking the test (i.e. a profile of the candidates);
who is using the test results and for what purpose;
who is teaching towards the test and under what circumstances;
what kinds of courses and materials are being designed and used to prepare candidates;
what effect the test has on public perceptions generally (e.g. regarding educational standards);
how the test is viewed by those directly involved in educational processes (e.g. by students, test takers, teachers, parents, etc.);
how the test is viewed by members of society outside education (e.g. by politicians, businessmen etc.).

It is hoped that aspects of this research will be reported on in this series.

Preface

The purpose of this research is firstly to investigate the ESP claim that tertiary level ESL students should be given reading proficiency tests in their own academic subject areas, and secondly to study the effect of background knowledge on reading comprehension. The study is set against a background of recent research into reading in a first and second language, and emphasises the impact that schema theory has had on this.

Students took two versions of the International English Language Testing System (IELTS) test, which has reading modules in three different academic subject areas. Analyses of variance showed that the reading subtests varied in their subject specificity: some were suitable for students in the relevant academic field, others were either too general or too specific. A Rasch analysis of the items revealed little bias against students who took an inappropriate reading module, and an investigation of the test content using a version of Bachman's Test Methods Characteristics scale showed that the test items did not seem to affect test specificity. Variation in the appropriacy of the reading passages was found to be partly due to differences in rhetorical function, partly to uncertainty among EAP teachers about concepts relating to academic and topic specificity, partly to the extent of students' familiarity with the subject areas, and partly to the fact that students vary widely in their background reading and cannot be accurately placed into three distinct subject areas.

Further analysis suggested that the relative importance of language proficiency and background knowledge in reading comprehension depended on the specificity of the reading passages, and an investigation into whether language ability affected the students' use of background knowledge supported the hypothesis that there is a threshold level below which learners have difficulty making use of this knowledge.

The book concludes by considering the implications of the findings for future test construction and research into reading.

Acknowledgements

I have to thank so many people for help with the research required for this book that I cannot do justice to everybody's contribution.

The project would not have been possible without the members of the ELTS Revision Project, Charles Alderson, Peter Hargreaves, David Ingram, John Foulkes and Gill Westaway, with whom I worked on the design and construction of the trial IELTS, and the many teachers and testers who wrote the Grammar Test and the Trial and Exemplar Versions of the Reading Modules.

The following people very kindly either administered or arranged for me to administer the tests at their language centres: Don Dunmore, Glenn Fulcher, Gail Langley, Kenneth James, Philip King, John Read, Margaret Rutherford, Chris Shaw, Philip Shaw, Anat Stavans, Susan Taylor, Alan Tonkyn, Carolyn Walker, Dianne Wall and Rosemary Zahn. I must also thank the 800 or so students who, in many cases voluntarily, spent about three hours of their time taking what proved to be very demanding tests.

I am most grateful to Carolyn Turner for providing me with the McGill system of clausal analysis, Michael Milanovic for arranging for the test papers to be printed and Lyle Bachman for allowing me to use an adaptation of his Test Methods Characteristics Rating Scale, and for answering all my queries about it. I also owe particùlar thanks to Barbara Adams, Joan Allwright and Nicki McLeod who spent several hours using the rating scale to assess the reading tests.

So many people have advised me about analysing my results that I cannot thank them all personally, but I would particularly like to thank Brian Francis of the Department of Applied Statistics at this university. I should also like to thank Tim McNamara, Sarah Cushing Weigle, Alastair Pollitt, Neil Jones and Mary Shegl for advice about identifying item bias, and Grant Henning, Fred Davidson and Dorry Kenyon for general advice over the years.

In particular I am indebted to Charles Alderson who encouraged me to start on the project, and inspired me to move in interesting directions.

Finally I must thank my daughter, Phoebe, for reading through the book, my son, Tom, for formatting the tables, and my husband, Christopher, for his advice and encouragement.

Caroline Clapham

1 Introduction

In recent years there has been increasing agreement among testers that language proficiency tests should, where possible, be related to candidates' future language needs. For example, if doctors are to be tested for their capacity to use English in an English-speaking hospital, it is considered only sensible to test them on the kinds of English that are used in the ward and the consulting room. Similarly, future air traffic controllers might be tested on the English needed in the control tower. No one is likely to question the good sense of such English for Specific Purposes (ESP) testing. The issue, however, becomes less clear cut when the proficiency test is aimed at a less sharply defined audience, such as students proposing to embark on tertiary education. Many testers consider that a language proficiency test for such students should contain samples of the kinds of language tasks required of them in their academic work, but it is not clear how much, if at all, these tasks differ from discipline to discipline, nor how much the subject matter of the test should vary according to the discipline of the examinee. The question here is whether there should be separate tests for students in the different academic disciplines, or whether all students should take a single test battery. There is some evidence to show that the language tasks in different academic disciplines are sufficiently similar for one set of test tasks to be appropriate for all (see Weir 1983 and Chapter 4 below), but it is not clear whether the subject matter of the tests should be different.

The results of research into the effects of field specific reading tests on EAP students' test performance have been somewhat contradictory, and no conclusive evidence has been produced either for or against the use of ESP tests. This book, therefore, reports on a large scale study into the effects of giving subject specific reading tests to future university students. The main aim of this study is to see whether an ESP approach to testing the reading proficiency of academic students is appropriate and feasible, and the secondary aim is to consider the effect of background knowledge on reading comprehension.

This first chapter briefly introduces ESP teaching and testing, and describes recent research into whether domain specific background knowledge affects test scores in English comprehension tests. Chapters 2 and 3 review research into the effect of background knowledge on reading in a first and a second language. Chapter 4 describes the construction of the reading component of the International English Language Testing System (IELTS) test, and Chapter 5 reports on a pilot study into the effect of subject area on test performance. Chapter 6 presents the research questions for the main study, and describes the tests, the questionnaire and the student sample. Chapter 7 describes a replication of the pilot study using a different set of tests, and gives the results of other investigations into the effect of subject area on test performance. These studies show that the reading subtests vary in their subject specificity, and Chapters 8 and 9 discuss the reasons for this variation. Chapter 10 looks at the effects of level of language proficiency on the use of background knowledge in reading, and also compares the effects of language proficiency and background knowledge on test scores. Chapter 11 summarises the main research findings, discusses their implications, and makes proposals for further research.

English for Specific Purposes

The main drive behind the introduction of ESP was practical rather than theoretical. With the rapid increase after the war in the importance of English for education, technology and commerce, increasing numbers of people around the world needed to learn English for clearly defined reasons such as reading academic textbooks or transacting business (see Hutchinson and Waters 1987). These changes coincided with developments in communicative methods of teaching, and led ESP course designers to base their materials on texts and activities which were tailored to suit students with specific linguistic needs.

Although there had been some awareness from the 1920s of the fact that learners in different jobs needed different kinds of language (see Widdowson 1983), the ESP movement only came into existence in the 1960s. In its early stages ESP researchers focused on register analysis – the analysis of sentence-level grammatical and lexical features to see what the distinctive features were between texts in different subject areas. These analyses often took the form of frequency counts of structures or verb forms (see for example, Barber 1962/1985), or clausal analysis (see Huddleston 1971), and as Swales (1985:59) said, although such analyses had descriptive validity, they had little explanatory force. Widdowson (1979:55) pointed out that the fact that English scientific texts had a relatively high proportion of some syntactic structures and a relatively low proportion of others did not reveal anything about scientific discourse as a whole, and Hutchinson and Waters (1987:10) said that few systematic differences were found between scientific and 'General English' texts. As the focus of linguistic research changed from being sentence-based to include research into how sentences combine to produce meaning, and as the increasing importance of sociolinguistics led to the study of language variation in different contexts, researchers such as Widdowson (1979) and Trimble (1985) began to apply rhetorical or discourse analysis to discover the main characteristics of Scientific and Technical English (EST) texts, and to see if there were differences between these texts and non-EST ones. Trimble, for example, built his studies and teaching round three rhetorical concepts:

a) the nature of the EST paragraph;
b) the rhetorical techniques most commonly used in written EST discourse; and
c) the rhetorical functions most frequently found in written EST discourse (Trimble, 1985:14).

So much ESP research has focused on EST that it is easy to think of ESP and EST as synonymous. However, EST is an offshoot of ESP, on a par with, for example, English for Social Scientists. Since there is some disagreement among ESP teachers and researchers about the hierarchy of ESP terms, I will explain how the term 'ESP' is used in this book and how it relates to English for Academic Purposes (EAP).

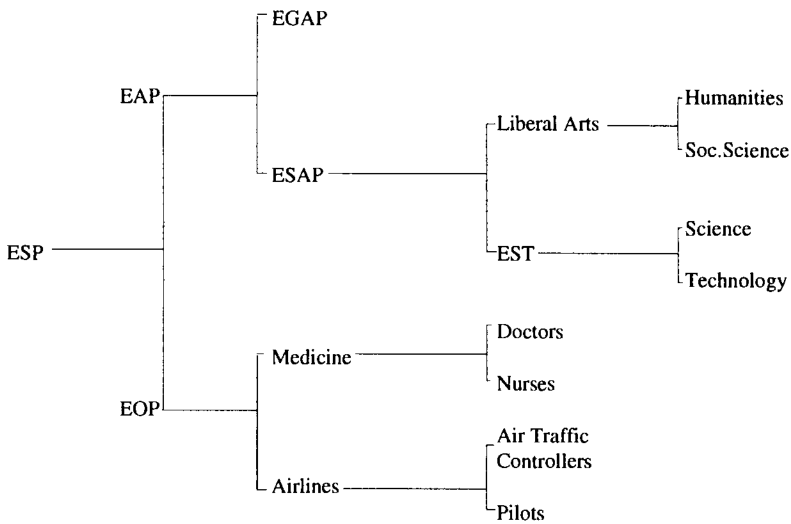
There seem to be two favourite ways of treating the concepts of ESP and EAP. Some ESP teachers consider that since many EAP courses are designed to suit students in all academic disciplines, EAP is too broad in scope to be considered a branch of ESP. These teachers think of it as the superordinate category from which spring increasingly specific types of ESP (see Jordan 1989). However, this takes no account of other types of ESP such as English for Occupational Purposes (EOP). Since EAP is itself a type of ESP because it is concerned with the English required for a specific purpose, that of studying at universities and colleges, it is more usual to think of ESP as the superordinate term, with EAP and EOP branching from it (see Jordan 1989 and Robinson 1991). EAP courses can be divided into those for English for General Academic Purposes (EGAP) and those for English for Specific Academic Purposes (ESAP) (see Blue 1993). ESAP courses can range from broad groupings of subjects, such as EST or Liberal Arts, to ones which are so highly specific that they are suitable only for single students or for small groups in one narrowly defined discipline. ('Discipline', 'Field of study' and 'Subject area' are used interchangeably in this book.) Figure 1.1 shows how the terms are used here, and gives examples of possible courses at the different levels of the hierarchy.

In recent years the focus of academic textual analysis has broadened to take account of different genres of writing ranging from academic articles to article abstracts and citations. Genre analysis studies not only the composition of texts, but also the roles that those texts play. According to Swales (1990), the academic world consists of a series of 'discourse communities', each of which uses a shared set of genres in order to achieve a common set of purposes.

> *By 'genre' is meant a typified socially recognised form that is used in typified social circumstances. It has characteristic features of style and form that are recognised, either overtly or covertly, by those who use the genre. Thus for example, the research article has a known public purpose, and has conventions about layout, form and style that are to a large degree standardised.*
> (Dudley-Evans 1987)

Figure 1.1

The ESP Hierarchy with Examples of Courses



EGAP = English for General Academic Purposes
ESAP = English for Specific Academic Purposes

I shall refer to genre analysis again in Chapter 9.

In a parallel development to the analysis of texts in different language domains, there was a growing awareness among teachers of the importance of students' individual needs, and during the late 1970s ESP course designers started to carry out 'needs analyses' of their students' future linguistic requirements (see for example, Candlin, Leather and Bruton 1976). These needs analyses were often expressed in terms of notions and functions (Van Ek and Trim 1975/1991 and Wilkins 1976). The most celebrated model for such a needs analysis was described by Munby (1978) in his *Communicative Syllabus Design* in which he presented a system for 'devising appropriate syllabus specifications from adequate profiles of communication needs' (Munby 1978:3). These profiles included the purposes of communication, the communicative settings, and the language skills, functions and structures required. This 'communication needs processor' was very influential for a few years, and was used as the basis for the English Language Testing Service (ELTS) test, and also for Weir's (1983) needs analysis for the Test of English for Educational Purposes (TEEP). (These are described further in Chapter 4.)

For a while this needs analysis approach was central to ESP. However, Hutchinson and Waters described the advent of the needs analysis era as 'a false dawn' (1987:12). They considered that the main problem with a Munby-style needs analysis was that it did not take account of the students themselves, their existing skills, their interests, and their language learning wishes. No account was taken of psycholinguistic elements, and the background knowledge that students brought to their language learning was ignored (see also Skehan 1984 and Alderson 1988b).

The Munby model was also criticised for being impractical (Mead 1982), for not being based on sound theoretical precepts, and conversely for not having been empirically verified (see Davies 1981a; Skehan 1984 and Alderson 1988a).

Munby did not intend his model to be used as the basis for test specifications, and it proved inadequate for this purpose. As Alderson (1988a) pointed out, any needs analysis based on it would produce a huge list of needs, most of which would not be convertible into test items, and since there was no indication of the respective importance of the different needs and skills it would be impossible to make a principled selection from those needs and skills which were testable.

By referring to the needs analysis era, Hutchinson and Waters (1987) implied that it was a thing of the past. However, this is not the case. Munby's model may not have heralded a new dawn, but the fact that it was so detailed and explicit meant that it was possible to level highly specific criticism against it, and thus to advance our understanding of what an adequate needs analysis should contain. We now know that such analyses can become too detailed, and also, paradoxically, too limited in scope. However, this does not mean they are unnecessary.

ESP Testing

The rapid expansion in ESP teaching was not accompanied by a similar increase in ESP testing.

At the time that ELTS was launched in 1980 there had been little or no research into the validity of giving academic students English proficiency tests based on different subject areas, and in a discussion on ESP testing, Alderson (1981) questioned many of the principles behind this approach. He agreed that since different university departments placed different demands on their students, there were some good arguments for including specific tests in an EAP test battery, and he felt that a comparison between performance on academically specific tests and the communicative needs of the relevant area might provide useful diagnostic information. He also accepted the fact that if a test looked appropriate for students in a given discipline it would have high face validity both for students and university lecturers. However, he questioned whether it was possible to produce a test which would be equally suitable for students in all branches of a discipline. For example, he wondered whether it would be possible to have a test for engineers that was of the same level of appropriacy for all engineers, regardless of their specialisation. This highlights one of the main difficulties with ESAP testing.

One of the main questions Alderson asked in 1981 (see also Alderson 1988b), was 'how specific is specific?'. Since it is at present usually impossible to give each student a test which is tailor-made for a unique set of circumstances, any ESP test must be a compromise, and in the case of EAP, where many disciplines must be subsumed under one broad subject area, these areas will cover so wide a field that some students will not fit into any of the groupings. Alderson (1981) cited the example of the student in urban studies who would not know whether to choose a test in science or in social studies.

Alderson (1988b) also asked what was meant by the term 'general text'. General to whom? Were 'general' texts so neutral that their subject matter was unfamiliar to all (see the reference to the Joint Matriculation Board examination on page 8) or were they *intended* to be neutral, but actually based on arts-based topics which might turn out to be more appropriate for arts than for science students?

Until there were answers to the above questions, Alderson wondered how much point there was in having specific EAP tests, since they were time-consuming and expensive to produce, and since it was so difficult to make equivalent tests in different subject areas genuinely parallel. The only way we could know, he said, was to carry out empirical studies (Alderson 1981:133).

Research into ESP testing

Since 1981 there has been some response to Alderson's plea for more research, and there have been several studies into the effect of background knowledge on EAP test performance. Before I describe these, it might be useful to review what we mean by background knowledge. Studies into the effect of background knowledge on ESL reading comprehension can be divided into two kinds, those concerned with world, that is, content and cultural knowledge, and those relating to knowledge of the formal or linguistic structure of texts. In this study I shall be mainly concerned with content knowledge, and in particular with subject or domain specific content knowledge – the knowledge acquired from schooling, interests and hobbies.

Three articles by Alderson and Urquhart (1983, 1985a and 1985b) aroused considerable interest and led to several follow-up studies. These articles described three studies carried out with students attending English classes in Britain in preparation for going to British universities. In each, Alderson and Urquhart compared students' scores on reading texts related to their own field of study with those on texts in other subject areas. In the third study, three groups of students in different disciplines – Business and Economics, Science and

Engineering, and Liberal Arts – took the Social Studies and Technology Modules of the ELTS test. The students' scores on the modules were somewhat contradictory. On the one hand, for example, Science and Engineering students taking the Technology module did better than the Business and Economics students who took the same test, and as well as the Liberal Arts students, although their language proficiency was lower. On the other hand, the Business and Economics students did no better than the Science and Engineering group on the Social Studies module. Since the authors used pairwise tests to assess the results, they were not able to test for the interaction between students and tests. However, they concluded that background knowledge had some effect on test scores, but that this was not consistent, and that future studies should take account of linguistic proficiency and other factors as well.

Koh (1985) had somewhat similar results with three groups of students – two in Science and one in Business Studies at Singapore University. Using analysis of variance to estimate the effect of background knowledge on cloze test results, she found that there was an interaction between student group and test, but that students did not always do best in their own subject areas. The Business students, for example, had their highest scores on the Science text. However, it turned out that half these students had studied Science previously, so it could be that prior knowledge was affecting their scores. The group with the highest language proficiency – one of the Science groups – did consistently better than the other two on all the texts, which were on Business, History, Politics and Science. She concluded that prior knowledge did affect test scores but that ignorance of the subject matter could be compensated for by high linguistic proficiency.

Shoham, Peretz and Vorhaus (1987) rejected the use of cloze in studies of the role of background effect, and used comprehension, referent and vocabulary-in-context questions for their study at Ben Gurion University. They used three-way analysis of variance to analyse their results but once again the results were inconclusive. While students in the Biological and Physical Sciences did better at the scientific texts, the Humanities and Social Science students did not do better on the test in their own subject area. (Peretz and Shoham, 1990 had similar results.) The authors' explanation for this was that the texts were only indirectly related to the students' specialised fields of study, and suggested that this might support Lipson's suggestion (1984) that 'a totally unfamiliar text is often easier to comprehend than a text with a partially familiar content' (Shoham, Peretz and Vorhaus, 1987:86). This contention of Lipson's is indeed radical. If it was supported by further research, there would be an almost unassailable reason for dropping ESP testing from university proficiency tests, since until each student has his or her own tailor-made test, ESP tests will have to be focused on fairly general subject areas, which will inevitably be only partially familiar to post-graduates in many highly specialised fields. If Lipson's idea was taken to its logical conclusion, of course, proficiency tests would have to contain material outside any candidate's experience. The JMB (Joint Matriculation Board) University Test in English for Speakers of Other Languages follows just such an approach, with passages in esoteric subjects such as silver markings and heraldic devices. Item writers have difficulty finding suitable texts and the ensuing materials are often excessively dull.

Hale (1988) commented on the inconclusive results of some of the above studies, and on the small sample sizes of some of them. For his study, he looked at all candidates taking the Test of English as a Foreign Language (TOEFL) over four administrations to see whether with the larger sample sizes there was a more consistent interaction between students' major field area and text content. The reading passages were all aimed at the general reader but were based on a wide range of topics in the arts and sciences. The numbers of candidates in the four sessions ranged from approximately 6,000 to 10,000. For the purposes of the main study, subjects were divided into two groups – Humanities and Social Sciences in one, and Biological and Physical Sciences in the other. Hale used analysis of variance and found that for three of the four test forms the effect of subject area was significant at .001. Students' reading performance was affected by a combination of their major-field area and the nature of the passages, but the effect was not large, possibly, Hale says, because the texts were taken from general sources rather than from subject specific textbooks. Hale's reason for using such large sample sizes was to provide a greater opportunity for any statistically significant effects to be detected. Using large sample sizes certainly has this effect, and once the number of subjects becomes really large almost anything can appear significant. It seems, therefore, that Hale's huge sample sizes may militate against his significant results being as informative as he might have hoped.

A question related to that of the effect of prior knowledge on test results is the question of whether tests in students' own subject areas are better predictors than more general ones. Tan (1990) used regression analysis to see whether familiarity with test content or level of language proficiency was the best predictor of ability in reading comprehension. Undergraduates at the University of Malaya were given 'prior knowledge' tests compiled by their own subject teachers, along with discipline-related cloze reading tests and a form of a 'general' proficiency test, the English Proficiency Test Battery (EPTB). In all subject areas under study – Medicine, Law and Economics – she found that comprehension of a discipline-related text could be predicted by both knowledge of the subject area and by language level, but that language level was the better predictor.

Jensen and Hansen (1995) used multiple regression analysis to compare the effects of prior knowledge and listening ability on university students' performance on academic listening tasks. They collected data over six administrations of a listening test based on an academic lecture. At each administration the students listened to a lecture on either a technical subject (for example, chemistry or biology) or a non-technical one (for example history or social anthropology).

The choice of lecture depended on students' prior knowledge, which was assumed to be appropriate if they said 'yes' to a question asking whether they had studied the topic of the lecture before. A different pair of tests was used for each administration. Jensen and Hansen found that prior knowledge had a significant effect on test scores on all but one of the technical passages, but on only one of the non-technical ones, and that in all cases the students' level of listening ability had a stronger effect than prior knowledge. There was no interaction between level of listening ability and the effect of prior knowledge. We do not know, however, how subject specific the lectures were, although, since they came from introductions to courses, they may have been fairly general in content. Nor do we know how much variation there was in the students' levels of listening ability: if the spread was only small, we would not expect the effect of background knowledge to differ between high and low scoring students. It is interesting that all but one of the significant subject effects related to the technical tests, and this suggests, as I shall be proposing later, that science students are better able to cope with non-science texts than non-scientists are with scientific ones.

Kattan (1990) wanted to see whether it was worth giving university students at Bethlehem University ESP proficiency tests based on the subject in which they were majoring, or 'whether a measure of a more generalised competence would do just as well' (1990:3). She compared the predictive validity of a 'neutral' reading test, with that of two subject specific tests – one for students majoring in English and one for those majoring in Nursing Studies. She correlated students' scores on both the neutral test and their own subject test with their grade point average (GPA) over a period of eighteen months, and found a correlation of .71 (n = 54) between the results of the test for English majors and their GPA. The correlation between the students' neutral test scores and their GPA was only .36. The nurses' correlations (based on only 19 students) were, on the other hand, not significant. The size of the correlation between the English students' subject specific test and their GPA is surprising, since the GPA is presumably based at least as much on subject knowledge as on linguistic proficiency. However, the test contained only sixteen items and its Cronbach alpha reliability index was .54. Since low reliability reduces the potential of a test to correlate with other measures, this suggests that the predictive validity of this English test was either exceptionally good, or that the English major at Bethlehem University places a stronger emphasis on English proficiency than do other majors, and that the students' GPA is partially, therefore, a reflection of their ability to use English. The English GPAs therefore might not be similar in composition to those of other subject majors.

It would be interesting to see whether a repeat study produced comparable results.

Yet another way of assessing the effect of subject area on test performance is to use bias analysis to see whether test items discriminate against students who are not familiar with the subject area of a text. O'Neill, Steffen and Broch (1994) used Differential Item Function (DIF) to see whether the proportion of correct items among students taking TOEFL reading tests was higher when the content of the texts was based on 'home-field' rather than 'non-home-field' subject matter. The researchers used the Educational Testing Service (ETS) DIF program, which is based on the Mantel-Haenszel statistical technique, and compares the odds of two groups answering an item correctly when ability levels are taken into account. The results agreed with those of several of the other studies reported above: students in biological and physical science performed better than the other students on the science-based texts, but humanities majors did no better than the scientists on humanities-based subjects. However, it must be remembered that all TOEFL reading passages are designed to be appropriate to all students, regardless of their field of study, and so a strong subject area effect would not be expected, and any differences might be too subtle to be detected by bias analysis. The fact, therefore, that, in spite of this, two groups of students did perform significantly better in their own subject area is interesting.

Several points emerge from the above studies. Firstly, language proficiency levels seem to play at least as important a role as background knowledge in the comprehension of reading texts. Secondly, background knowledge itself is not easily assessed: a student who is in Business Studies may well have previously worked in another discipline such as Science, or may have scientific interests in his or her spare time. Thirdly, although the above studies were in many ways inconclusive, there did seem to be a tendency for science students to perform better than other students at science-based tests, but to perform as well as the humanities students on humanities based ones. Finally, the level of specificity of the subject-based texts probably varied widely in the different studies, but this was not fully taken into account in the studies.

In the next chapter I review recent research into the effect of background knowledge on reading in the first language. I describe some recent models of reading comprehension, and pay particular attention to schema theory, which has been used as a basis for much recent research into reading and listening comprehension (see Chapters 2 and 3). Schema theoreticians provide accounts of how people might store and activate knowledge, and show how important such knowledge is for comprehension. It might be expected that ESP would now be linked with schema theory, since ESP instructors focus on language and material which are designed to be similar in form and content to those used by their students in their chosen field of work or study. Indeed, Widdowson (1983) uses schema theory in his search for a theoretical model for ESP, and recently ESP researchers such as Jensen and Hansen (1995) have started referring to schema theory in their writings.

2 Reading in a First Language

Reading Research

The process of reading has been intensively studied, and has interested researchers in disciplines as disparate as anthropology, philosophy, psychology, education, artificial intelligence and linguistics. Not surprisingly, therefore, there have been many different approaches to it, and the term 'reading' has been given many interpretations. The following definition of reading provides a useful basis for this study, although, as we shall see later, H. M. Robinson (1966) and parallel processing theoreticians might query the use of the word 'subsequently'.

> *Reading can be defined loosely as the ability to make sense of written or printed symbols. The reader uses the symbols to guide the recovery of information from his or her memory and subsequently uses this information to construct a plausible interpretation of the writer's message .*
> (Mitchell 1982:1)

Although most researchers accept that the process of reading covers all stages from the initial observation of written marks to the comprehension of whole texts (see, for example, H. M. Robinson 1966, Walcutt 1967 and Stanovich 1991), some have concentrated on the decoding of written symbols (for example, Fries 1963), whereas others, such as Adams and Collins (1979), have ignored the decoding stage altogether.

In this study I am concerned with the 'higher' levels of processing, that is, the levels of processing beyond decoding (see Craik and Lockhart 1972, Hulstijn 1991 and Segalowitz, Poulsen and Komeda 1991). I shall concentrate therefore on those theories or models that attempt to account for the comprehension of words, sentences and texts. Some of the research that is relevant to this study relates to listening rather than reading comprehension since listening and reading share many cognitive processes, and since, as Kintsch and van Dijk (1978) pointed out, many of the main differences between reading and listening occur at the decoding level. Of course there *are* differences between listening and reading at the higher levels of processing: listeners and readers usually have different aims when they attend to the discourse, spoken and written texts are generally presented in different styles to suit the medium of presentation, and readers can re-read a text whereas listeners cannot re-listen (see Buck 1990).

However, there are enough similarities between the two receptive skills for research into the higher levels of listening comprehension to apply to reading as well.

The greater part of reading research over the years has been devoted to the decoding of symbols and single words, but some researchers at the start of the 20th century looked at the wider aspects of reading, and indeed put forward ideas about the process of reading which are similar to those being mooted now. In 1908, for example, Huey came to the conclusion that experienced readers read words, phrases and sometimes sentences as complete entities, and that this process inhibited the recognition of the constituent letters or words. In 1917, Thorndike showed how much reading comprehension depended upon the reader's powers of reasoning, and in 1948, Gray emphasised the effect that background knowledge had on reading comprehension. However, as the influence of behaviourism, with its concentration on observable events outside the individual, spread during the 1920s from psychology to the social sciences in general, there was little research into cognitive processes, and research into reading became chiefly restricted to the relationship between stimuli, such as words, and responses, such as word recognition. What research into cognitive processes there was (for example, Bartlett 1932), received little recognition. However, in 1957, two parallel events occurred which had a profound effect on research into reading. The first was the publication of B. F. Skinner's *Verbal Behavior* (1957), which was an attempt to explain language from a behaviourist point of view, and the second was the publication of Chomsky's *Syntactic Structures* (1957), with its notion of deep and surface linguistic structures and the thesis that the study of language and the study of the mind were intimately connected. This book, and Chomsky's (1959) critical review of *Verbal Behavior*, led to the realisation that behaviourism by itself could not account for all human cognitive activity, and that mental processes not only existed but needed to be studied. Then, in 1967, Goodman produced his seminal paper, 'Reading: A Psycholinguistic Guessing Game', which suggested, as indeed Horn (1937) had thirty years earlier, that the reader was not just a passive absorber of information, but was an active participant in the process of reading. This interactive approach to reading, where reading is an active dialogue between the author and the reader, caught the imagination of psycholinguists and cognitive psychologists and led, as Samuels and Kamil (1988) put it, to a 'burst of model-building activity' and 'a geometrically accelerating body of empirical evidence about basic (reading) processes' (1988:22).

Because so many disciplines have been carrying out research into reading, and because these have sometimes depended on each other's work, and sometimes been carried out independently, it is difficult to give a clear chronological account of developments in our understanding of the effect of background knowledge on the reading process. I shall therefore pick out those theories and models of reading which seem to be most closely related to my study, and shall relate them to each other as much as possible.

According to Jenkinson (1972), the earliest model of the complete reading process was drawn up by Gray (1960) and revised by H. M. Robinson (1966). In this model, reading consists of four activities – word perception, comprehension, reaction to what is read, and assimilation of what is read through the fusion of old and new ideas. The last of these emphasises the role of background knowledge in reading, but it is interesting that 'comprehension', which consists of such skills as 'understanding relevant facts', 'following directions' and 'securing the main idea', is considered to be a different aspect of reading. Gray describes comprehension as relating to the literal and the implied meaning of the text, in other words reading the lines and 'reading between the lines' (Gray, 1960:11). He does not relate it directly to 'assimilation' where readers relate what they have read to their background knowledge. We shall see shortly that when cognitive psychologists approach the processes of reading they see assimilation as part and parcel of comprehension. One interesting point about Robinson's (1966) revision of the model is that she emphasises that the four activities operate simultaneously. This is anticipating the concept of parallel processing which will be described later.

Robinson (1966) makes another important point. She says that it is important not to confuse three interrelated aspects of reading: the reading process, the skills and abilities used in reading, and the procedures used to teach reading, and she says that the Gray/Robinson model relates to skills and abilities rather than processes. It is sometimes difficult to keep processes and skills distinct, but Strang (1972) describes abilities and skills as being the *products* of reading, and Alderson (1990a) says that they 'underlie or contribute to the reading process' (1990:425). Typical examples of skills and abilities are: 'recognising details', 'identifying main ideas', and 'distinguishing fact from opinion'. I shall be discussing reading skills in more detail in Chapter 6.

Models by educationists such as Gray and Robinson are very elaborate but include little real detail. They are therefore difficult to test empirically, and there is little evidence to support or reject them.

Models of Reading Comprehension

While most educationists were looking more at reading skills than at the process of reading, psychologists, artificial intelligence researchers and some linguists were seeking to learn more about the cognitive processes involved. In 1977, for example, Carver describes reading as being a linear process from graphic symbols to meaning responses, and says that readers check words individually, and sound them out phonetically. Models such as these are now described as bottom-up models, where the reader passively perceives input which progresses from the lowest level of reading – the interpretation of symbols – to the higher levels such as the assigning of meaning. This flow of information is considered to be very fast, and is little affected by information stored in memory. Gough (1972), too, believes reading to be bottom-up: according to him readers follow five stages: eye fixation, absorption of the visual stimulus, letter identification, phonological representation, and understanding of words serially from left to right. Gough specifically says that the reader is not a guesser: 'He plods through the sentence, letter by letter, word by word' (Gough 1972: 354). Gough expressly presents his views strongly and unambiguously so that they are open to empirical confirmation. Indeed so one-sided are his statements that it is tempting to think that he was writing as devil's advocate. Certainly many experiments have shown that his bottom-up explanations of the reading process are inadequate. In a convincing discussion of Gough's model, Rumelhart (1977a) shows that bottom-up models fail to allow for the fact that the comprehension of letters, words and sentences are all affected by higher level processing. For example, semi-legible hand-written words are interpreted as different words according to the phrases which precede and succeed them (see Nash-Weber 1975), ambiguous words change their meaning according to the syntax of the encompassing proposition – note the different meanings of 'eating' in 'The children are eating apples' and 'The juicy red ones are eating apples' – and the meaning of a sentence is affected by the context in which it appears (see Bransford and Johnson, 1973). In all these cases higher level processes are invoked to elucidate lower level input. The results of these and similar experiments make it clear that the process of reading cannot be a simple linear progression from low level processes upwards.

Goodman's approach to reading is very different. He suggests that reading is driven by hypotheses – the reader follows a cyclical procedure of sampling the text, predicting what will come next, testing predictions, and adjusting or confirming them (Goodman 1975). Since the reader samples the text solely in order to test hypotheses, this model which is 'concept driven' rather than 'data driven' is described by many researchers (for example, Stanovich 1980) as being a top-down model. In top-down models, the decoding process is relatively unimportant. The reader, far from being a passive receiver, plays an active part in the text interpretation, using background knowledge to form inferences, and decoding symbols only when it is necessary for comprehension. Smith (1988), another top-down theorist, describes reading as follows:

> *Features of sequences of words may be analysed but the letters themselves do not need to be identified when the reader's objective is the identification of words. And features of words may be analysed without the words themselves being identified when the purpose of reading is to find specific kinds of sense in the text. Readers can go straight to meaning in the text by means of prediction. Reading is not a matter of identifying word after word.* (Smith 1988:285)

Stanovich (1980) says that top-down theorists such as Goodman and Smith base their ideas on a belief that the actual decoding process is slow, and that it is therefore quicker and more economical for the reader to make hypotheses about what the text will contain next, so that the slow decoding process can to some extent be avoided. He admits that it is now generally accepted that readers do form inferences while they are reading, and that background knowledge has a powerful effect on comprehension, but says that it is false to imply that the generation of hypotheses concerning subsequent words in a text is quicker than processing the words according to purely visual information. He cites experiments which show the importance of the visual clues and says that reading is an interactive process, which involves a combination of top-down and bottom-up processes. Goodman (1981) and Smith (1988) both agree that their models of reading are top-down, but neither of them accepts that they are purely top-down. Goodman (1973) accepts that reading must start with a graphic display, but says that the efficient reader 'touches as few bases as possible to get to his goal' (1973:23). Smith (1988) says that 'no top-downer would want to claim that reading is not an interaction with the text' (1988:218); however, it seems here that 'interactive' is being used in two different ways. As Grabe (1988) points out, the expression 'interactive' is sometimes interpreted to refer to the interaction between the text and the reader, and sometimes to the interaction between top-down and bottom-up reading processes. Smith appears to be using the former meaning. However, since it would be impossible to interpret a text without at least some input from the visual stimuli, this supposed schism between 'top-downers' and 'interactionists' seems to be somewhat contrived. Any difference that there is must be one of degree, with top-down theorists giving precedence to the importance of readers' attitudes, purposes and knowledge, and interactionists perhaps giving more emphasis to bottom-up approaches.

There have been several interactive models of reading comprehension. One of these was developed from a bottom-up model, which had been used for demonstrating how words and symbols can be processed automatically (see LaBerge and Samuels 1974). Samuels (1977) modified it and turned it into an interactive model with feedback loops which allow the reader to move back and forward between the different levels of processing. This model is still based on the idea that the initial stimulus comes from the low level, visual stimuli, but it does allow for context and background knowledge to affect the interpretation of the text.

Another interactive model was devised by Rumelhart (1977a). 'Graphemic input' and 'orthographic', 'lexical', 'syntactical' and 'semantic knowledge' all enter a 'pattern synthesiser' at the same time, and all interact to produce 'the most probable interpretation of the graphemic input' (1977a:588). Rumelhart points out that interactive models are difficult to portray using the traditional information processing tools such as flow charts, and he accepts that in his own model it is not clear what goes on in the pattern synthesiser. However, he shows how computer scientists are developing 'formalisms' for the representation of parallel processes, and has since published many works on parallel distributed processing.

Spiro (1980) emphasises that different people process text in different ways, depending on their purposes, attitudes, interests and background knowledge, and Stanovich (1980) too, in his interactive-compensatory model, takes account of differences among readers. His model is based on Rumelhart's, but allows for the fact that if readers do not recognise a word or phrase because it is unfamiliar, they can compensate for this by using a top-down method of guessing. Similarly, if the topic is unfamiliar, they can apply bottom-up processes. This intuitively satisfying model may account for differences not only among first language readers, but also among proficient and less proficient readers in a foreign language. However, Freebody and Anderson (1983) found that when children were given texts which varied in topic familiarity and vocabulary difficulty they did not compensate for weaknesses in either area. Smith (1988), without giving any reasons, feels that this compensatory model only applies when the reader is aiming at accurate word recognition, and that it would not apply when people are doing 'purposeful, meaningful reading' (1988:285).

In all top-down and interactive models of reading it is presumed that humans depend on memory or previous knowledge of some kind when they interpret written cues. Without previous knowledge they would not be able to take an active part in comprehending a text. For example, they would not be able to make inferences or hypotheses about what was coming next. Fundamental to all these models, therefore, must be some system of storing and retrieving past knowledge. The group of theories which attempt to account for this come under the general umbrella term of 'schema theory'.

Schema Theory

In his 1932 book, *Remembering*, F. C. Bartlett describes a series of experiments related to different aspects of memory. In some studies he asked subjects to introspect as they reproduced diagrams and illustrations which they had been shown only briefly. During experiments, subjects frequently gave the pictures names:

> *The name was given immediately and unreflectingly; for the presented visual pattern seemed at once to fit into or 'match' some preformed scheme or setting.*
> (Bartlett 1932:20)

As the illustrations became more complex, these names affected the ensuing reproductions. If the name closely matched the original illustration, the subject's drawing was similar to the original; if it did not, the drawing was changed to fit in with the name. For example, a rectangle connected by two small lines to a small rectangle within it was drawn correctly by a subject who described it as 'two carpenter's squares', but wrongly by another who called it a 'picture frame'.

Bartlett also demonstrated the role of inferencing in perception. Subjects were shown a series of illustrations containing gradually increasing or decreasing detail, and were asked to describe what they saw. Once they felt they knew what the final illustration would represent they would say, for example, 'It will be a crown', and if this was correct, they would be able to infer the changes in the next illustrations. They built up a set of expectations, and these frequently overrode the actual visual clues. For example, an illegible notice on a closed gate was persistently remembered by most of the subjects as saying 'Trespassers will be Prosecuted'.

Bartlett made no mention of the Gestalt psychologists, but he seems to have been working on very similar lines. One of the basic tenets of their philosophy is that humans seek to make coherent any pattern that they perceive (see Wulf 1922). Gestalt psychologists concentrated mainly on the visual image, but recent cognitive psychologists, for example Bransford (1979) and others, have extended that principle to written and spoken discourse. Bransford's research into the effect of context on comprehension, which is discussed in more detail below, repeatedly shows how people attempt to make sense of material which is initially incomprehensible.

When Bartlett later proposed a theory of remembering to account for the results of his experiments he appropriated the term 'schema'. This had been adopted by Head (1920), who, in a study of the human cortex, wanted to account for our ability to remember previous postural movements when embarking on new ones.

> *(Previous impressions) may rise into consciousness as images, but more often, as in the case of spatial impressions, remain outside central consciousness. Here they form organised models of ourselves which may be called schemata. Such schemata modify the impressions produced by incoming sensory impulses in such a way that the final sensations of position or locality rise into consciousness charged with a relation to something that has gone before.*
> (Head 1920:605)

Bartlett was not convinced by some of Head's theories about the cortex, but he found a redefinition of the term 'schema' helpful in accounting for some aspects of memory:

> *'Schema' refers to an active organisation of past reactions, or of past experiences, which must always be supposed to be operating in any well-adapted organic response. That is, whenever there is any order or regularity of behaviour, a particular response is possible only because it is related to other similar responses which have been serially organised, yet which operate, not simply*

> *as individual members coming one after another, but as a unitary mass. ... All incoming impulses of a certain kind, or mode, go together to build up an active, organised setting: visual, auditory, various types of cutaneous impulses and the like, at a relatively low level; all the experiences connected by a common interest: in sport, in literature, history, art, science, philosophy and so on, at a higher level.*
> (Bartlett 1932:201)

A most convincing advocate of the importance of the reader's own input is the psychologist John Bransford. His work owes much to Bartlett's findings, and although he is not himself a schema theorist, he makes many of the same assumptions. I shall briefly describe the work of Bransford and his colleagues, before describing some of the schema theory models.

Like Bartlett (1932) before him, Bransford uses memory in the form of free recall to test subjects' understanding, since he says 'Poor understanding result(s) in poor memory performance, (and) it might be tempting to assume that the better the comprehension, the better memory performance will be' (Bransford 1979:154). He points out, however, that making wrong assumptions while comprehending leads to apparent memory errors. He and various colleagues describe a series of experiments in which they asked subjects to describe or remember pictures or texts. Bransford and Johnson (1973), for example, gave one group of listeners an untitled text which is now much quoted in the literature. It starts as follows:

> *The procedure is actually quite simple. First you arrange things into different groups. Of course one pile may be sufficient depending on how much there is to do. If you have to go somewhere else due to lack of facilities that is the next step; otherwise, you are pretty well set. It is important not to overdo things.*
> (1973:722)

Most of the group said that the passage was incomprehensible, and were unable to remember it accurately. However, another group, who were given the title, 'Washing Clothes', before they heard the text, found the passage much easier to understand, and remembered it better. The effect of adding this simple context to such a text is dramatic, and the authors deduce two things from this. Firstly, readers need background knowledge to make sense of a passage, but secondly, and most importantly, they must be able to activate this background knowledge. It is no use knowing about the processes involved in using a washing machine, if you do not know that you should bring this knowledge to bear. Interestingly, Bransford and McCarrell (1977) also show how the wrong background knowledge can be activated so that readers are led into confusion. Subjects who were asked to read a passage about a space trip remembered it much better under its correct title than when the passage was wrongly headed 'Watching a peace march from the 40th floor'. In the same way that Bartlett's subjects changed what they saw in order to make sense of it, Bransford and McCarrell's subjects struggled to make sense of the wrongly titled passage, and by altering the meaning, managed to convince themselves that the text was at least partly comprehensible.

Alba *et al.* (1981) replicated some of Bransford and Johnson's experiments, asking students to read rather than to listen to the texts. They found that if students were asked to recall the passages, the results were the same as in the original experiments, but if they were asked to recognise individual sentences from the passages, they performed equally well whether they knew the topic of the passage or not. The researchers presumed that what affected the subjects' recognition of sentences was the cohesiveness of the text rather than the provision of a text title, and this led them to deduce that the presence of a context only affected recall, not understanding. It was because they suspected that recall might not be a suitable tool for testing comprehension that they used sentence recognition as well. While it is true that asking students to recall passages may well be an inappropriate method of testing comprehension (see the discussion at the end of this section), it is not clear that recognition of individual sentences is any better, since that may depend on rote memory rather than high level understanding. Certainly the recognition of the wording of single sentences does not seem to have been testing understanding in Alba, Alexander, Hasher and Camiglia's experiment, since, as any reader of Bransford and Johnson's texts will agree, their passages, which are difficult or impossible to understand without a context, are suddenly comprehensible once the context is provided.

Bransford and his colleagues intentionally designed obscure texts that would have the maximum potential for demonstrating the effect of context. Tannen (1979) points out that the texts are unnaturally full of pronouns, and Haviland and Clark (1974) show that the Washing Clothes passage does not follow the usual Given/New distinctions which are required in natural discourse. Haviland and Clark's Given-New Strategy model rests on the assumption that language is mainly used for imparting new information, and that the speaker anticipates what the listener already knows, and uses syntactic clues to highlight the new information. For example, in the sentence 'It was Einstein who searched in vain for the unified field', the Given material is 'Someone searched in vain for the unified field', and the New is 'That someone is Einstein'. The listener's success in understanding depends on whether the given information matches something in his or her memory. In the Washing Clothes passage there is no opportunity for the listener to match the Given with an antecedent in memory, since the text starts directly with 'The procedure ... '. What procedure? The essential Given material is missing. In Bransford and Johnson's terms, the essential previous knowledge has not been activated. This Given/New concept accounts for many misunderstandings in hearing and reading texts.

In attempting to make their texts obscure, Bransford and Johnson perhaps make their scripts unnecessarily ambiguous, and thus limit the extent to which one can extrapolate from the results. Parts of the Washing Clothes text, for example, are difficult to understand even when a context has been provided. For example, 'It is important not to overdo things' presumably means, 'It is important not to overload the machine'. Here the problem does not just lie with a misuse of cohesive devices: the sentence is in an inappropriate register for a set of instructions, and the phrase 'overdo things' is not only vague but misleading. To interpret this the reader surely requires not only background knowledge, but also imagination. It would be interesting to know how well listeners, even if they knew the context of the passage, remembered that particular sentence. Of course an authentic text might not demonstrate Bransford and Johnson's point so well, but the fact that the Washing Clothes text is so deviant means that some critics are reluctant to relate these findings directly to the comprehension of 'natural' texts. For example, Grabe (1988) points out that one of the important parts of reading is the recognition of text genres and distinct text types which are deliberately exploited by writers. The specially devised experimental texts, Grabe thinks, 'may not tap into particular genre or linguistic information that is available to readers when processing longer text segments' (1988:65). However, Anderson *et al.* (1977a) say that just because a text is bizarre, there is no reason to think that its readers make more use of extralinguistic knowledge to interpret it than they would usually, and Bransford, Stein and Shelton (1984), while accepting some criticisms, use examples to show that we also depend upon our activation of the relevant general knowledge in order to understand 'natural' language.

It is now generally accepted that background knowledge does affect reading comprehension, but there is as yet no consensus as to how it does so. It is not known how knowledge is stored in the mind, nor how it is accessed. One theory of how information is stored is proposed by the artificial intelligence researchers, Collins and Quillian (1969). They describe a semantic-memory structure which consists of a set of hierarchical systems. Under the superordinate term, 'animal', for example, are fish and bird, and under bird are ostrich and canary. Each term in the hierarchy is stored with its distinctive features, so that an animal has skin and can breathe, a bird has wings and can fly, an ostrich has long legs and cannot fly and a canary is yellow and can sing. For 'cognitive economy' each feature is stored at the most general possible level, so that, for example, characteristics that are common to all birds are stored at the 'bird' level. What is interesting about this system is that Collins and Quillian have shown that people answer questions such as 'Does a canary sing?', where the question relates to the distinctive features of the canary, more quickly than questions such as 'Can a canary fly?' where the relationship is one step away in the hierarchy. Similarly, 'Can a canary fly?' is answered more quickly than 'Does a canary breathe?', where the connection is yet further up the hierarchy. It seems that in order to answer this last question, people have to go through the intervening steps of the hierarchy: the canary is a bird, a bird is an animal, animals breathe, therefore canaries breathe. This model has some critics: Bransford (1979) wonders whether the speed with which people answered the questions was more related to frequency of association than to positions in a hierarchy, and Anderson and Ortony (1975) show that the hierarchies are too rigid to cope with problems of polysemy. How, for example, would such a hierarchy account for the different meanings of 'piano' in 'Pianos can be pleasing to listen to', and 'Pianos can be difficult to move'? However, this model has been very influential in artificial intelligence and cognitive science circles (see, for example, Rumelhart and Ortony 1977, and Pearson and Fielding 1991).

The 1970s was a vintage decade for the introduction of new models attempting to account for the formal structures that underlie knowledge, and although the models differ in many ways, they have enough features in common to be called schema-theoretic models.

> *A schema theory is basically a theory about knowledge. It is a theory about how knowledge is represented and about how that representation facilitates the* **use** *of knowledge in particular ways. According to schema theories, all knowledge is packaged into units. These units are schemata. Embedded in these packets of knowledge is, in addition to the knowledge itself, information about how this knowledge is to be used.*
> (Rumelhart 1980:33)

Two of the most celebrated schema-theoretic models are those by Minsky (1977) and Schank and Abelson (1977). I shall briefly describe these two models here, in order to provide some concrete examples of schema models, before going on to a more general description of schema theory.

Minsky (1975) introduced what he called 'frame system theory'. According to this, each person has, stored in memory, a host of interrelated frames which represent stereotypical situations; when people meet a new situation, such as entering a room, or going to a party, they choose the appropriate frame and change it as necessary to fit the new situation. Minsky explains his frame as follows:

> *We can think of a frame as a network of nodes and relations. The 'top levels' of a frame are fixed, and represent things that are always true about the supposed situation. The lower levels have many* ***terminals*** *– 'slots' that must be filled by specific instances or data. Each terminal can specify conditions its assignments must meet. (The assignments are usually smaller 'subframes'.) Simple conditions are specified by* ***markers*** *that might require a terminal assignment to be a person, an object of sufficient value, or a pointer to a subframe of a certain type. More complex conditions can specify relations among the things assigned to*

> *several terminals.*
> *Collections of related frames are linked together into* ***frame-systems**... . **Different frames of a system share the same terminals** ... (which) **are normally already filled with 'default' assignments**. These default assignments can easily be displaced by new items that fit better the current situation ... .*
> *Once a frame is proposed to represent a situation, a* ***matching*** *process tries to assign values to each frame's terminals consistent with the markers at each place.*
> (Minsky 1975:355; Minsky's emphases)

As an example, Minsky imagines someone walking into a room and describes the frame structure that will be summoned up. This will include default value expectations concerning the shape of the room, doors, windows, furniture and so on, and Minsky shows how these will be adjusted as the actual room is perceived.

Minsky also shows how frames can provide the background knowledge that is required for understanding stories, and how if, because of lack of context, the right frame is not called up, a sentence can be misunderstood. For example, 'She wondered if he would like a kite', is not easy to understand, unless we know that 'Jane was invited to Jack's birthday party'. (This relates directly to Haviland and Clark's Given/New model, 1974.) Minsky suggests that there may be different kinds of frames: syntactic frames, which are mainly concerned with verb and noun structures; semantic frames, which centre on the meaning of words; narrative frames, which provide the basis for stories; thematic frames for topics; and scenarios for events.

Minsky's ideas were an inspiration to artificial intelligence researchers who were trying to program computers to simulate human methods of understanding. Schank and Abelson (1977) found, however, that the scope of the frames model was too great to be easily applied to computer programming, and they limited their approach to a more specialised area which could be tested on the computer. They chose an area similar to Minsky's scenario, and called it a 'script'. A script is:

> *a structure that describes an appropriate sequence of events in a particular context. A script is made up of slots and requirements about what can fill those slots. The structure is an interconnected whole, and what is in one slot affects what can be in another. Scripts handle stylised everyday situations ... . A script is a 'predetermined, stereotyped sequence of actions that define a well known situation'. A script is in effect a very boring little story.*
> (Schank and Abelson 1977:422)

As an example, the authors cite a four-sentence description of a visit to a restaurant:

John went to a restaurant.
He ordered a hamburger.
It was cold when the waitress brought it.
He left her a very small tip.

These four sentences adequately cover the story because we fill in the missing details ourselves. If we did not have background knowledge of the standard scenario of a visit to a restaurant we should not understand the story. Only if we already have stored, for example, information about tipping, and that tips are given for prompt and satisfactory service, do we understand how 'He left her a very small tip' follows on from 'It was cold when she brought it'.

Schank and Abelson devised a computer program called SAM, which was supplied with a restaurant script. When it was presented with the following story,

John went to a restaurant. The hostess seated John. The hostess gave John a menu. The waiter came to the table. John ordered lobster. John was served quickly. John left a large tip. John left the restaurant.

it was able to use its programmed script to fill in the stages in the story that the text omitted. For example, it paraphrased part of the story as follows:

The waiter got the lobster from the chef. The waiter served John the lobster. John ate the lobster.

From this it was able to make deductions, and to answer questions which were not based on the actual text content. For example it could answer the question, 'What did John eat?'

Schank and Abelson increased the dimensions of their work by adding 'plans', which are responsible for people's deliberate behaviour, such as the withholding of a tip, and 'goals' which are people's intentions. Even with these additions the model is far too simple to be able to interpret texts of any complexity, and it must be remembered that this program was designed for machine rather than human processing. However, it does highlight the importance of inferencing, and shows how vital prior knowledge is for even the simplest comprehension task.

I shall not describe the other well known models here (for example, Winograd 1975, and Rumelhart and Ortony 1977) but shall give a brief overview of the main features that schema-theoretic models seem to have in common. My description is based mainly on the accounts in Rumelhart and Ortony (1977), Schallert (1982) and Anderson and Pearson (1988).

According to schema theory, knowledge is stored not in lists, but in hierarchies. Within these hierarchies are schemata which are embedded in other schemata, and which themselves contain subschemata. These schemata vary in their levels of abstraction, and represent all sorts of knowledge, such as objects, academic topics, rules, events, routines and social situations. They represent knowledge, rather than definitions, so they are not language based, but are symbolic representations of knowledge which may be used for understanding language. Schemata are not static, but fluid; they change according to the input. Schemata can be refined and new ones can be developed by the process of accommodation, that is, the modification of previous schemata in the light of new information (see Bransford 1979).

Each schema has a specific set of variables, some of which are obligatory. For example, Anderson and Pearson (1988) give as an example the schema 'SHIP'S CHRISTENING'. The variables of this schema would include a celebrity, a new ship and a bottle. On different occasions the variables would take different values, which would be under certain constraints, so that the celebrity, for example, would have a certain standing in the community, and might therefore be the Prince of Wales or the Secretary of State for Defence. If no value is provided from the incoming text, the variable is assigned its default value. For example, if a sentence reads 'The mayor broke the bottle on the *Endeavour's* bows', readers would presume that the bottle contained champagne. As long as they have prior knowledge about ship's christenings, that is, the SHIP'S CHRISTENING schema, they can supply that information for themselves. The assigning of values to the variables is called 'instantiation'. Some values are more typical than others, and these are the ones that are most likely to be instantiated (see Anderson *et al.* 1976). For example, in the above sentence, 'the new ship' is more likely to be instantiated as a naval ship than as a liner. In this case there has been interaction between the SHIP'S CHRISTENING schema and another schema, possibly SHIP NAMES, which has constrained the interpretation and acquisition of the input information. The instantiation of variables is therefore affected by context, and comprehension proceeds as the variables of the schemata are instantiated.

So far in this chapter, the term 'context' has been used very loosely, since different authors use the term to mean different things. I interpret it to refer to any of the following:

a) the text which precedes and succeeds the material under scrutiny,
b) the setting of a text – topic, period etc.,
c) the purposes, attitude and interests of the reader and the writer.

There has been some disagreement about which levels of language processing are covered by schema theory. Although it is generally considered to be related to high level semantic processing, some researchers have widened its scope. Van Dijk and Kintsch (1983), for example, say that schemata can be found at local or global levels, for example verb and narrative schemata, and Adams and Collins (1979) expect schema theory to account for all stages of reading comprehension, including letter identification and sentence parsing.

In a review of Adams and Collins' 1979 article, Schank (1980) complains that by widening the notion of schema theory in this way, the authors have in effect reduced it to nothing. He points out that their model for sentence parsing, which they describe as being part of schema theory, is just an 'apparatus' for processing things, and is not at all the same as a high level mental structure. Schank sees schema theory as accounting for those high level mental processes such as inferencing and predicting of which we are aware. For example, if we are presented with a restaurant story we can, at least partially, explain what deductions we have made on the way to comprehension. Lower level processes, on the other hand, such as letter recognition or sentence parsing, are generally unconscious (see LaBerge and Samuels 1974 on automaticity) and should not come under the heading of schema theory.

Since schema theory is still so conjectural, and since different models vary widely, the choice of the levels of processing no doubt depends on the opinions of those devising the models, and so will vary according to the theoretician.

One area of schema theory which overlaps with background knowledge is that of formal schemata, that is, the stored knowledge of text structure. Structures of stories, for example, have been much researched. Thorndyke (1977) devised a simple story grammar based on one by Rumelhart (1975). According to Thorndyke, stories consist of several unique components (setting, theme, plot and resolution) which are conceptually separable, and are organised by narrative syntax rules which are independent of the linguistic content of the story. Thorndyke says that if people are able to match a particular story to a previously learned organisational framework, they will be able to use that framework to help them understand the text. When he tested this theory, he found that subjects were better able to remember a story if it had a familiar structure, than if it had familiar content. However, Garnham (1985) and Garnham and Mason (1987) are not convinced by his results. They argue that Thorndyke's findings were not necessarily caused by familiarity with the text structure, but could have been caused solely by knowledge about the world. Indeed they say that story grammars are redundant because the organisation of stories into episodes reflects the way people store events that occur both in stories and in the real world.

This same criticism could be levelled at Kintsch and Greene (1978), who asked Americans to write summaries of an Apache folk tale and a Grimm fairy tale. When they found that the Americans had more difficulty in writing the Apache summary than the Grimm one, they presumed that it was because the students were unfamiliar with the structure and style of Apache folk tales. However, once again, this discrepancy could have been caused by differences in the subjects' cultural or content knowledge (see, for example, Steffensen and Joag-Dev 1984).

On the whole, first language research into formal schemata has concentrated on narratives, and there has been less work on expository schemata. Most research into expository texts has concentrated on the analysis of text structure and the reader's processing of that structure. The main researchers in this area have been Kintsch, van Dijk and Meyer, and although the effect of background knowledge on comprehension is not central to their work, it plays an important part. For example, in Kintsch and van Dijk's model of comprehension (1978), propositions are connected to form macropropositions, and macropropositions are connected to form the text macrostructure which reflects the gist or the topic of a text. A reader of this text will make inferences about this topic, using activators such as the title, thematic words and world knowledge. Kintsch and van Dijk say that many discourse types have a conventional, schematic structure that organises the macropropositions in the text, and they refer to narrative schemata such as those mentioned above, and schemata for newspaper articles and research reports. Although the authors do not discuss the role of world knowledge schemata in any detail, they describe the strategies that would be needed to activate them.

Meyer (1984) has produced a model of reading comprehension which describes the interaction of text variables (such as topic, content, structure and emphasis) with writer variables (such as knowledge of audience and topic), and reader variables (such as world knowledge, perspective, values and purpose). She says that the dominant reading strategy followed by skilled readers is based on text structure. Readers approach a text looking for patterns that will tie all the propositions together. For example, a text's first sentence may match the readers' problem/solution schemata. Meyer (1975, 1985) lists five kinds of expository discourse which affect the reader's expectations while reading – collection (this shows how ideas or events are related), description, causation, problem/solution and comparison – and says that these are all top-level structures which can be thought of as types of schemata. Meyer shows how an easy-to-remember expository text will be well structured so that the reader is led to selecting appropriate schemata which will generate expectations about what will come next, and will instantiate most of the new propositions and transfer the results into long-term memory. Meyer and Freedle (1984) gave students texts which were based on the same content, but which were presented in different discourse styles, and found that students remembered text which was organised in a 'causation' or 'comparative' style better than when it was written in a simple 'descriptive/collective' style. The researchers deduced that more organised types of discourse led to more efficient processing, and that this affected recall. (See also Urquhart 1984, who found that readers recalled events which were described in order of occurrence better than ones which were not.)

It should be noted that Meyer and Freedle (1984) used recall protocols to enquire into memory *not* comprehension; so far in this review of schema theory I have barely distinguished between the two, since knowledge, memory and comprehension are so closely linked that it is often difficult to do so.

Comprehension is required for the storage of new knowledge, memory is required for the recall of that knowledge, and memory is also required for the comprehension of new material. However, it is important not to confuse comprehension and memory. As we have already seen, Bransford and Johnson (1973) used text recall to assess students' reading comprehension and memory, but they maintained the distinction between the two. Bransford (1979) points out that 'comprehension does not guarantee optimal free recall' (1979:34). Although it is presumably impossible to recall something accurately (unless it has been rote learned) without having initially understood it, it is not the case that everything that is understood is remembered.

Recall methods depend on memory, and are not, therefore, ideal for investigating people's understanding. However, Swaffar, Arens and Byrnes (1991), writing about English as a second language, argue that most reading comprehension items only test bottom-up skills and 'fail to measure the *way* a reader understands' (1991:153). They therefore argue that reading tests must 'reveal that students are able to comprehend textual meaning *and* utilise or express views about that meaning' (1991:158), and they, like Bernhardt (1991), advocate the use of recall protocols in tests of reading comprehension, although they have certain reservations about the validity of the generally used marking schemes. The pros and cons of using recall protocols should be held in mind when considering the studies described in Chapter 3, since much research into ESL reading processes is based on the use of such techniques.

Limitations of Schema Theory

Schema theory has opened the doors to a mass of further research into human learning and memory, and is generally seen to have an important role in human comprehension. As Wilson and Anderson (1986) say, schema theory serves the following functions: it provides ideational scaffolding; it directs allocation of attention; it enables inferential elaboration; it allows orderly searches of memory; it facilitates editing and summarising; and it permits inferential reconstruction (1986:53). However, not surprisingly, it does not satisfy everyone. Alba and Hasher (1983), for example, think that schema theory is too limited, and that the stored record of any event is far more detailed than schema theory implies. Beers (1987) says that the computer programs based on schema theory are too restricted in scope, and lack the flexibility of the human mind. Conversely, he says that as schema theory is so flexible and can conform to any number of different situations, it is not easy to verify empirically: 'It is too vague to have interesting testable consequences' (Beers 1987: 371). (See also Mandler 1984.) Van Dijk and Kintsch (1983) think that schema theory cannot account for our understanding of a text, since, as well as general knowledge, most discourses bring in new information which is not just an instantiation of a stereotypical script or frame. McNamara, Miller and Bransford (1991) agree, saying that although there is no doubt that people do use schemata in comprehension, it is not clear what they do with them. McNamara *et al.* say that a more satisfactory approach, which can cover scripted and unscripted activities, comes under the theory of mental models (see Johnson-Laird 1983 and Garnham 1985). Readers construct a mental model, which consists of mental tokens:

> *arranged in a structure that depicts the situation described by a text. ... The mental model can give rise to images, although mental models can also contain nonperceptual information, such as goals and causal relationships.*
> (McNamara, Miller and Bransford 1991:494)

It is quite possible, says Johnson-Laird (1983), that schemata are a special case of procedures for constructing mental models.

McClelland, Rumelhart and Hinton (1986) say that one of the greatest difficulties with schema theory has been finding ways of coping with the complex interaction of schemata that have to take place during any simple attempt at understanding. 'A vast number of microsteps' (1986:9) are needed for any single act of cognition, and sequential models of human cognition are too slow, especially when each act of cognition is circumscribed by constraints. They found that schema theory even failed to account adequately for the task of understanding simple stories, and they have come to think that the answer lies in parallel distributed processing (PDP).

It seems most likely that some sort of parallel interactive processing does take place during the comprehension of new material, but how it does so is not yet known. In the meantime it seems that schema theory plays a part in both mental and PDP models.

Research into the Effect of Content and Cultural Background Knowledge on Reading

Researchers vary in the amount of effect which they think readers' background knowledge has on comprehension. Clarke and Silberstein (1977) say that readers bring more to the text than writers because of their 'formidable amount of information and ideas, attitudes and beliefs' (1977:49), but Eskey (1988) considers that this effect is exaggerated. He accepts that successful reading depends on relating received information to prior knowledge, but says that such reading cannot be carried out without successful decoding, and that at least some understanding of the text can take place without the activation of related schemata.

Whatever the truth may be, many researchers have studied the effect of prior knowledge on comprehension, some comparing the performance of students from different cultural or educational backgrounds, and others comparing students with different amounts of knowledge of a topic. All the following experiments are based on schema theory.

In a study into the effect of background culture on interpretation, Tannen (1979) showed young American and Greek women a six minute film about a boy stealing pears and then asked them to re-tell the story. In their recalls, the subjects showed that there were many ways in which they organised and changed the content of the film to fit their expectations, and Tannen found that many of these seemed to be culturally determined. For example, many of the Americans showed awareness of the requirements of media productions, and commented on the film's technique, strange colour and unusual sound effects. None of the Greeks commented on the technical side of the film, or criticised it in any way; they concentrated on interpreting the meaning of the story.

Steffensen, Joag-Dev and Anderson (1979) asked Indians and Americans to read and then recall two letters, one describing an Indian wedding, and the other an American one. The subjects read the letter relating to their own culture faster, and produced appropriate elaborations to the story, that is they added facts which were not in the original letter, but which made sense. When they recalled the other passage they misinterpreted it, adding culturally based distortions. The authors concluded that cultural schemata showed a pervasive influence on comprehension and memory.

Anderson *et al.* (1977a) presented students in weight-lifting or music classes with two texts, each of which could be read in two ways. The first could be understood as a wrestling match or an escape from prison, and the second as a game of cards or a woodwind rehearsal. The weight-lifting students took it for granted that the first text concerned a wrestling match and that the second one was about a game of cards. The music students thought the first text described someone escaping from prison and the second one was about a music rehearsal. Although either reading of the texts led to some anomalies, most of the students managed to fit these into their understanding of the texts, and a startling 80% never realised that there might be different interpretations.

In a rather different approach, Spilich *et al.* (1979) chose students who had a high (HK) or a low (LK) level of knowledge about a topic, baseball, and gave them a very technical account of a baseball match. The HK students recalled the passage more coherently than the LK ones, and scored higher on a set of questions about the passage. They also produced more elaborations of the input and gave graphic accounts of the game. The LK subjects gave very short accounts which were often out of order, and included many irrelevancies. From the results the authors assumed that the subjects matched input with their knowledge structures, and that because the HK subjects had more knowledge structures related to baseball they were able to process the information more readily.

Vesonder (1979) had similar results in a study based on Kintsch's (1974) theory of text processing. He presented science and non-science students with two scientific texts, and found that the science students recalled the texts better than the non-scientists, and were more accurate at recognising statements which had appeared in the texts. Vesonder found that his Kintsch-based grammar, which did not take background knowledge into account, was useful as a structural processing mechanism, but was unable to explain the differences between the two groups of students.

Freebody and Anderson (1983) gave students explanations of a familiar and an unfamiliar game. The texts were carefully matched to contain similar topics, and the grammar and the vocabulary were almost identical. There were easy and difficult versions of the texts. In the difficult versions one third of the content words were exchanged with less common ones. The students were asked to summarise the texts, recall them freely and answer sentence recognition questions. Topic familiarity accounted for three times as much of the score variance as vocabulary difficulty, and so the researchers assumed that prior knowledge was more important than text difficulty.

Symons and Pressley (1993) asked groups of students at different stages in a psychology course to search through a textbook for material that would be relevant to a given topic. The students were given ten low-inference questions which could be answered by reference to explicit statements in the text. Students did progressively better at the task as their course progressed and they became more familiar with the subject matter. They did not improve in the same way when they were given similar tasks related to earth sciences. The authors presumed firstly that, as Pichert and Anderson (1977) say, the incoming information fills slots into an activated schema and is processed more easily than is information that does not fit the schema, and secondly, that attention is directed towards information which is considered relevant and important, and that information which is consistent with prior knowledge has a storage advantage over unfamiliar information during encoding and is thus more likely to be recalled.

One educational researcher who, without benefit of schema theory, had a strong influence on research into the effect of prior knowledge on reading comprehension, was David Ausubel, who described how learners drew on previously acquired concepts when they were trying to understand new information (Ausubel, 1963). He introduced the expression the 'advance organiser' which is now widely used in education, and he argued that by providing students with advance organisers teachers could help them to assimilate new information better. In order to test whether an advance organiser helped comprehension, Ausubel (1960) chose students from eight different academic disciplines and gave them a multiple-choice test on the contents of a passage on metallurgy. Five days before this test, half the students had been given an introductory text on the composition of metals, and the other half a text on the history of the processing of iron and steel. Ausubel's expectation was that the text on the composition of metals would work as an advance organiser and that students who had read it would score higher in the metallurgy test than those who had only read the historical passage. From Ausubel's point of view the results of the study were disappointing as the students who read the introductory script did little better than the other group. However, the results are interesting because they appear to have been partly affected by the students' major field of study, that is, their prior subject knowledge. In a later experiment, in which the text was on a subject unknown to any of the students – endocrinology – there was again no significant difference between the experimental and control groups overall, but there was a significant difference ($p = .01$) among those students who had the lowest marks in a verbal ability score. Ausubel concluded that students at a higher level of language ability could spontaneously organise new material, whereas those at a lower level could not.

The Effect of Background Knowledge on 'Good' and 'Bad' Readers

One point which has not yet been mentioned is the fact that readers, even in their first language, have different levels of reading ability, so that some can loosely be described as 'good' and some as 'bad' readers. Perfetti and Lesgold (1977) say that three of the sources of these individual differences are variation in people's speed of verbal coding, differences in short term memory capacity, and varying sensitivity to discourse structures. Interestingly, different theorists have produced diametrically opposed views on the effect of level of reading ability on students' dependence on prior knowledge. For example, Perfetti and Lesgold (1977) argue that whereas good readers can take advantage of background knowledge and context, readers who are slow at decoding symbols and words overburden their short-term memory and cannot call up the appropriate schemata. However, Stanovich (1991) says that the quicker word recognition skills of the better reader are not due to superior context skills, since in many reaction-time studies it has been shown that poor readers often use context more than good ones. Stanovich thinks that as word recognition efficiency improves, the effects of background knowledge and context dependency decrease. I shall be discussing this further when I look at the role of schema theory in reading in a second language.

Conclusions about Reading in a First Language

Although there are some disagreements about how people process reading in their first language, it is possible to draw the following conclusions from the above discussion.

The way a reader understands a text depends to a great extent on that reader's purpose in reading the text, and also on that reader's knowledge and beliefs about the world. Therefore, the process of reading is interactive, in that comprehension of a text depends not only on the writer's input but also on the reader's; the writer makes certain assumptions about the prior knowledge of the reader and this is apparent in the text. The process of reading is a combination of bottom-up and top-down procedures which interact with each other. At least some of the lower level processes, such as the decoding of letters and words, are mostly automatic for skilled readers. The reader brings a set of schemata to bear on the reading process. These relate to the lexical system, the syntactic system and the semantic system, and are affected by attitude and culture. They enable the reader to make the many inferences required by the writer. Generally schemata can be adjusted to accommodate new information, but if the reader's schemata are inadequate because of a lack of the appropriate background knowledge, then comprehension breaks down. Finally, it must not be presumed that all readers read in the same way. There are differences in the ways they process text, and some of these differences may account for the fact that there are 'good' and 'bad' readers. However, little is yet known about how, why and when these differences occur.

The question now is whether readers in a second or foreign language read in the same way.

3 Reading in a Second Language

In this chapter I shall first discuss prevailing views on whether readers in a second or foreign language (L2) read in the same way as native speakers, and then discuss the importance of L2 readers' background knowledge in their processing of written texts. Before we go any further, however, it will be as well to decide whether the present study is concerned with second or foreign language learners since there is a traditional distinction between the two. The former live in countries which use the language for at least some of their day-to-day activities, and/or use it as the medium of instruction in the secondary schools. The latter learn it in a country where it is not commonly spoken. In the case of a language learnt for academic purposes, students may have learnt it as a foreign language in their home countries, but are preparing to move, or have already moved to a country where, at least at tertiary level, that language is the medium of instruction. In this case the students are presumably transferring from using the language as a foreign language to using it as a second language. The research studies that I shall report in this section sometimes refer to second languages, sometimes to foreign languages and sometimes to both. However, as my research relates to language for academic purposes, I shall refer to all non-native speakers as speakers of a second language (L2).

Comparison of L1 and L2 Reading Processes

Many researchers have asked whether L2 readers read in the same way as native speakers: Alderson (1984), Devine (1988a), Barnett (1989), Bernhardt (1991), Carrell (1991) and Bossers (1991, 1992) discuss the question in some detail. There seem to be two different beliefs. One is that the processes are indeed the same and that anyone who is a 'good' reader in their first language (L1), will also be a good reader in the L2. Block (1986), for example, found no differences in the strategies used by poor L1 and L2 readers, and Sarig (1987) discovered that any differences there were owed more to individual methods of processing reading than to differences between first and second language speakers. However, other researchers have found, not surprisingly, that inadequacies in the second language have prevented L2 readers from using the same skills and processes that they use in their first language.

Coady (1979) says that a great number of reading skills transfer automatically from the first to the second language, and that the advanced level student reads in the same way as a native speaker. However, he says that the lower level reader will be prevented from using such skills as inference and prediction by an inability to decode the language. He suggests that as learners become more proficient they gradually put less and less emphasis on decoding processes, and more on cognitive strategies. Eskey (1988) considers that accurate decoding is essential for all readers, and that this has to become automatic before ESL readers can read in the same way as native speakers. Carrell (1988) says that most ESL learners put too much reliance on bottom-up processing, and Barnett (1989) thinks that bottom-up models such as Carver's (1977) (see Chapter 2) are fair representations of the way in which ESL beginners read. Wolff (1987), on the other hand, has found that the more difficult the text, and the less familiar L2 students are with the lexis and syntax, the more they depend on top-down processes.

It is now generally accepted that low level language learners do not read in the same way as native speakers, although it is not known in what way they differ. Nor is it known at what stage, if any, the second language reader becomes sufficiently proficient in the L2 to be able to read it in the same way as an L1 reader. If advanced proficiency learners do read L1 and L2 texts in the same way, at what stage does this change from second-language-like to first-language-like reading take place? Is there, as Clarke (1980) and Eskey (1988) suggest, some sort of linguistic threshold that students have to reach before they can bring their first language strategies to bear?

Clarke (1980) carried out two studies with low level ESL students who spoke Spanish as their first language. He gave them cloze tests in English and Spanish, and found that good L1 readers, defined as those who did well at the Spanish cloze tests, seemed to depend on semantic clues, in contrast to the poor readers who depended on syntactic clues. However, in the second language both groups used the same proportion of syntactic clues. The author presumed that poor language ability prevented the good readers from using their usual reading skills when reading in the foreign language. Their limited control of the language forced them to revert to poor reader strategies, that is, it 'short-circuited' the reading process. Clarke confirmed his findings with a miscue analysis where he asked two readers to read a passage aloud, and then analysed their errors. The good reader produced fewer miscues in Spanish and English than did the poor reader and those few miscues were either semantically acceptable, or later corrected. However, the differences between the two were smaller in English than in Spanish, and the good reader seemed 'less able to focus on semantic cues in the target language than in the native language' (1980:200). Clarke concluded that a low level learner of a second language cannot decode enough of the graphic and lexical symbols to be able to bring top-down processing systems to bear.

Cziko (1980) used miscue analysis to assess the reading strategies of native French speakers at intermediate and advanced ESL levels. When the students read aloud French and English texts Cziko found that while the native speakers and the advanced learners made semantically appropriate deletion and insertion errors, the intermediate learners tended to make substitutions which resembled the original words in appearance but did not make syntactic or semantic sense. Although Cziko had reservations about the use of miscue analysis for this research he felt that the results showed that reading strategies varied according to readers' competence in the language, and that they showed that native speakers and advanced learners used interactive reading processes, while the intermediate learners inclined towards bottom-up skills.

McLeod and McLaughlin (1986) argued that in order to master complex cognitive tasks, readers must be able to process the less complex ones automatically (for an explanation of 'automatic' see McLaughlin, Rossman and McLeod, 1983). If tasks such as decoding letters and words were not automatic, the processing demands themselves took up so much time that there was none left for more complex activities. In their study, the authors used native and non-native speakers, and supplemented miscue analysis with what they called an 'oral cloze' test. The ESL speakers were classed as beginners and advanced level students according to a language placement test. All subjects were given two passages to read aloud, and were marked according to 'meaningful' errors, that is errors which did not change the meaning of the text, and 'non-meaningful' errors, which did. They were then given a ten-item oral cloze test in which they were asked to read aloud two to three lines of a text before predicting what the next word might be. The miscue analysis showed that the beginners made mostly non-meaningful errors, and were able to make almost no correct predictions in the cloze test. The advanced learners were better at making predictions in the cloze test, but still made mostly non-meaningful errors in the miscue task. They did not seem to interact with the text in the way that the native speakers did. The native speakers made many meaningful insertions, deletions and substitutions. McLeod and McLaughlin concluded that the advanced learners had still not reached the stage where they automatically restructured text, although they had now mastered many of the mechanical aspects of reading. They accepted that the advanced readers' concentration on pronunciation might have interfered with their performance, but said that this did not account for the differences between the beginners and the advanced learners.

Bossers (1992) wanted to find out whether L1 reading ability or L2 knowledge had more effect on students' reading comprehension. He gave Turkish students, who had at least an intermediate level of Dutch proficiency, a Dutch multiple choice reading comprehension test, and used multiple regression analysis to see whether L1 reading ability or L2 knowledge accounted for more of the test score variance. He found that both variables contributed significantly, but that L2 knowledge was a far more important factor than L2 reading ability, in spite of the fact that many of the students were highly proficient at Dutch. He therefore says:

> *It is simply not true that the influence of L2 knowledge is limited to the initial phases of the L2 acquisition process: L2 knowledge is strongly related to L2 reading comprehension even in advanced learners.*
> (Bossers 1992:185).

There are many similarities between these and McLeod and McLaughlin's results. The fact that Cziko's advanced students, however, appeared to process reading in a native-like manner, could be due to the different levels of proficiency of the 'advanced' students in the two studies. Possibly Cziko's students were more advanced, and therefore had more native-like proficiency.

However, many criticisms can be levied at the Clarke, McLeod and McLaughlin and Cziko pieces of research. Firstly, Clarke's results are compromised by his choice of reading test: although some cloze tests have been shown to correlate highly with reading test scores (see Oller and Conrad 1971), cloze tests demand productive as well as receptive skills, and are not therefore suitable for assessing reading ability (see also Alderson 1983). Secondly, the number of subjects in Clarke's miscue analysis is too small for generalisations to be made – there were only two subjects – one good and one poor reader - and these two readers were not necessarily typical of other good or poor readers. Finally, and most importantly, as Cziko acknowledged, miscue analysis is not appropriate for identifying the reading processes of second language learners. The technique is based on asking subjects to read a text aloud, which is a very different activity from silent reading, and may induce the reader to attend more to the technical aspects of the task than to the meaning of the text. This is so for L1 speakers as well, of course, but in the case of L2 readers the problem is exacerbated by the subjects' worries about pronunciation. Most second language learners, unless they are very advanced or very confident, are self-conscious about their pronunciation, and in a reading aloud exercise are likely to concentrate on the sounds they are producing rather than the meaning of the text (see comments in McLeod and McLaughlin 1986). Probably what these studies show is that at some stage in their L2 learning readers become sufficiently confident of their pronunciation to be able to spare some thought for the meaning of what they are reading.

In the next section I shall describe some studies into the effect of background knowledge on L2 comprehension. Many of these, too, use elicitation techniques which are not necessarily appropriate for their purposes. Apart from cloze tests and miscue analysis, and recall protocols, which I have already discussed in Chapter 2, the most common techniques seem to be multiple-choice and open-ended comprehension questions, summaries and verbal introspections. Some of the problems with these will emerge as the studies are described, but one general point must be made. In many of the articles in which these studies are reported the authors do not give examples of the elicitation techniques used, and in the majority of the reports there are no descriptions of how, if at all, the techniques were validated. Most of the articles do give inter-marker correlations for subjectively marked recall protocols (all of which are over .90), but none of the studies described here gives reliability indices for objective tests. How then can the reader know whether research results are due to the independent variable, or to inconsistencies in the elicitation technique? This lack of test validation seems to be a general weakness with research into reading, and is strongly criticised by Bernhardt (1991:67).

In addition to the above problems it is difficult to draw any general conclusions from the following studies as they involve so many different variables. For example, not only are ages, and the linguistic and cultural backgrounds of the subjects in the various experiments different, but the experiments relate to the learning of different second languages: English, German, Spanish, French. In addition, the experiments are underpinned by no universally agreed definitions of levels of linguistic proficiency, and so the proficiency levels of the different groups of subjects are in no way comparable. Students who are described as 'intermediate' by one researcher may be 'advanced' according to another. Some research reports do cite recognised ratings such as those of the Science Research Associates Reading Lab (for example, Hudson 1982), but most classify their students according to a local placement test, or the level of the students' language class. Even in the case of studies which report TOEFL proficiency scores (for example, Floyd and Carrell 1987), it is not clear how useful these scores are, as they relate to levels of general, rather than reading, proficiency. (For more about the lack of direction in L2 reading research, and the problems of lack of compatibility, see Bernhardt 1991.)

There is also the problem of how to define subjects' background knowledge, and here again Bernhardt (1991) feels very strongly. She is concerned about the assumptions L2 researchers make about the knowledge of their subjects, and the fact that they are inclined to assume that all members of a group have the same background knowledge. For example, she cites a study in which 50 Catholic Spanish learners of English are compared with 50 Islamic Arabic-speaking learners, and points out that the fact that these people have the same general cultural background knowledge does not mean that they have the same background knowledge; she says that 'assuming knowledge or lack thereof on the basis of ethnic heritage is a rather naive view of knowledge' (1991:95). Bernhardt says that researchers into reading processes should take more care to assess the different types of knowledge that readers bring with them, and she tried to do this herself by asking students to free-associate, that is to write a list of everything they could think about a topic, before they were given the related text to read. From scanning the lists she was able to get a good idea of the students' knowledge about the topic, but unfortunately the three-point scale which she used to mark these free associations was too crude to distinguish adequately between the different levels of knowledge, and she was not therefore able to come to any conclusions about the effect of this content knowledge on the students' reading comprehension. Some of the following examples of research do check the students' background knowledge fairly carefully (for example, Levine and Haus 1985), but it will be seen that some follow the approach that Bernhardt so much condemns.

The Effect of Context and Background Knowledge on L2 Reading Comprehension

It is difficult to classify the following studies in any logical way since their aims do not dovetail neatly, but their main concerns lie with the effect on L2 comprehension of one or more of the following: cultural, content or formal background knowledge; the effect of activating schemata by supplying context or topic related information; and the rival effects of linguistic versus background knowledge. Some are concerned with comparing the performance of L1 and L2 readers, and others with comparing the performance of L2 learners at different levels of proficiency. Most of the studies are based on specially created, somewhat abnormal texts, but a few are based on authentic ones.

Threshold Levels

Hudson (1982) accepts Clarke's (1980) short circuit hypothesis, but argues that students can override these short circuits if they are encouraged to call up the relevant schemata. He gave reading texts to academic students in advanced, intermediate and elementary ESL reading classes. The classes were divided into three groups, each of which was given reading passages followed by reading comprehension multiple-choice tests. Before reading the passages, one group at each level was introduced to the passage's reading topic by means of pictures and discussion, one was introduced to the relevant vocabulary, and the third read the texts without any introduction. The treatments had different effects according to the students' reading levels. For example, the elementary students who were given the schemata-activating preparation did significantly better than those given vocabulary training. On the other hand, the schemata-activating training had no effect on the advanced learners. Hudson concluded that the advanced learners already had systems for calling up the relevant schemata, and so were unaffected by the preparation tasks. He therefore deduced that much of the research into L1 effects of schemata is applicable to L2 learning, and that there is a short circuiting problem with low level language learners, but that this can be at least partially overridden by activating schemata which are otherwise inaccessible.

Laufer and Sim (1985) were interested in finding if there was a threshold level below which students were not sufficiently proficient at the L2 to be able to understand academic texts. They asked students to answer questions in their mother tongue, Hebrew, about three academic texts in English, and then to retrospect about how they arrived at their answers. In order to identify the students' threshold level, they gave the passages to groups who were progressively more proficient in English. They found that there was a threshold level below which the students tried but failed to use their good L1 reading skills to interpret the English texts. What appeared to prevent the lower level students from understanding the passages was not their knowledge of English structure, but firstly their limited vocabulary and secondly their background knowledge. These findings contrast with Freebody and Anderson's (1983) research into L1 users, where they found that knowledge of vocabulary did not necessarily improve understanding of a text. The difference here must lie in the fact that proficient L1 speakers recognise almost all the vocabulary in a text already, and are therefore able to guess the meaning of any unknown words from context. As Johnson (1981) says:

> *It may take a high percentage of difficult vocabulary items to have a significant effect on readers' comprehension of a text. The normal redundancy in a text may enable readers to cope with unfamiliar words without too much disruption in their understanding. Readers seem to be able to construct a text from memory based on inferences made while reading. This familiarity with the topic of the passage and general background knowledge of the theme may allow the reader to construct highly plausible meanings for unfamiliar vocabulary words.*
> (1981:54)

Low level L2 speakers, on the other hand, know so little of the vocabulary that they are not in a position to make use of context.

Research Based on Recall Protocols

In spite of the differences between memory and comprehension, many researchers have used recall protocols to investigate L2 readers' comprehension processes.

Carrell (1983) looked at the reading skills of ESL readers at different levels of proficiency, and discovered that neither 'high-intermediate' nor advanced ESL learners processed text in the same way as native speakers. She identified three different kinds of background knowledge, and found that native speakers used all three kinds in their processing of text: context (prior knowledge of what the text would be about), transparency (the presence or absence of specific, concrete lexical items such as 'the clothes' instead of the non-transparent 'the things') and familiarity (readers' familiarity with the subject matter). She gave a linguistically heterogeneous group of advanced and upper intermediate ESL university students adaptations of Bransford and Johnson's (1973) Washing Clothes (described as a familiar topic) and Balloon Serenade (an unfamiliar topic) texts to read, with or without a title or picture, and with or without transparent terms. Each student was given one version of each of the texts, one with or without a context, and one with or without transparent textual clues. After reading the texts the subjects were asked to write down, in English, all they could recall of the passages. The text recalls were double marked and scored according to the number of idea units present (idea units 'correspond to simple sentences, basic semantic propositions or phrases' [Carrell 1983:183]). The results showed that the native speakers recalled the familiar text with context and the familiar text with transparent terms better than the other versions. Familiarity, context and transparency all had a significant effect on recall scores. On the other hand the advanced and the intermediate ESL students were not affected by either context or transparency, and the intermediate students were not affected by familiarity either. Carrell concluded that the ESL learners were not efficient at either top-down or bottom-up processing because neither context nor textual transparency gave them enough help.

Lee (1986) suspected that asking subjects to recall the texts in English rather than in their native tongue might conceal some of the processes taking place during the reading, so he replicated Carrell's study, translating the texts into Spanish, and asking the subjects, third year undergraduates learning Spanish, to write their recalls in their mother tongue – English. This time the effect of context, the interaction of context and familiarity, and the three-way interaction of context, familiarity and transparency all had significant effects on the students' recalls. Lee concluded that all three kinds of background knowledge affected the subjects' reading comprehension, but that they did not do so uniformly. The interaction was very complex, and Lee presumed that it was only by asking the students to recall the texts in their own language that this complexity emerged. There were, of course, other differences between the two studies which may have led to the different results: the students had different linguistic and cultural backgrounds, the texts were presented in different languages, and the levels of second language proficiency of the two samples were probably different. However, Lee's point about recall tasks in the second language is important. In a second language only highly competent students, probably too competent to be attending English courses, would voluntarily go into much detail in their recalls. Most students would tend to keep their recalls short and simple in order to avoid production difficulties in the second language, and would thus omit much of the material that they had remembered.

Donin and Silva (1993) reinforce Lee's views about recall in the L2. They gave student nurses who were learning French at an intermediate level three medical texts which were presented in either English, which was their mother tongue, or French. Each student read one text in English, and two in French. They recalled the English and one of the French texts in English, and recalled the other French text in French. The students' reading times were slower in French than in English, and their French recalls implied that they were reading in a bottom-up manner. However, when these same students did their recalls in English, they made use of their medical background knowledge to understand the texts and appeared to be using inferencing in the same way as they did in their L1.

Bransford and Johnson's invented texts have been used for many studies into the effect of context on reading. Roller and Matambo (1992), used the Washing Clothes and the Balloon Serenade in a partial replication of Carrell's and Lee's studies, and Wolff (1987) gave German secondary school children four texts of which one was the Balloon Serenade. In Wolff's experiment, the passages were presented on video, and in some cases were accompanied by illustrations relating to the text, and sometimes not. The children watched the videos and were asked to recall what they had just heard in German. Wolff found that the inclusion of illustrations in an easy text made no difference to the students' recalls, but that in difficult texts, for example the Balloon Serenade, it improved recall. For these difficult texts, therefore, Wolff argued that the students must have been using top-down processing. Wolff confirmed his findings when he interviewed the students after their recalls, and asked them to report, in German, on their listening processes. One of the texts concerned a marriage, and students who did not realise this wrongly interpreted it as concerning a birthday party, or a cowboys and Indians film, or drug addiction. Once these students had decided on a particular context, they kept to it, doing their best to make the whole story fit the script they had chosen. This seems a clear example of top-down processing, and also shows the useful insights into reading processes that can be achieved by using qualitative data such as verbal introspections.

Some findings in Carrell's and Roller and Matambo's studies do not agree with other studies into the effect of background knowledge on comprehension. According to the researchers' expectations, students would remember the familiar topic better than the unfamiliar one. However, the native speakers in Carrell's study, and the non-natives in all three studies remembered the unfamiliar topic, the Balloon Serenade, better than the familiar topic, Washing Clothes. The Balloon Serenade describes a modern-day Romeo using balloons to carry a microphone up to a high-rise flat so that his beloved can hear the music from his guitar. It is more fun than the Washing Clothes passage. It could be for this reason that this topic was remembered better than the other; it is in Carrell's terms 'more salient'. It may also be that interesting passages are sometimes more memorable than familiar but dull ones, although this does not, of course, mean to say that they are necessarily easier to *understand*. On the other hand it could be that the students' existing washing clothes schema interfered with their understanding of the contents of the Washing Clothes text (see Lipson 1984). Or, as Roller and Matambo suggest, it could be that the Balloon Serenade was more easily remembered than the other because it contained more concrete nouns which were easily 'imageable' without a picture, and that the text had a more consistent formal structure than that of the Washing Clothes text. Roller and Matambo suggest that the familiar formal structure of the Balloon Serenade passage together with the use of easily imageable concrete nouns made this text easy to recall using rote memory. Clearly, if this is the case, there must be a complex interaction taking place between topic, text structure and linguistic features.

Although in Wolff's experiment the students understood the Balloon Serenade better when it was contextualised, thus supporting Bransford and Johnson's original conclusions about the role of context in comprehension, in Carrell's, Lee's and Roller and Matambo's studies they did not. In all three studies the contextual clues helped students understand the Washing Clothes text but not the Balloon Serenade, and in the Roller and Matambo study the contextual clues actually made comprehension of the Balloon Serenade more difficult. There are various possible explanations for this. Firstly, it is true that the title – Balloon Serenade – is not very enlightening and might not clarify the text. Secondly, in the case of Roller and Matambo's study it may be that the students do not have a serenading schema and therefore would not understand the picture. This would support Lee's contention that the picture itself is so obscure that it cannot help explain the text. However, I am not convinced that this is the whole answer. Unless the illustration was more badly drawn in all three studies than it was in the original Bransford and Johnson research, it is difficult to believe that, at least for students with a European culture, it would fail to elucidate such an initially incomprehensible text. Perhaps this is a case where the picture improves comprehension of the story but not memory.

The Effect of Contextual Clues on the Comprehension of Unknown Words

Adams' research (1982) used an adaptation of Washing Clothes among others in an attempt to find out whether the presence of contextual clues improves the understanding of unknown words. She devised six passages in which target words related to the topic were replaced by nonsense words. The passages were written in French and translated into English. American college students were given these texts in one or other of the languages. Half the students in each language group were told what the texts were about – the others were not. The results showed that the students who were told the context understood more of the nonsense words than those who were not. There was no significant interaction between the language in which a text was written and the presence or absence of the contextual clues, and Adams suggested that this was because the clues would be more useful to low level than to high level students. The higher the language proficiency of the students the less impact the clues would have, as the students were increasingly more able to create a context from linguistic cues in the text. This would, of course, agree with Hudson's (1982) findings (see above). Unfortunately Adams did not know the comparative proficiency levels of her subjects and so was unable to test this hypothesis.

The Use of Authentic Texts

In most of the above studies, the research was based on ambiguous or invented texts. Levine and Haus (1985) criticised this approach, and used a genuine newspaper account of a baseball game. Basing their research on Anderson *et al.* (1977a) and Chiesi *et al.* (1979), they gave the passage, which was in Spanish, to ninth grade children in two different ability Spanish classes. They tested the children's understanding with twelve multiple-choice questions of which four were factual and were designed only to require direct recall from the passage, and eight were 'scriptally implicit' and were intended to require inferencing as well as reference to the text. The students' prior knowledge of baseball was also tested, and the subjects were classified as HK or LK. The HK students did significantly better at all the questions: for the direct questions $p = < .05$, and for the inferencing ones, which were designed to require background knowledge, $p = < .01$. The level of language proficiency had no effect on the answering of the direct questions, but did on the inferencing questions ($p = .01$), and there was a significant interaction between language level and background knowledge. Although background knowledge helped all the students, it appeared to help the higher proficiency level students more.

Kozminsky and Graetz (1986) also used authentic texts. They gave Hebrew-speaking psychology students attending advanced English courses a sociology text presented in either English or their mother tongue, and asked them to make notes before summarising the text in Hebrew. The authors found that the quality of their notes and their summaries were much poorer in English than in Hebrew, and that they showed evidence of high knowledge but low language skills. The students appeared to use top-down skills, but were unable to integrate these fully with bottom-up processes. The authors agreed with van Dijk and Kintsch (1983) that 'what is really wrong with poor readers is that they recognise isolated words inaccurately and too slowly, and compensate for their lack of decoding skills with context-dependent guessing or hypothesis testing' (van Dijk and Kintsch 1983:24).

The Effect of Cultural Background Knowledge on Comprehension

Wolff's experiment (1987, see above) produced a nice example of students first activating a schema, and then clinging to it tenaciously. Other researchers have studied the effect of specially activating readers' schemata, much in the same way as Ausubel introduced his advance organisers (see Ausubel 1960 and also Hudson 1982, above). Carrell, Pharis and Liberto (1989) were interested in activating background knowledge by giving students extra tuition in the topic areas related to their reading material. They first gave EAP students a pretest containing three reading passages with three multiple-choice comprehension questions and two open-ended questions. Two experimental groups were given different kinds of schema activation training while a control group was given none. When the reading passages and the test were administered a second time there were no significant gains in the scores of the multiple-choice items, but both experimental groups showed significant improvement in their open-ended answers while the control group did not. Once again it may be that the choice of test method affected the results. The multiple-choice questions may not have been subtle enough to elicit the sorts of answers which would distinguish between activated and non-activated schemata. On the other hand, it may be that because there were so few questions, the differences in the results were due to chance.

Floyd and Carrell (1987) studied the effect of activating students' *cultural* schemata. EAP students of different nationalities were given a letter about the 4th July celebrations in Boston, a town which none of the students had visited. There were two versions of the letter: both included the same idea units, but they varied in syntactic complexity. Half the students in both the experimental and the control groups read the letter in its complex version and half in the simple form. All students were asked to answer some multiple-choice questions which were designed to test their ability to draw cultural inferences, and they also had to produce a free written recall of the text. In spite of Lee's criticisms of the earlier Carrell study (Lee 1986) these recalls were written in English. The students in the experimental group were later given a training session in which they were shown a map of Boston, and discussed 4th July celebrations in general. These students also went to a typical American barbecue. The control group had no further introduction to the 4th July celebrations, and no barbecue. After the experimental group's training sessions all the students were given the same version of the letter and were asked to answer the questions and recall the passage again. The experimental group's scores on the re-test and the recall task improved significantly more than the control group's, and this effect held good for both versions of the letter. The authors concluded that background knowledge, and this could be either cultural or content, had more effect on the results than syntactic complexity.

These findings agreed with two previous studies by Johnson (1981, 1982) who wanted to discover whether a text's level of linguistic complexity or its cultural content had more effect on ESL learners' reading comprehension. In the first study, Johnson gave Iranian and American students two reading passages, one an Iranian folktale, and the other an American story. The two stories were each presented at two linguistic levels, one being a simplified version of the other. Multivariate analysis of the students' recalls and answers to multiple-choice questions disclosed that the level of semantic and syntactic complexity of the texts had less effect on the ESL students' recalls than did the cultural background. Similarly, in study two, where a heterogeneous group of EAP students were given a text of which the first half was culturally familiar and the second half was not, and where only some students were given help with the intentionally difficult words, the students' recall protocols and sentence recognition answers showed that familiarity with the content had a greater effect on recall than did the glossing of the vocabulary.

The Effect of Formal Schemata on Comprehension

Carrell has also turned her attention to the effect of formal schemata on ESL reading comprehension. In one study (Carrell 1984a) she based her research on Meyer and Freedle's (1984) experiment. EAP students were asked to read a text which had been modified in one of four ways. Each version of the text was written in one of Meyer's (1984) discourse types: 'collection of descriptions', 'causation', 'problem/solution', and 'comparison'. Carrell found that the style in which each text was written led to significant differences in the number of idea units recalled, and, as with Meyer and Freedle's study, the 'collection of description' text was generally less well remembered than the other three types. Carrell (1985) has since shown how ESL reading comprehension can be improved if students are explicitly taught about expository text structure.

In another study, Carrell (1984b) found, like Thorndyke (1977), that students who were familiar with the form of an English folktale understood such tales more easily than those who were not. However, since she accepted that her results could have been due to content rather than formal knowledge, she later attempted to tease out the different effects of formal and content schemata (Carrell 1987a). She gave two groups of high intermediate EAP students, of which one was Moslem and the other Roman Catholic, the biographies of two little-known religious personalities. Both the texts were written in a historical narrative style. All students read both biographies, but half the students in each group were given versions of the biographies in which the events were no longer in chronological order. Students had to answer 14 multiple-choice comprehension questions and recall the texts in English. Content knowledge proved to be a stronger predictor of performance than either rhetorical structure or the interaction between the two. However, there seems to be a grave problem with this research: the supposedly unfamiliar rhetorical style is not a style at all – it is just a jumbled-up version of a genuine style, presenting what Garnham (1985), describing other jumbled order studies of this kind, describes as a 'bizarre sequence of events' (1985:177). What the research seems to show is that if readers have prior content knowledge they are able to make sense of a text even if its propositions are presented out of order. This may not be exactly what Carrell intended to show, but it does emphasise the importance of content knowledge in reading comprehension.

Overall Conclusions about Reading in a Second Language

From the above section it has emerged that, in spite of my reservations about the elicitation methods involved and the lack of validation processes, there has been enough agreement about how L2 readers process text for some conclusions to be drawn. For example, it seems clear that there is a certain level of language proficiency below which L2 readers are unable to use their L1 reading strategies efficiently, although it is not certain whether low level students use predominantly top-down or bottom-up processes. It does seem that L2 readers use or attempt to use many of the same processes as native speakers but, because of their decoding problems, 'the harmony between top-down and bottom-up processing is disturbed' (Wolff 1987:313). There is also evidence that for L2 readers as well as L1, background knowledge, whether related to culture, content, or formal structure, plays an essential part in reading comprehension, though cultural and content knowledge may be more important than knowledge of form. Indeed Johnson (1981), Hudson (1982), Laufer and Sim (1985) and Floyd and Carrell (1987) suggest that cultural and content background knowledge is more important than syntax. This may be the case for advanced level learners, but at lower levels of language learning knowledge of syntax might be expected to have a much stronger effect. Indeed Bernhardt says: 'As a reader's linguistic knowledge grows it begins to override knowledge-driven inferencing. In other words, a reader begins to rely more on the *language* and less on what he/she *thinks* the language contains' (1991:170). However, there is not enough evidence yet to show whether she is right.

One set of studies produced results which ran against the main trend. The studies based on the Balloon Serenade lead me to wonder whether there is another factor which has at least as important an effect on comprehension and memory as does background knowledge, and that is salience or novelty. If a text is particularly interesting or unusual is it more memorable than one which is familiar but dull, and if this is the case how does this fit in with schema theory? It would be interesting to see the results of a study specifically designed to study the comparative effects on comprehension and memory of salience and prior knowledge.

Outstanding Research Issues

The research described in this chapter shows that background knowledge does affect the reading comprehension of ESL readers, but it also shows that comparatively little is known about the manner in which it does so. The rival effects of content and formal knowledge, for example, need more study, and so does the effect of familiar as compared to 'salient' passages. The present study, however, is devoted to the three following issues, which will be discussed more fully in Chapter 6:

1 If it is accepted that background knowledge does have some effect on reading comprehension, should this be explicitly taken into account when EAP proficiency tests are devised? Should students intending to study in different academic areas be given reading tests in these different subject areas, so that they are not disadvantaged by a lack of appropriate background knowledge?

2 Although Johnson (1981) and Floyd and Carrell (1987) found that background knowledge appeared to have more effect on comprehension than did syntactic or lexical knowledge, the EAP studies reviewed in Chapter 1 imply that students' levels of language proficiency have at least as important an effect on test performance as does background knowledge. This study will therefore research the comparative importance of language proficiency and background knowledge.

3 The third, related, question arises from the fact that it is not clear whether students with low levels of language proficiency depend more on background knowledge for their interpretation of written texts than do high level students. Clarke (1980), Cziko (1980) and McLeod and McLaughlin (1986) all think that at the lower levels students are unable to decode written language sufficiently to be able to bring top-down processing to bear, but Wolff (1987) found that low level students used background knowledge to help to make sense of incomprehensible material, and Bernhardt (1991) suspects that background knowledge becomes less important as students become more linguistically proficient. My third aim, therefore, is to see at which levels of proficiency students seem to be most affected by their background knowledge.

Bernhardt's 1991 book, *Reading Development in a Second Language*, which has already been much referred to, devotes many of its pages to a critical review of recent research in Reading in a Second Language, and the author makes detailed recommendations for how future research should be carried out. I shall list four of these here, and then discuss them in more detail. In brief, Bernhardt recommends that:

1 Basic research 'should be conducted on second language readers delineated by native language group'.

2 The research should be 'based on authentic materials'.

3 Multiple measures should be collected on each subject involved in the study.

4 Reports of research studies should give details of the procedures followed. (1991:225).

Language Group

It is true that students with different linguistic backgrounds may process reading in different ways (see, for example, Carrell 1984a), and it is true that if members of these different backgrounds are grouped together these differences may obscure important reading effects. It is certainly important, therefore, that at least some second language research should be carried out on linguistically homogeneous groups. However, the choice of sample must depend on the purpose of the research. If its aim is to enquire into aspects of the reading process which are considered to be universal, it makes perfectly good sense to use heterogeneous samples, as it does if the purpose of the research relates to the testing of multinational EAP students. However, since variation in the subjects' first language and culture may obscure findings, it is desirable, where possible, to study one or two subsets of linguistically homogeneous subjects as well. For this reason, in my pilot study (see Chapter 5) I not only investigate the reading comprehension of a heterogeneous group of ESL students, but also of a group of linguistically homogeneous Indonesian students.

Authentic Materials

The use of anomalous, artificially designed texts like those by Bransford and Johnson cannot give us full knowledge of how readers process 'real' texts, but the choice of text type must depend on the purpose of the research. Adams (1982), for example, was interested in the effect of background knowledge on the ability to guess lexical meaning. She researched this by inserting invented vocabulary into a suitably adapted text which was perfectly appropriate for her study. However, if the research is specifically concerned with the comprehension of natural texts, then it should be based on such texts. In the research into EAP reading described in Chapter 1, all the reading passages were authentic, as indeed they are in my own study.

Multiple Measures

Since there is no perfect measure of reading comprehension, and since different testing techniques affect results, it is only sensible that, where possible, multiple measures should be collected from each subject involved in a reading comprehension study. Bernhardt recommends a combination of four techniques: recall protocols, a grammatical achievement test, responsive writing and a retrospective interview. As I have already shown, I have reservations about recall, since it depends on memory, and writing tasks are unsuitable because of the impossibility of disentangling productive from receptive skills. However, a range of different types of comprehension item should be used, and, where possible, these should be supplemented by introspective or retrospective interviews. In my own study there is a variety of item types.

Details of Research Procedures

Bernhardt says that researchers should report their studies in more detail. She criticises the fact that many research papers include incomplete information about the subjects in the experiments, the tasks, the responses and the methods of data analysis. She argues that experimental studies into reading as a second language are now so complex that they cannot be adequately covered in ten-page reports. I can only agree.

4 The Development of the IELTS Reading Modules

The reading tests used in this study come from the International English Language Testing System (IELTS) test, the latest in a British Council sponsored series of test batteries designed to assess the English proficiency of students wishing to attend English medium universities and colleges. The major part of this chapter focuses on the construction of these tests, but in order to understand the factors that affected their design it is necessary to know something about IELTS's predecessors. The first section of this chapter, therefore, gives a brief description of these, tracing the development of the reading components, and explaining why users' reactions to the rationale and practicality of the test battery which directly preceded IELTS led to the decision that IELTS should have reading modules in three different subject areas. The second section describes the construction of the first versions of the IELTS reading tests, including the design of the draft specifications and tests and their subsequent content validation. This section is intended to show what steps were taken to make the three reading modules as appropriate as possible for students in the relevant subject areas. The third section briefly reports the piloting and trialling of the reading modules, and the final section describes the tests' underlying construct.

The Precursors of IELTS

The English Proficiency Test Battery (EPTB)

In the early 1960s British universities and colleges became dissatisfied with the level of English of many of their overseas students (Davies 1965). A set of procedures called the British Council Subjective Assessments, for which British Council officers produced their own test materials, was not screening students satisfactorily (see Moller 1981 and Alderson and Clapham 1992), and the British Council therefore commissioned a language proficiency test, the EPTB, which was to identify students who would not be able to cope effectively with their academic studies. This battery was constructed by Alan Davies, and was introduced in 1964. It included discrete point tests of grammar and sound discrimination, which were presented in a multiple-choice format so that the test could, when necessary, be administered and scored by unskilled examiners. These tests belonged to what Spolsky (1977) called the psychometric-structuralist era of language tests. However the EPTB also contained two reading tests which did not fit so easily into the discrete point testing mould. One of these was a test of reading speed, the 'intrusive words' test, where misfitting words had been inserted at random positions in a reading passage, and candidates had to identify as many of these words as possible in a given time. (This test method has since been called cloze-elide, and is described in Manning 1987.) The other reading comprehension test contained two passages on 'general topics'. In each of the passages candidates had to restore selected words which had been partially deleted. These reading tests might be described as integrative rather than discrete point since, instead of aiming at testing one linguistic element at a time, they required candidates to use several aspects of their linguistic knowledge concurrently. For the deletion tasks, Davies had wanted to use texts which were similar to those that the students might meet in their studies, and since he found that there were gross differences in the texts used in scientific and non-scientific university courses, he initially devised two versions of this test, one for science and one for non-science students (Davies 1965:64). However, since he did not trial these two tests on a single sample of students, he had no data on the comparative difficulty of the two versions. When, therefore, trial centres complained that the tests were not of equivalent difficulty, and when a trial on a sample of native English speakers bore this out, these two forerunners of subject specific testing were abandoned. It is interesting to note that by taking into account students' need to read quickly, and by acknowledging that students in different disciplines read different kinds of texts, Davies was looking towards more specialised testing. Indeed he recommended (Davies 1965) that linguists should provide descriptions of the English used in varying academic situations, so that testers could take this into account when writing academic proficiency tests; he thus paved the way for the needs analyses of the 1980s. (For more about the EPTB, see Davies 1965, Davies and Alderson 1977 and Moller 1981.)

The English Language Testing Service (ELTS)

During the mid 1970s, as changes in linguistic theory and teaching took place, and the emphasis on discrete point structures in teaching and testing gave way to notions, functions and communicative competence (see Hymes, 1972), the EPTB came to be considered outdated:

> *In the latter half of the 1970s the British Council was faced with the need to introduce a new or modified English proficiency testing system geared to the changes in ELT developments, notably in ESP, and to the changes in the needs of sponsored students seeking to come to England.*
> (Seaton 1981:121)

In 1977 six teams of teachers and consultants were commissioned to draw up plans for a new test, the English Language Testing Service (ELTS) test. I shall describe this test battery in some detail since IELTS is a revised version of it, and since a discussion of its design raises many issues which are relevant to this study.

In 1978 Brendan Carroll presented the British Council with the test designers' report, *Specifications for an English Language Testing Service* (later published in Alderson and Hughes 1981). (To avoid confusion, it should be noted that the term 'Specifications' in this report refers to linguistic profiles of students. Although it makes proposals for the form and content of the new test, it is in no way a blueprint for it, and many of its proposals were subsequently dropped.)

In the introduction to these specifications, Carroll says:

> *... language teaching and testing methods have shifted their emphasis from atomistic language features, such as uncontextualised phonemic discriminations ('hit-pit'), to broader features of linguistic communication. The trend now is, as exemplified in the present report, to postpone consideration of language realisations until the communicative needs of the users have been clearly determined, broadly-speaking a sociolinguistic approach.*
>
> *(These trends) have also encouraged the development of programmes in English for Specific Purposes (ESP) so that fewer people are now engaged in devising tests and teaching programmes which aspire to meet equally well the needs of all users, regardless of the purposes for which they will need the language ... .*
>
> *Our problem is not just whether the present test can encompass the needs of these, and many other, diverse study courses, but whether* **any** *single test can do so. And we have adopted the hypothesis that the solution to our testing problem, and the way to improve the testing service, is through a process of diversification of test instruments to meet the diversity of the test situations.*
>
> *...In designing our testing service, then, we will need to specify the communicative demands of a variety of courses, of different levels, types and disciplines, and to devise workable instruments to measure how far applicants can meet those demands.*
> (Carroll 1981:66)

Carroll and his colleagues were therefore adopting an ESP approach. Using Munby's communication needs processor as a model (Munby 1978), they drew up profiles of six students whose language needs were supposedly typical of business studies, agricultural science, civil engineering, medicine, laboratory technicians, and academic social survival. The aim was to:

> *build up the profiles of the communicative needs of a number of students on study programmes in Britain in such a way that we will be able to identify common and specific areas of need upon which an appropriately diversified test design can be based ... .*
>
> *Each specification will provide information about the communicative needs each participant will have in studying his programme and in living in an English speaking community.*
> (Carroll 1981:66)

These specifications were used to draw up a network of the linguistic relationships between the six subject areas, and to show similarities and differences between them. However, instead of being based on empirical research, these profiles were based on the intuitions of a small group of language teachers. This is unfortunate since all the conclusions drawn about relationships between the six subject areas can only be considered speculative. (For further criticisms of the specifications, see Clapham 1981, Criper 1981, Hamp-Lyons 1987, Criper and Davies 1988.)

ELTS Construct

Since I shall later be discussing the construct of the reading component of IELTS, and since this has been affected by the theory underpinning ELTS, I shall briefly spell out the constructs on which ELTS appears to have been based. I say 'appears', because no final test specifications are available. However, as far as can be judged from the Carroll *Specifications* and from the contents of the test itself, the battery was based on three constructs which are described in Hamp-Lyons (1987). I shall describe these, and shall briefly comment on them and their implementation in the ELTS test battery. For further discussion, see Criper and Davies (1988) which I shall be referring to in more detail below.

1 Language is divisible into the four skills of reading, writing, listening and speaking. The skills are in turn divisible into finer language skills (or functions) such as 'understanding conceptual meaning' with its related micro-skills such as 'quantity and amount', 'comparison and degree' (Munby, 1978:126).

Interestingly, in his communication needs processor Munby does not accept this division of language into four skills:

> *... the profile of communication needs for a participant is interpreted in terms of the language and skills required for its realisation. The customary division of language skills into listening, speaking, reading, and writing ... and their treatment as macro-concepts, will not serve here.*
> (Munby 1978:116)

In the ELTS test the four language skills were tested separately. Carroll does not discuss this discrepancy between Munby's model and the test. Presumably he had to manipulate Munby's model to suit the proficiency test demands. However, this makes the matching between test items and Munby's macro- and micro-skills difficult. When researchers tried to relate the test items to Munby's taxonomy of language skills, they frequently found no direct relationship (Criper and Davies 1988). As is the case, I suspect, with many proficiency tests, it is possible that the test writers composed what they thought were good items, regardless of what the specifications (if any) required. And indeed, even if they had rigorously followed the specifications, there might have been different interpretations of what they were testing. (Alderson 1993b, has shown that experienced testers and language teachers often disagree, sometimes wildly, on what a given item is testing.) In Chapter 6 we shall see whether the IELTS reading test items appear to be any closer to their test specifications.

2 Basing his findings on the six intuitive profiles described earlier, Carroll describes some of the factors underlying the language proficiency of the future ELTS candidates:

> Factor I: *'general' factor, accounting for a sizeable proportion (perhaps half) of the variance, representing the common communicative requirements and characteristics (intelligence, motivation, academic aptitude) of all participants.*
>
> Factor II: *an 'academic study' factor reflecting the ability to use the communication/language skills necessary for handling academic discourse of a relatively neutral attitudinal nature.*
>
> Factor III: *a 'personal relationships' factor representing the non-study relationships with contacts in field or clinical work.*
>
> Factors IV+: *specific or small-group factors representing the special additional requirements of odd-man out programmes.*
> *(Carroll 1981:83)*

In a large-scale testing programme it is not feasible to take account of all the factors, but the first three factors should be testable. However, in ELTS, only Factors I and II were tested. The battery took account of Factor I by including a General component which was taken by everyone, and which assessed general reading and listening skills. For Factor II there was a study skills reading test which examined such skills as the ability to read an index and to understand a bibliography.

3(i) The test is based on a needs analysis of students in different academic and non-academic fields of study; this analysis is modelled on Munby's communication needs processor.

Whereas in Munby's needs processor each student profile is specific to that particular student, each of the profiles in the ELTS *Specifications* covers the needs of many students in a range of loosely related subject areas. This was necessary, because financial and practical constraints made it impossible for the ELTS battery to design tests for each academic discipline, but it does raise doubts about the suitability of the Munby model, where the description of the needs of any 'participant' are specific not only to the discipline, for example 'agriculture', but also to the 'central area of study', for example 'cattle breeding' (Munby 1978:205).

3(ii) Since the language used in different disciplines varies, there should be different tests for students in these different disciplines.

Carroll cites little proof for the above. Although he found much in common in the language used in all the profiles in his study:

> *... it still remains that the spoken and written discourse of the ... disciplines are very different indeed; their linguistic and diagrammatic realisations have very different appearances. Can we then test different disciplines with identical test material, selected to test their common communicative requirements? Or will we, in so doing, use over-generalised language/diagram realisations which may favour candidates in one particular discipline or, worse still, be equally irrelevant to all the disciplines? We are not yet in a position to answer these questions, so we propose to continue in a pragmatic fashion by preparing tests in different disciplinary areas.*
> (Carroll 1981: 82)

ELTS therefore contained academic modules in five different subject areas, and there was one general academic module for students whose area of study was not covered by any of the other five (see Table 4.1.).

After describing the linguistic relationships between the different subject areas, Carroll also says:

> *The conclusion we draw from these relationships is a perfectly clear one,* ***that language skill requirement patterns cut right across disciplinary boundaries*** *(author's emphasis)* (1981:82).

This is pure speculation on Carroll's part, but if he is right it would seem, as Hamp-Lyons (1987:181) says, to argue against the inclusion of the different subject modules.

The Structure of ELTS

The final form of the academic version of ELTS is shown in Table 4.1.

Table 4.1

Contents of the ELTS Test Battery

G (General)	M (Modular)
G1 Reading	M1 Study Skills
G2 Listening	M2 Writing
	M3 Interview
	Modules
	Life Sciences
	Medicine
	Physical Sciences
	Social Studies
	Technology
	General Academic

All students took the General tests, G1 and G2, and the three modular tests in the academic subject area nearest to their own.
G1, G2 and M1 were multiple choice. M1, M2 and M3 were based on an academic source booklet.
Scores were reported on a 1-9 band scale, with a profile score for each of the five subtests.
An overall band score gave the mean band score of the five subtests.

The ELTS Validation Study

ELTS was introduced in 1980 and was generally well liked. However, since there was little known about its validity and reliability, the British Council and the University of Cambridge Local Examinations Syndicate (UCLES), who were joint sponsors of the test, commissioned Edinburgh University to conduct a validation study (see Criper and Davies 1988). This study covered aspects of the test's practicality, reliability and validity which are too extensive to be described here. However, two of its recommendations which had a direct impact on the ensuing ELTS revision were that the test should be simplified, and that it should take less time to administer. The report also made several comments about the academic subject modules. For example, it reported that examinees had difficulty selecting appropriate modules:

> *In our evidence there is clear indication of the difficulty testees (and administration) have in making an appropriate choice. And yet if this matching is problematic, much of the rationale for the complexity of ELTS disappears. The principle underlying ELTS is that true English proficiency (i.e. a learner's 'true score') is best captured in a test of specific purposes. If it is the case that matching student to module (or testee to test) is so uncertain, then ELTS loses the very advantage it was designed to maximise. Only if there is reasonable certainty about such matching, viz., that the*

> *testee is provided with a test which is recognisably appropriate for him and which he agrees is appropriate, can matching be properly taken advantage of.*
> *Two comments are relevant here. First that the number of mismatches in terms of testee unease or administrative perplexity may be so small that the problem disappears as noise in the system. The second is that, if we take seriously the claim of matching (that a testee's 'true score' is best/most truly achieved by a test of specific purposes) then we are allowing a proliferation of test types which can only lead to a situation of one test–one testee. Such an outcome is not only a denial of the group function of tests, it is also an invitation to wholesale impracticality which has been called 'pseudo-procedure', i.e. a device for improvements which can never be realised.*
> *We escape from the throes of logical and practical dilemma pragmatically thus: never mind the implications of a matching principle, rather secure a workable test (in terms of time, materials, organisation) which goes some way towards fulfilling the matching principle. How far it goes depends on these two factors; the practical organisational one and the statistical configurations which indicate gains in prediction. In other words, there is really no point in maintaining the present ELTS structure (6/7 module choices) since they are expensive in practical terms.*
> (Criper and Davies 1988:108)

At a consultative meeting to discuss this validation study (see Hughes, Porter and Weir 1988) it was agreed that ELTS should be revised, but that this revision should not be too radical. The test should if possible be shortened, its administration simplified as much as possible, and the reliability of marking increased.

IELTS - Designing the Reading Modules

Collecting Views from Advisers

In August, 1986, the ELTS Revision Project, under the directorship of Charles Alderson, was set up to control the design and construction of the revised test battery. (For a list of the members of the Project Steering Committee see the Acknowledgements.) I joined the project in January 1987 as Research Co-ordinator, and from then on was concerned either centrally or peripherally with all the stages of the revision process. When I refer to the activities of members of the Steering Committee after January 1987, I shall therefore use the pronoun 'we'.

During November and December, 1986, the ELTS Revision Committee sent out questionnaires to ELTS users such as test administrators, EAP teachers and testers, and receiving institutions such as universities and colleges, asking for comments on ELTS and for advice about the content and format of the future test battery. After the results of these questionnaires had been analysed we interviewed British Council Headquarters staff, and held discussions with EAP teachers and testers. We also studied 1,000 past candidates' test report forms since these showed, among other things, what disciplines the candidates intended to enter, what degree or qualification they were seeking, and which test module they had taken. We asked applied linguists for their views on how academic language proficiency could best be tested, and we also studied the literature relating to students' linguistic problems and study needs. The results of these surveys are described in Westaway, Alderson and Clapham (1990), Alderson and Clapham (1992, 1993), and Foulkes and Hargreaves (1988). The feedback we received on the number and type of subject modules to be included in the revised ELTS is of direct relevance to this study, and will be briefly reported here.

Respondents' ideas on whether IELTS should have a modular component were very mixed and ranged from demands for an increase in the number of subject modules to suggestions that the whole test should be one of general proficiency with no modular component.

Many teachers and testers felt, as indeed the ELTS Validation Study had reported, that once candidates knew that they could take a subject module which was related to their own field of study, their expectations would be raised and they would expect to have a test exactly suited to their own chosen subject area. Since in ELTS there were only five subject areas plus General Academic many of the students were inevitably disappointed:

> *Brendan Carroll in his Specifications for an English Language Testing Service (1981) identified 20 important student categories or fields of study (Page 107). However, whilst this list was fairly detailed in some subject areas it failed to break down other subject areas such as engineering and sciences in sufficient detail. Within our own faculty of engineering, for example, there are departments of aeronautics and astronautics (presumably covered by aviation in Brendan Carroll's list), civil engineering (running postgraduate courses in structural engineering, transportation planning and engineering, irrigation engineering and soil conservation), cryogenics, electrical engineering, electronics, computer science and information engineering, engineering materials, mechanical engineering, ship science and sound and vibration research (running postgraduate courses in audiology, sound and vibration studies and automotive engine and vehicle design technology). If*

one were to design subject-specific modules for each of these disciplines it would clearly be a very large undertaking. On the other hand, it seems unfair on an electronics engineer to expect him to answer questions on, say, casting, which presumably a student of engineering materials would be able to answer with no difficulty at all.
(Blue 1987)

The evidence of the Test Report Forms supported this. The academic subjects on which candidates were embarking ranged from aquaculture to digital communication to periodontology, and within some of the disciplines there are many branches. For example, as Blue might have expected, 34 different branches of engineering were listed, as well as five each of accountancy, agriculture and economics (Alderson and Clapham 1993: Appendix 12).

Such a variation within a subject area inevitably meant that one subject module was by no means specific to or suitable for all the students doing that subject. Most of the engineering candidates, for example, chose the ELTS Technology Module, but some chose General Academic, some Life Sciences and some Physical Sciences. This variation in the choice of module was reflected in other subject areas: of seventeen accountancy students, nine took Social Studies and eight General Academic; of eleven students studying agronomy, nine took Life Sciences, one took General Academic and one Technology.

A test with five subject modules, therefore, could not be described as subject specific. These modules, however carefully devised, could not cover the range of the students' specialisations, and even if there were twelve or twenty-four modules this would still be the case. In addition, as we have seen in Chapter 1, there is as yet no body of evidence to support the ESP testing claims that different disciplines consistently demand different linguistic skills, and that students are disadvantaged if they take a test which is not in their subject area.

In his concluding comments on the construction of TEEP, Weir says:

In our investigations of the language events and activities overseas students have to deal with in British academic environments and the difficulties they encounter therein, we discovered much that was common between students of different disciplines and at different levels. This did not remove the possibility though that the subject content of texts employed in our test tasks might unduly affect performance. Whilst we attempted to take account of this in our sampling, we were unable to produce any conclusive evidence that students were disadvantaged by taking tests in which they had to deal with texts other than those from their own subject area. The case for a variety of ESP tests therefore remains unproven.
(Weir 1983:549)

Because of the shortage of concrete evidence, and since it would in any case be impossible to satisfy all students, however many modules there were, some of the language testers and teachers who responded to the questionnaires felt that there should be no subject specific modules at all. However, almost all participants in the ensuing meetings felt that one of the attractions of ELTS was the choice of subject modules. In addition, the receiving institutions, in their questionnaire, indicated that they were very much in favour of them.

Alderson and Urquhart's (1985b) findings do offer some support for an ESP position. They found that the ELTS scores of students in certain subject areas tended to cluster together, e.g. business studies and humanities, and engineering and science. The ELTS Validation Study researchers, too, found some support for ESP in the results of a test/retest study that they conducted. They reported on the differing correlations which they found between different subject areas, and said:

> *But perhaps our most important finding here is that modules behave differently, even very differently, from one another. We do not know whether this means that the modules are different by accident and should not be, or whether they are different by design, i.e. that they properly reflect the different nature of the subject areas they represent. What is interesting here is that there is no uniform set of results for modules. They do behave differently, and this is, indeed, a justification of ELTS as a test of ESP. One interpretation of the ELTS construct would be that it means that everyone (all modules) is different in the same way; another interpretation is that everyone is different but in different ways. It begins to look as though the second of these interpretations is more likely to be correct.*
> (Criper and Davies 1988:111)

Because there was some evidence that candidates could be disadvantaged if they took a test which was too far removed from their own discipline, and since the majority view of those consulted was that ELTS should not be changed more than necessary, it was decided that the new version would still have a modular component.

The choice of how many subject modules there should be, however, was a very difficult one and inevitably depended on which subject areas these modules were designed to cover.

Surprisingly, only a fifth of the test administrators felt there should be fewer subject modules, but a representative from the British Council's Technical Co-operation Training Department said that:

> *Experience overseas suggested the test was cumbersome to administer, too time consuming, and probably unfair to candidates*

> *since the selection of modules according to subject of study was fairly arbitrary, and certainly difficult to make.*
> (ELTSREV 1987a)

Although most of the receiving institutions liked the idea of six modules, they thought that a three-module test would be adequate. Some of them suggested a conflation of the present modules so that the new ones would be:

> *1) General Academic + Social Studies*
> *2) Physical Sciences + Technology*
> *3) Life Sciences + Medicine.*

Most of the EAP teachers who filled in a questionnaire about the revision of ELTS, did not agree that there should be fewer subject modules and several of them felt there should be more. However, when these teachers later met to discuss the test's future in more detail, they came to appreciate the fact that the more subject specific a test became, the higher the candidates' expectations and ensuing disappointment might be, and they also became aware that a great increase in the number of subject modules would pose practical problems in the generation of parallel tests. By the end of the meeting there was some agreement that the number of modules should be reduced rather than expanded (ELTSREV 1987b).

The language testers felt, on the whole, that since it was impossible to have enough subject modules to suit all disciplines, it would be better to have just two or three broad categories of specialisation. Some testers wanted two modules, Science and Non-science, and some wanted a third, Arts and Humanities, because the requirements of courses in this area, with their strong focus on critical writing and subjective commentary, were very different from those of the social scientists (ELTSREV 1987c).

However, a search through the 1,000 Test Report Forms revealed that only nineteen students intended to take subjects that fell within the scope of Arts and Humanities. Of these nineteen, more than half were heading for courses in English, linguistics, and modern languages. Such students were not the ones for whom ELTS was specifically designed as their English was in almost all cases very good. It seemed, therefore, that although the needs of students in Arts and Humanities might well be different from those in other disciplines, the number of candidates taking that ELTS module would be so small that it would not be worth the time and money spent on the construction.

The question, therefore, was whether there should simply be a two-modular structure divided into Science, and Arts and Social Science. The problem here was that the Science module would be expected to cover such a wide range of subjects that the selection of texts might pose a real problem for the test writers. To find something equally familiar to both, say, construction engineers and veterinary scientists might be so difficult that the test writers would have to choose something that was so bland that it would be equally suitable for social scientists.

A count of the subject areas represented in the Test Report Forms revealed that the candidates were roughly divided into thirds, one third intending to take subjects in Business Studies and Social Science, one third Physical Science and Technology, and the remaining third taking Life and Medical Sciences. These three areas are the three recommended by those receiving institutions which suggested a conflation of the existing six modules. In the absence of any strong evidence from research as to the ways subject areas cluster, such a conflation seemed to provide a practical solution. and it was decided, therefore, that the revised battery would consist of three subject modules:

1) Business Studies and Social Science
2) Physical Science and Technology
3) Life and Medical Sciences.

There was some feeling from ELTS users that the battery should also provide different tests for under- and post-graduates as undergraduates did not in all cases belong to a single discipline. However, since the revised ELTS would no longer be divided into many subject specialities, we felt that undergraduates should no longer find it difficult to choose a module. In addition, no research findings had yet shown that undergraduates' needs were radically different from those of postgraduates. We therefore decided that undergraduates and postgraduates would take the same test battery.

The Structure of IELTS

Table 4.2 shows our plans for the structure of the revised test battery, IELTS, which was now jointly sponsored by the British Council, UCLES and the International Development Program of Australian Universities and Colleges (IDP). This overall design remained largely unchanged up to the test's launch in 1989, but the Grammar component was dropped after the main trials (see Alderson 1993a).

Table 4.2

Proposed Structure of the Academic Version of IELTS

General Component

Test	**Time**	**Marking**
Grammar	45 minutes	Clerical
Listening	30 minute tape	Clerical
Speaking	11-15 minutes interview	Trained rater

All candidates to take the same three tests, regardless of their future course of study.

Modular Component

Test	**Time**	**Marking**
Reading	55 minutes	Clerical
Writing	45 minutes	Trained rater

Three Reading and Writing modules to be based on three broad academic subject areas: Business and Social Science (BSS), Life and Medical Science (LMS) and Physical Science and Technology (PST). Candidates to take the module closest to their future field of study. Reading and Writing to be integrated so that the writing component would depend to some extent on passages used in the Reading test. The main disciplines to be covered in these three modules are listed in Appendix 4.1.
Scores to be reported on a 1-9 band scale, with a profile score for each of the five subtests. An overall band score would give the mean band score of the subtests.

Designing the Draft Reading Modules

There is now a body of knowledge concerning students' academic needs upon which test constructors can draw. Pre-sessional teachers have amassed data and experience from their classes, and there is also relevant empirical research into students needs and problems (e.g. Bridgeman and Carlson 1983; Geoghegan 1983; Powers 1985; Weir 1983). However, the transformation of this knowledge first into specifications and then into test items poses problems for the test designer, especially when the needs of different disciplines have to be collapsed into, say, two or three subject areas. There is a danger that by the time the test is written it will have only tenuous links with the original content analysis and specifications. In the case of TEEP, for example, Cyril Weir spent two years conducting a detailed needs analysis of the requirements of tertiary level students (Weir 1983). He toured tertiary education colleges finding out what the students' needs were from both the teachers' and the students' viewpoints. This data is and will continue to be invaluable for years to come, but, according to Alderson (1988c), when Weir tried to make use of it for writing items he found he could not. The information he had was so diverse, and there was such variation even within disciplines or departments, that he was unable to make useful generalisations which could be turned into test items.

For IELTS, it was decided that no new needs analysis should be carried out but that once test writers had designed draft specifications and tests these would be distributed to subject specialists as well as to teachers and testers for detailed comment. In this way, the new tests would undergo content validation after, rather than before or during, test construction (Alderson 1988c).

Draft Specifications and Tests

In the latter part of 1987, three teams of item writers, each with three members, of whom at least one was an EAP teacher, and one a tester, met to discuss the content and format of the draft academic modules. Since no firm idea of a test construct had emerged during the consultative process (see Alderson and Clapham 1992), the teams were asked to devise their own, referring to needs analyses and other recent testing research as they wished. They were encouraged to be innovative, but had to keep within the overall design constraints (i.e. the test battery had to be shorter than ELTS and simpler to administer, all four language skills had to be tested, and the tests of listening and reading had to be clerically markable).

Two of the draft sets of specifications – Life and Medical Sciences (LMS) and Physical Science and Technology (PST) – turned out to be very similar in their approach. Both prescribed the number of texts to be used (PST said one, two or three and LMS said four) and both described the possible sources of such texts and recommended a variety of reading tasks. The PST texts were based on potential or actual engineering disasters, and the LMS ones were on food safety, the AIDS virus, and the effect of diet on cancer. Both sets of specifications also described what were seen as the main reading activities of students in the relevant subject area. For example, LMS listed four reading 'purposes':

1 reading for specific information
2 reading for salient points
3 reading to evaluate (a) evidence and (b) writer's thesis
4 reading to identify topic or theme

and PST listed two 'macro-purposes':

1 to acquire information
2 to increase understanding

with a set of micro-purposes:

1 to extract main ideas
2 to scan for information
3 to review and interpret evidence in order to take action, reach a conclusion, form a hypothesis

4 to understand how to do something
5 to understand how something works
6 to find out about other people's experiences
7 to find out about a sequence of events
8 to confirm or reinforce one's understanding or memory.

Both draft tests included a wide range of test types including gapped summary, information transfer, diagram completion, multiple-choice and open-ended questions.

The BSS draft specifications were very different, and took the form of a rationale for the test's content rather than a prescription for test writers. The rationale proposed that the tasks in the test should be as similar as possible to those that students would actually perform during their studies. Candidates were, therefore, to be presented with a large amount of reading – six passages with a total of approximately 9,000 words. A few small reading tasks were to be included in the draft test, but the major part of the candidate's assessment was to be based on an essay drawing on material from the reading passages. To make the essay task authentic, one of the passages was to be irrelevant, so that students would have to show that they could select suitable texts. (For more about the BSS draft specifications see Coleman 1988.)

Content Validation

In mid January 1988 we sent out the draft specifications and tests, accompanied by questionnaires, to subject specialists, applied linguists, language testers and teachers in Australia, Britain and Canada. In addition, we asked three 'moderators' to comment on the proposed academic modules. We collated and summarised the ensuing reports and questionnaire responses, and our findings led to a radical revision of all the specifications. These findings are described below.

Subject Specialists

Each of the three academic modules and specifications was read by twenty or more subject specialists in the relevant subject areas. In most cases respondents did not complete the questionnaire themselves, but were interviewed by members of the ELTS revision project team or by the authors of the draft specifications, so that potentially interesting answers could be followed up in detail. (See Appendix 4.2 for a copy of the questionnaire.)

It has to be said that in all three subject areas the interviews varied in value. Some lecturers were interested in the project and willingly volunteered an hour of their time to offer carefully thought-out answers and suggestions. Others, though, did not appear to have read the material until just before the interview and then gave only cursory replies. Study of the questionnaires has revealed, too, that some lecturers commented adversely on item types if they disliked the passage on which the questions were based. Their answers to questions about reading tasks, therefore, were coloured by their views on the texts. However, on the whole the interviews were useful and many valuable suggestions were made by the respondents.

Life and Medical Sciences

Twenty-two lecturers responded to the LMS questionnaire. They came from many different universities in Australia and Britain and between them they covered a wide range of disciplines, as will be seen from the attached list of subject areas (see Appendix 4.3). In some disciplines there was more than one respondent. For example, there were five agriculturalists, so that it was possible to check whether there was any unanimity among lecturers in the same subject area but in different universities.

On the whole the LMS subject specialists approved of the draft specifications and draft test paper. Eleven out of seventeen felt that the reading passages, which came from academic journals and were mostly on the subject of irradiation of food, were appropriate for their undergraduates, and seventeen out of twenty that they were suitable for postgraduates. However, some did not like the fourth passage, which was on the AIDS virus, partly because of the topic itself, and partly because the associated exercise, labelling a diagram, was difficult. The other reading tasks consisted of a gapped summary, matching headings to text, multiple-choice questions, and labelling a flow chart. These tasks were generally felt to be suitable. Two respondents liked neither the reading texts nor the tasks, and also thought their students would be disadvantaged by them, but these two were both psychologists and it quickly became evident that the LMS module was not appropriate for them; BSS would have been more suitable.

The five agriculturalists were almost unanimous in their answers. One out of the five felt that the texts were not suitable for undergraduates because there were too few tables, and another said the texts were inappropriate for postgraduates because the topics were wrong. With these exceptions all the agriculturalists were satisfied with the specifications, texts and tasks.

There was only marginally less agreement when all the life scientists were grouped together. One biologist liked the texts, whereas the other biologist did not, and the forestry respondent wanted texts related to land use, but otherwise their answers were similar to those of the agriculturalists. The medical scientists too, showed no sharp differences from the rest. One veterinary scientist, however, felt that newly arrived students might not have seen a report on current research before, and there is some evidence from individual comments that medical scientists expect their undergraduates to start reading articles from journals later than do life scientists. However, there is not enough data to confirm this.

In addition to their direct answers to the survey questions, the life and medical scientists made many extra suggestions and comments. Most of these related to the writing tasks but they did suggest two for reading: 'complete a list of authorities pro and con an idea from the text' and 'select information relevant to a particular issue'.

Physical Science and Technology

Thirty-three subject specialists answered the PST questionnaire. Their responses were more varied and revealed wide disagreement between lecturers in different groups of subject areas. There seemed, for example, to be a split between chemistry, physics, mathematics and computer science on the one side and engineering on the other. However, this split was not clear cut – the two mechanical engineers often did not agree with each other, and frequently did not agree with the civil engineers. It is possible that the great divergence among the respondents was due, at least in part, to the choice of texts for this module. There were two problems with these. Firstly, they were both in a narrative style, and came from general magazines of the *New Scientist* or *Scientific American* genres. Although some lecturers thought they were appropriate for their undergraduates, some felt they were too journalistic, and that they were too short on logical argument, description, symbols and tables. Secondly, both texts were reports on potential or actual engineering disasters, and, although satisfying most of the engineers, satisfied very few others. It seems that the texts did not fully match the specifications, which required the texts to contain 'different kinds of writing' and gave as examples 'review, description and discussion'. Most of the respondents felt that the specifications themselves did take account of what was known about reading in their disciplines, and more than half said that the criteria for selecting texts were specified in such a way that the most appropriate texts would be chosen. However, many respondents felt that in spite of this the most appropriate texts were not chosen. Indeed three respondents specifically said that the criteria were satisfactory but that the texts were not.

Whether it was indeed the particular choice of texts that led to the disagreement is not clear, but it is certain that PST is a difficult area to cater for. Seven of the lecturers in physics, chemistry, mathematics and computer science said or implied that their students would read less wordy texts, and the physics and mathematics lecturers said that their texts were mainly specialised mathematical ones with a preponderance of symbols and mathematical formulae. Indeed the five lecturers who thought the specifications were inadequate all came from these four disciplines. Since there are such different textual requirements for PST, it is not certain whether the problem can be solved, even if the texts are more varied, and include more logical arguments, symbols and tables.

Fortunately, most of the respondents did not think that their students would actually be disadvantaged if they were presented with the PST texts. Only three thought their undergraduates would be disadvantaged, and four thought their postgraduates might be.

Business Studies and Social Science

The specifications for BSS, as was described earlier, were markedly different from the other two sets, and academic purposes were not listed in any way. Of the twenty-seven BSS subject specialists who responded to the questionnaire, fifteen felt the specifications did take account of what was known of reading and writing in their area, but two suggested that purposes should be listed as in the other specifications, and several listed purposes of their own. For example, a lawyer suggested that 'reading narrative' should be listed as a purpose and a political scientist suggested 'critically evaluating opposing theories'. More than a third felt that the criteria for selecting texts were inadequate, and one, in public administration, recommended that texts should include tables and diagrams.

When it came to the choice of texts there were some markedly opposing views. The religious studies lecturer, for example, said there were too many tables, and that the text should be completely linear, whereas a business analyst wanted the text to be broken up into headings, boxes, tables and diagrams. The test contained six thematically linked reading passages, which ranged from newspaper articles to extracts from academic papers. The specifications required that these should be thematically linked rather than representative of a wide range of disciplines. The group that was most satisfied with the texts comprised the social scientists – politics, economics, psychology and education. They all felt the texts were suitable for their undergraduates, and almost all that they were right for their postgraduates. The business studies group, however, was less satisfied: three out of seven felt the texts were too difficult and two felt there were not enough graphs and tables.

Unfortunately this dissatisfaction with the texts was so strong that almost half the lecturers thought that their undergraduates would be disadvantaged by them, and a quarter that their postgraduates would. Some were worried by the style and content of the texts, some thought the passages were too long, and some just said that they were too hard.

The reading tasks consisted of open-ended comprehension questions. These were generally considered to be appropriate for the students, although a few lecturers stressed that their students were expected to write, not to do reading exercises. This supports the approach taken by the BSS team, who felt that their test would be more authentic if the students were given no reading tasks, but were only asked to produce a piece of writing based on the reading passages. However, this is of little help for testers who need to be able to assess reading separately from writing in order to know whether, regardless of their writing skills, students are capable of comprehending academic texts.

Summary of the Subject Specialists' Views

Of the three academic modules, it was the LMS module which was considered most appropriate by its respondents. There did not seem to be the split in the requirements of different subject areas which was so obvious in the other two.

In PST and BSS it seemed to be the texts that caused the problems, and it may well be that if PST had included a non-engineering text, and if BSS had had shorter texts in a wider range of disciplines, the problems would have been reduced.

Of all the sections of the questionnaire it was the questions on the reading passages which proved the most useful. The subject specialists clearly found it easy to compare the sample texts with those they gave their own students, and the questionnaire answers were to the point and enlightening. The least helpful comments were on the tasks. Since these tasks had to be clerically markable, they were, inevitably, not in a form that most lecturers would use with their students. Although the majority of the lecturers said they were appropriate, and one or two wished their students were given such tasks, many did not appreciate that their purpose was to assess reading. When they were asked, therefore, what other reading tasks their students were given, some said 'write essays', 'write summaries' or even 'take notes at lectures'. The responses to these questions, therefore, although confirming that on the whole the tasks were suitable for the students, did not produce useful suggestions for other tasks.

Language Teachers, Testers and Applied Linguists

In addition to seeking the views of subject specialists, we also sent questionnaires to language teachers, testers and applied linguists in Britain and Australia asking for comments on all the draft specifications and tests, general and modular. Thirty answered the questionnaire, but since some of the respondents were not concerned with English for Academic Purposes, only nine answered the questions relating to the subject modules. (See Appendix 4.4. for a copy of the questionnaire.)

Because there were so few responses it is difficult to generalise in any way about them. I have therefore attempted to report the flavour of the answers, without giving a detailed breakdown.

Physical Science and Technology

Most of the respondents felt that the PST specifications did take account of what was known about reading in this subject area, and that the 'purposes' were sufficiently detailed to cover the kinds of reading done by students in this area. One respondent, however, suggested another four purposes which were not included in the draft specification, and these four were later incorporated in the final specifications. Again, most of the respondents said that the criteria for selecting texts were adequately specified, but one said that passages from textbooks would be more appropriate, and one commented on the fact that there was no guidance on the required complexity of the texts. Most thought that the texts chosen were appropriate for their intended readers, but three respondents said that the texts were too general, and two said that the language used was too journalistic. Only one of the nine thought that students in this broad subject area would be disadvantaged if they took a test containing the draft reading passages.

The respondents were quite unable to rank the texts according to their suitability for the students. Some refused to try, and those who did all produced different rankings.

Comments on the test tasks varied, but the general impression was that there was a sufficiently wide range of task types, although one respondent suggested that one of the tasks should consist of a flow chart completion, and another suggested that candidates should be asked to identify reasons, definitions, methods, results, and to distinguish between statements of evidence and interpretation.

Life and Medical Sciences

On the whole the comments on the LMS module were very similar to those relating to PST, but one respondent said that, unlike the PST specifications, the LMS ones were not specific enough. All the respondents considered that the reading passages were appropriate, although one was not sure whether they would be suitable for undergraduates, and wondered whether non-medical students would be at a disadvantage. All the respondents also felt that the tasks were suitable for the candidates, and one thought that the tasks were exactly those which the students would have to do in their academic studies.

Business Studies and Social Science

Although most of the respondents said that the language used in the BSS module was suitable for this subject area, they were almost uniformly concerned about other aspects of the reading passages. Two were worried that there was too much reading material, and one felt that the texts were so difficult that few native speakers would manage the test. In what way they were difficult was not specified. One felt that the texts would only be appropriate once students had embarked on their course and had learnt the appropriate study skills, and two felt that the texts were too heavily oriented towards Great Britain. All respondents said that the students in this subject area *would* be disadvantaged if they took a test containing these reading passages.

Moderators

We also sent the draft specifications and tests to three moderators. These were university applied linguists who were experienced at teaching and testing English for Academic Purposes. They were not sent a questionnaire, but were asked to write detailed reports on the three subject modules.

Unlike the language teachers and testers, the three moderators all liked the BSS team's approach. One of them thought that the test design proposals were by far the most interesting of the three modules, another liked the overall model better than those of the other two modules, and the third appreciated the team's attempt to ask candidates only to carry out tasks which would be appropriate for students in an academic setting. This moderator felt that of the three modules this contained the most authentic test of academic reading and writing. However, although the three moderators liked the team's arguments and innovatory ideas, they felt that the execution of these ideas was inadequate. One of them said that the specifications lacked rigour, and that the text characteristics were not sufficiently specific. All three felt that there was too much reading material, and that the inclusion of an irrelevant reading passage would confuse the candidates. None of the moderators were satisfied with the somewhat sketchy reading tasks.

On the whole the PST specifications and draft test were liked, although one moderator felt that the reading tasks did not mirror the good ideas contained in the specifications. Two of the moderators questioned the use of a gapped summary as a test of reading comprehension, and two found some of the reading tasks over complicated. There were some criticisms of the macro- and micro-purposes, one reader saying that they were too vague, and another that they were somewhat repetitive. There was also some concern about the clause in the specifications which allowed the three reading passages all to come from the same publication. It was felt that if all the texts came from one article there might be too little variation in the text types, and it would be difficult to avoid favouring students in one academic discipline.

The comments on the LMS drafts were similar to those of PST, although there was considered to be a greater load on reading in this module. Two moderators thought that the texts were too difficult and too closely related to the medical sciences, with no suitable texts for students in disciplines such as botany and forestry. One moderator felt that three of the test 'purposes' were indistinguishable, and in any case irrelevant since on this occasion all the students would have one purpose only in reading the texts – that of showing they understood the texts well enough to pass the test. One moderator liked the gapped summary task which was more easily marked than the ones in PST and BSS.

Overall Summary

The general feeling given by all our advisers (that is, the subject specialists, the teachers and testers, and the moderators) was that on the whole the PST specifications were closest to those required for the IELTS academic modules. Most people liked the idea of the macro- and micro-purposes even if they were not satisfied with the individual descriptions, and the texts and tasks were considered generally suitable even if there was an unacceptable mismatch between the specifications and the reading passages. The subject specialists, however, felt that the texts were biased towards engineers, and they made it clear that at least one passage ought to be more suitable for physicists and chemists. LMS was generally deemed appropriate by all the respondents (the consistently adverse comments from the psychologists helping to define the boundaries of this subject area) but some respondents would have liked macro- and micro-purposes listed for reading, instead of the more general 'purposes'. The BSS specifications worried the subject specialists, teachers and testers; the moderators, although intrigued by them, did not feel that they converted well into texts and tasks.

Redrafting the Specifications

Once we had analysed the results of the content validation exercise, we revised the specifications to take as much account as possible of the advisers' views. We had to reject the BSS rationale partly because of the difficulty in implementing the specifications and partly because one of the unalterable IELTS test constraints was that reading and writing had to be given equal prominence. We used the PST specifications as a foundation for the new ones because they seemed to be the closest to what was needed, and because they were the most clearly presented. The macro- and micro-purposes were converted into 'academic tasks' such as 'identifying structure, content, sequence of events and procedures' and were modified according to subject specialists' comments, and to the findings of needs analyses such as Weir's (1983). What was interesting was that as the revised specifications took shape, they gradually became more and more similar to each other. For example, as an academic skill was added to one set of specifications it became clear that it was also a required skill in the other two subject areas. 'Identifying the underlying theme or concept', for example, is required in all three subject areas, and so is 'identifying, distinguishing and comparing facts, evidence, opinions, implications, definitions and hypotheses'. Eventually the final list of academic tasks was identical for all three subject areas.

The specifications were changed gradually over time, and some of the final changes were not made until the results of the pilot and trial tests had been analysed. By that time the only sections in which the specifications differed from each other were those describing the target audience for each of the modules, and those specifying the sources and types of reading passage. Even here the differences were minimal. In all three modules the specifications now required item writers to select at least one text from each of the two main subject areas, so that, for example, in PST there had to be one text relating to the physical and one to the technical sciences. In BSS one text had to relate to business studies and one to the social sciences. The only difference in the specification of text content between BSS and the two science modules was that for BSS no mathematical formulae were required.

Draft versions of the revised specifications were sent to the item writing teams to guide them when they revised their draft tests. Extracts from the LMS test draft specifications as they were in December 1989, at the end of the ELTS Revision Project, are presented in Appendix 4.5. The PST and LMS modules were extensively revised, and BSS was completely redesigned. By September 1987 the three modules were ready to be piloted.

IELTS Trials

To round off the description of the steps that were taken to make the IELTS reading modules as appropriate for their candidates as possible, here is a brief description of the pilot and main trials. Detailed reports of these trials are described in Clapham and Alderson (forthcoming).

The IELTS Pilot and Main Trials

The pilot reading modules were given to 152 BSS, 80 LMS and 228 PST students, in Algeria, Australia and the UK. All the students had English as a second language. The aim of these trials was to identify major problems with the tests, contents, procedures, time allocations, instructions and layout so that these could be corrected before the major trials took place. As far as possible each student took the reading module that was appropriate for his or her field of study, and each student filled in a questionnaire giving background details and commenting on the suitability of the tests.

The questionnaire results were tabulated, the tests' descriptive statistics were calculated, and classical and Rasch item analyses were used to identify weak items. The three tests were then revised. LMS and PST needed few alterations, but BSS, which was only in its first draft, needed more radical changes. The number of reading passages was reduced from four to three, and the number of test types from seven to five. All the reading passages were related to education, and it was not until a new version of the test was in the process of being written that the specifications were altered to require the inclusion of a passage on business studies.

One encouraging finding was that there was no evidence that students and administrators had any difficulty matching students' disciplines to the appropriate reading modules.

After the tests had been revised, major trials using ESL students were run by Australia and the UK. 1,146 students took BSS, 633 LMS and 779 PST. The Australian trials took place in Hong Kong, Indonesia, Thailand and Australia (see Griffin 1990), and the British ones in Bangladesh, Cyprus, Egypt, France, Germany, Hong Kong, Hungary, Jordan, Malaysia, Rwanda, and Great Britain. For a comparison of the Australian and the British results, see Alderson (1993a). Some of the results of these trials will be described in Chapter 5. After these trials the tests were given a few final alterations before they were launched as the first 'live' IELTS reading modules in October 1989.

Native Speakers

In the above trials all the students were speakers of English as a second language. There was some disagreement among members of the IELTS Project Steering Committee as to the value of trying out academic proficiency tests on native speakers. Some members argued that it was unreasonable to expect non-native speakers to succeed at a test if native speakers could not answer the questions easily, and that the IELTS tests should therefore be tried out on native speakers of English. Others said that there would be no point in doing so because so many factors apart from language would affect their scores that the native speakers would not necessarily achieve high marks. This issue remains unresolved in language testing: although in the era of discrete point testing, native speakers were expected to achieve scores of over 90% (see Hamilton *et al.* 1993), such scores are not considered to be so easy to achieve on communicative, academic tasks.

In spite of these reservations, the trial versions of the reading modules were given to a small group of English L1 students who were at the end of their first year of a two-year course leading to English A Levels. The students took the academic modules most appropriate for their subject areas. The results are shown in Table 4.3.

Table 4.3

Native Speakers' Trial – Descriptive Statistics

	BSS	LMS	PST
Number of Students	29	16	10
Number of Items	38	39	33
Mean Raw Score	27.52	32.25	27.50
Raw Standard Deviation	3.42	4.39	3.5
Mean Score as %	72.41%	82.69	83.33
Percentage Score S.D.	9.00	11.25	10.62
Mode	76.30%	84.60%	84.80 and 87.90
Median	73.70%	84.60%	86.40%
Max/min	86.80%/55.30%	94.90%/56.40%	93.90%/57.60%
Mean Discrimination*	.20	.28	.27
Reliability (KR20)	.53	.79	.72

* $E_{1\text{-}3}$

As can be seen, these statistics are based on only a small numbers of students, but even so the results are interesting. For example, the range of scores is surprisingly high (see the max/min row in the table). In all three groups the distribution of scores is negatively skewed, with most students clustering at the top end as expected, but unexpectedly in LMS and PST the few lowest scoring students are well spread out. In BSS the skew is not so marked, but in the LMS and PST groups, about 75% of the students achieved scores of 84% or over. This distribution accounts for the mean discrimination indices which are high for a native speakers' trial. The fact that there was such a difference between the majority of students who all clustered at the top of the distribution and the low scoring students at the other end, meant that some items inevitably had high discrimination indices. (Of course it must be remembered that since the sample sizes are so small, especially for the PST trial, one low scoring student will have a strong influence on the results.)

Without knowing more about the students who took these tests, it is not clear how one should interpret these results. It might be that the lower scoring students had difficulties with either the language or the subject matter. The fact that these students were still at least a year away from their college careers, and were therefore below the academic level of the target audience for these tests, may have been one of the causes of the low scores. One would expect those same students to have higher scores if they took the test a year later; not only would their subject specific background knowledge have increased as they advanced in their studies, but their general background knowledge and experience of study skills should also have improved.

Predictive Validity

Finally, one important aspect of these reading modules which has not been discussed is that of the tests' predictive validity. Are the tests succeeding in distinguishing between those students who will be able to cope with their academic reading in English, and those who will not? A thorough predictive validation study would involve gathering measures of post IELTS students' reading ability later in the year to see whether these agreed with earlier IELTS scores. One of the difficulties of such a survey would be that the student sample would need to include IELTS successes and failures. The successes could be evaluated within their new institutions, but it would be difficult to follow up the failures (see Criper and Davies 1988). I shall not discuss the problems of predictive validation here, but see the comments on the ELTS Validation Study made by Pollitt (1988) and Clapham and Hughes (1988).

Since the major IELTS trials were not run until early in 1989, and since the IELTS Project finished later that year, we were not able to set up a full-scale predictive validation study of the IELTS trial tests. However, in order to get some idea of the new tests' predictive validity, we ran a small study on the pilot tests. In September/October of 1989, approximately a year after the pilot tests had been administered, questionnaires were sent out to the subject tutors of students who had taken the pilot versions of IELTS, asking them to report on their students' language proficiency and academic success. The questionnaire was based on one used in the ELTS Validation Study (Criper and Davies 1988). As it was difficult to discover the academic destinations of some of the students, only 56 questionnaires reporting on academic students were completed, and of these only 29 related to BSS students, 11 to LMS and 16 to PST. The numbers are therefore very small and the results of the study are inconclusive. An attempt to replicate the ELTS Validation Project's comparison of ELTS scores with students' final college success or failure (see Criper and Davies 1988) was not successful because 35 of the 56 students had not yet completed their courses. However, one of the questionnaire items did produce results which it is worth reporting. This question asked tutors to rate, on a scale of 1 – 9, their students' general language ability at the end of the year. Table 4.4 gives the Pearson Product Moment correlations of these ratings and the students' raw scores on the academic reading modules.

Table 4.4

Predictive Validity Correlations between Scores and Ratings

	General English
BSS (n = 29)	.53*
LMS (n = 11)	.52
PST (n = 16)	-.07

*= significant at .01

In a predictive validation study of this sort the correlations would not be expected to be high, since a year had elapsed between the test and the time when the tutors assessed their students, and since such ratings by untrained tutors are likely to be unreliable (see Pollitt 1988 and Clapham and Hughes 1988). In addition, the tutors were rating the students on their general language ability rather than on their reading proficiency. In these circumstances the BSS correlation with the General English rating of .53 is surprisingly high. It would be ironic if the test which was most changed after the pilot trials proved to have been the module with the highest predictive validity. It is interesting that the LMS correlation, although not significant because of the small sample size, was similar. However, this similarity could well be due to chance.

Construct of the IELTS Reading Modules

Now that the stages in the metamorphosis from ELTS to IELTS have been reported, it is appropriate to describe the construct underlying the IELTS academic reading modules. We had asked applied linguists for advice on current theories of language proficiency on which we might base the IELTS test battery. However, the applied linguists' responses were varied, contradictory and inconclusive, and provided no evidence for a construct for EAP tests on which we could base the test:

> *We were obliged to take an eclectic approach to the establishment of specifications for our test writers. This meant that we selected those aspects of the different responses which we judged to be practicable, to fit our brief (in particular, to maintain a degree of continuity with the existing test) and to correspond with our other sources of information and opinion. The result is very far from being a theoretically pure model of language proficiency, and perhaps the most that we can claim for our underlying construct is that it does not appear to contradict or conflict in any serious way with what theorists and empirical research have revealed as to the nature of language proficiency.*
> (Alderson and Clapham 1992:164)

The theoretical foundations on which the IELTS academic reading modules are based are listed below.

1 Although it is accepted that academic students will give the best evidence of their academic reading ability if they are given reading tests in their own field of study, it is administratively impossible to give each student a test directly related to his or her discipline, and since the ELTS choice of five subject areas led to many cases of students either taking an inappropriate module, or being disadvantaged by inappropriate subject material in their academic module, IELTS has only three broad subject areas, BSS, LMS and PST. These capture, it is hoped, the main differences between the three subject areas, without leading students to expect content closely related to their own branch of an academic discipline. The theory underlying this test is therefore that academic students will give the most accurate evidence of their academic reading ability if they are given reading tests in their own broad subject area. IELTS is an ESP test in so far as it is specifically designed for students proposing to undergo academic study. It is an ESAP test in that it provides tests in three broad subject areas. (See Chapter 1 for an explanation of ESAP.)

2 The IELTS reading passages are intended to be authentic texts for students in the relevant academic disciplines. They must come from authentic sources, but they can be modified to remove ambiguities or grammatical errors.

> *Texts must be realistic and in modern English and must appear to be authentic, even if the original texts have been modified.*
> (Appendix 4.5: Draft Specifications)

Since each module covers two broad subject areas, each module is intended to contain at least one text from each of these areas. For example, LMS must have at least one text from the Life and one from the Medical Sciences.

3 The tasks are intended to be as authentic as possible, given the requirement

for clerical marking. Within any one module there should be a variety of item types, but as there do not seem to be any major differences between tasks in the three broad subject areas, the same types of items are suitable for all three modules.

4 The reading tests are intended to sample the students' ability to perform a string of tasks for academic purposes, for example:

(i) identifying structure, content, sequence of events and procedures
(ii) following instructions
(iii) finding main ideas which the writer has attempted to make salient.

Since it is difficult, however, if not impossible to know what a given item is testing (see Alderson, 1993b), no single item can be definitively described as testing one or more of these tasks.

It is not implied or assumed that these (tasks) can or must be tested in isolation or independently of each other.

The tasks are listed for the guidance of the item writers, and may not bear any direct relationship with the items themselves.

In Chapters 7, 8 and 9 of this book I shall be investigating these constructs in more detail.

Conclusion

The aim of this chapter has been to show our attempts to make the IELTS reading modules as suitable as possible for students in the three broad subject areas of BSS, LMS, and PST. This was a difficult task because each broad subject area covers two not perfectly compatible narrower ones, so that PST, for example, has to be suitable for both chemists and engineers although the texts required in their two disciplines are different both in subject matter and style. (I shall be discussing the different text types in Chapter 9.) However, this is only a problem when selecting text types and topics; from our content validation study it appears that the academic reading skills required are the same in all three areas, and the test types, although not on the whole typical of the sorts of tasks students would do, are equally appropriate for all three subject areas. One advantage of having three subject areas instead of the ELTS five is that few students are expected to have difficulty selecting the appropriate module. Students are less likely to be disadvantaged, therefore, by taking modules which are outside their subject area.

In Chapter 5, I describe my pilot study into the effect of students' subject area on test performance. This study is based on data collected during the trialling of these first IELTS reading modules.

5 The Pilot Study

As Chapter 1 showed, evidence concerning the effect of subject area on EAP test performance is mixed (see, for example, Alderson and Urquhart 1985b, and Koh 1985). It does not always seem to be the case that students achieve higher reading comprehension scores on tests that are based on familar subject matter. The first aim of this pilot study therefore was to see whether students taking the IELTS reading modules were at an advantage if they took the reading module in their own academic field of study, and to get some tentative ideas as to why this might or might not be so. The second was to try out the intended method of data analysis, and to get information about the amount of data and the number of students required for the main study.

Research Question

This study was based on one question: do students score significantly higher in a reading test within their subject area than in one outside it? This does not address the question of background knowledge per se, but addresses the question of whether reading tests in different subject areas should be included in university language proficiency tests.

The Tests

The tests used in the pilot study were the trial versions of the three academic reading modules which were described in Chapter 4. The BSS module contained three reading passages, all related to Education, and the item types included matching headings to paragraphs, a gapped summary, true/false questions and the identification of sentences in a text which contained the answers to a set of questions. The LMS module had four reading passages of which three were on the topic of food irradiation, and one was on the relationship between diet and cancer. The item types included a gapped summary, the identification of phrases which had been omitted from the reading passage, some multiple-choice questions, and a flow chart completion exercise. The PST module contained three reading passages, one on recycling resources, one on the possible collapse of a railway tunnel, and one on the Chernobyl disaster. The passages were accompanied by open-ended questions, a problem matching task, and a gapped summary. For the purposes of this study these trial versions will be referred to as T(BSS), T(LMS) and T(PST). Each student took two of the reading modules, one of which was in the appropriate academic subject area. Table 5.1 gives statistics from the complete trials of the revised versions.

Table 5.1

The Trial Tests – Descriptive Statistics

	T(BSS)	T(LMS)	T(PST)
N. of Students	1146	633	779
N. of Items	38	39	33
Mean as a Percentage	46.1	42.2	51.0
Percentage Scores S.D.	21.2	23.5	23.9
Reliability (KR 21)	.88	.91	.90

The Examinees

The examinees were non-native English speakers progressing to undergraduate or postgraduate studies at English medium universities. Some were already in Australia or Britain taking preparatory English classes, some were still in their home countries. They formed a heterogeneous sample with different linguistic and cultural backgrounds and different levels of language proficiency. Unfortunately, not all the 1,146 students who took the tests could be included in the study because some had failed to provide details of their fields of study. In addition the data set is heavily weighted towards Business Studies and Social Science students. Table 5.2 lists the first languages of those students who provided the data base.

Table 5.2

Students' First Languages

Language	N	Language	N
Chinese	30	Malay	13
French	79	Minangkabau	1
German	24	Portuguese	1
Indonesian	58	Spanish	1
Japanese	14	Sundanese	1
Javanese	7	Thai	10
Korean	13	Turkish	1

The students were classified into three groups according to which of three broad discipline areas they would be studying next. Their future, rather than their past, subject areas were chosen as it is a student's future field of study which dictates the choice of IELTS subject module. Universities are interested in information about how students will fare in their future university course, and therefore, in the case of IELTS, it is the student's ability to cope in the future subject area which is tested. The classification of students is not always easy, as subjects such as Computer Studies and Architecture border on different subject areas. For consistency's sake the classification was made according to the list supplied in the IELTS Administrators' Manual.

It can be seen from Table 5.2 that there was one homogeneous language group – the Indonesians – which was sufficiently large for separate analyses to be carried out on that group. (There was also a large group of French students, but too few of them took a combination of the LMS and PST modules for a balanced analysis of the results to be possible.) All but five of the Indonesians were attending English classes in Indonesia, before going on to postgraduate studies in Britain. The other five were attending language classes in Australia. Since the use of a heterogeneous group may well affect findings in unexpected ways, results from the Indonesian group are reported as well as those for the Whole Sample.

Table 5.3 gives the means and standard deviations for both the Whole Sample and the Indonesians.

Table 5.3

Distribution Statistics – Whole Sample and Indonesians

	Whole Sample			**Indonesians**		
	T(BSS)	**T(LMS)**	**T(PST)**	**T(BSS)**	**T(LMS)**	**T(PST)**
N. of Students	174	177	155	46	44	44
Mean as a %	43.7	44.1	44.2	37.8	40.4	44.1
Percentage Scores S.D.	14.0	20.8	18.0	10.8	15.3	10.7

Analysis

Since each group of students took two tests, one in their own subject area and one not, and since an inspection of the distribution statistics showed that the variances of the different groups were homogeneous*, a repeated measures analysis of variance design was used, with student group and test as the two independent variables. Using SPSS Repeated Measures Manova (SPSS, 1990:362) the results were analysed to see whether there were significant differences between the group scores, the test scores and the interaction of the two.

*The largest difference in the standard deviations of any pair of groups was 14.00 – 9.00; it is only if one S.D. is twice the other that it is necessary to carry out an F ratio test into the homogeneity of variance. See Kerlinger (1973:287) for a discussion of the effect of heterogeneous variance on F test results.

The analyses were carried out in pairs: Groups BSS and LMS with Tests T(BSS) and T(LMS), Groups BSS and PST with Tests T(BSS) and T(PST), and Groups LMS and PST with Tests T(LMS) and T(PST).

The Results

In order to show the direction that any differences in mean scores may be taking, the results of the three main analyses are shown in the form of tables of means. These give the mean scores for each group and each test, the overall means, and the means of the diagonals. The diagonals show the interaction of group and test and therefore the effect of subject area on test scores. The means of the diagonals are given underneath the tables of means.

Table 5.4 shows the results, for the Whole Sample and for the Indonesians, of the LMS and PST groups who took Tests T(LMS) and T(PST). Below the tables of means are the F values and the levels of significance.

Table 5.4

Tables of Means (%) – LMS and PST Groups

Whole Sample

N	Group	T(LMS)	T(PST)	Total
23	LMS	41.6 [1]	43.3 [2]	42.5
17	PST	34.5 [3]	43.1 [4]	38.8
40	All	38.6	43.3	41.0
	39.1	Diag. Mean		42.3

Indonesians

N	Group	T(LMS)	T(PST)	Total
11	LMS	48.7 [1]	43.5 [2]	46.1
10	PST	31.8 [3]	43.3 [4]	37.7
21	All	40.7	43.4	42.0
	37.9	Diag. Mean		46.1

Differences between Means – Significance of Results

Whole Sample

	Means		F ratio	P values
Groups	(LMS) 42.5	(PST) 38.8	.43	NS
Tests	T(LMS) 38.6	T(PST) 43.3	4.20	.05
Subj. Area Effect	39.1	42.3	1.83	NS

Indonesians

	Means		F ratio	P values
Groups	(LMS) 46.1	(PST) 37.7	2.47	NS
Tests	T(LMS) 40.7	T(PST) 43.4	1.13	NS
Subj. Area Effect	37.9	46.1	7.83	.01

NS = Not Significant

In the Whole Sample, the LMS group (mean 42.5) did better overall than the PST group (mean 38.8) though not significantly so. In addition, Test T(PST) was significantly easier for both groups together (mean 43.3) than was T(LMS) (mean 38.6). Looking at the interaction of group and test effect, we would expect students to perform better at the tests in their own subject areas. We would therefore expect the diagonal mean of boxes 1 and 4 together, where LMS students took T(LMS) and where PST students took T(PST), to be higher than the diagonal mean of boxes 2 and 3. These diagonal means will from now on be referred to as means of advantage and disadvantage, since students are supposedly being advantaged if they take tests in their own subject area. In this instance the mean of advantage, the diagonal mean of boxes 1 and 4, is the higher (42.3 as compared to 39.1) but it is not significantly so. Subject area, that is the interaction of group and test effect, is not therefore significantly affecting test results. This may not be considered surprising in this instance as the two tests are both based on scientific subjects, and since all the reading passages are inevitably fairly general in content as they have to be appropriate for a wide range of disciplines within the Physical and Life Sciences.

In the case of the Indonesian sample, however, the results are different. Here there is no significant difference between the two groups, nor between the two tests, but there is a significant subject area effect. Not only is it significant but it is significant at .01, with the effect being in the expected direction: the mean of advantage is higher than the mean of disadvantage. It seems in this case as if students are doing better in their own subject module than in the other. It is of course possible to get such a result by chance – the numbers of students in the Indonesian groups are small, and this was only one of many similar investigations. However, it shows that subject area may be having an effect, and is clearly worth investigating further.

Table 5.5 gives the results of the BSS and PST groups taking T(BSS) and T(PST). For the Whole Sample the results are similar to those already discussed. There is no significant difference between the groups, but there is between the tests – T(PST) is easier than T(BSS). Subject area again has no significant effect, but this time the mean of disadvantage is slightly higher than that of advantage. Students are doing marginally better on the tests outside their own subject areas, but not significantly so.

This time the Indonesians, too, show no significant subject effect, although the contents are this time based on what might be considered more disparate subject areas, Business Studies and Social Sciences, and Physical Science and Technology. This surely throws some doubt on the importance of the previous significant result.

Table 5.5

Tables of Means (%) – BSS and PST Groups

Whole Sample

N	Group	T(BSS)	T(PST)	Total
65	BSS	44.8 [1]	49.0 [2]	46.9
10	PST	34.5 [3]	43.0 [4]	38.8
75	All	43.4	48.2	45.8
		47.1	Diag. Mean	44.6

Indonesians

N	Group	T(BSS)	T(PST)	Total
14	BSS	39.5 [1]	44.2 [2]	41.9
9	PST	35.1 [3]	45.5 [4]	40.3
23	All	37.8	44.7	41.3
		40.6	Diag. Mean	41.8

Differences between Means – Significance of Results

Whole Sample

	Means				F ratio	P values
Groups	(BSS)	46.9	(PST)	38.8	2.91	NS
Tests	T(BSS)	43.4	T(PST)	48.2	7.61	.01
Subj. Area Effect		47.1		44.6	.96	NS

Indonesians

	Means				F ratio	P values
Groups	(BSS)	41.9	(PST)	40.3	.16	NS
Tests	T(BSS)	37.8	T(PST)	44.7	7.5	.05
Subj. Area Effect		40.6		41.8	1.07	NS

The final group to be analysed was BSS and LMS (see Table 5.6).

The results here are generally similar to those above, though in this case the two groups in the Whole Sample are significantly different from one another. The subject area effect is once again not significant but in this case the mean of disadvantage is strikingly higher than the mean of advantage.

Once again the Whole Sample test means are significantly different. For this group of students there seems to be a consistent effect whereby T(PST) is easier than T(LMS) which is easier than T(BSS). (This is not the case for the complete trial test population, where T(BSS) is easier than T(LMS) [see Table 5.1].)

One reason for the general lack of a significant subject effect might have been that some students were embarking on new fields of study. As was the case in the Koh (1985) study, some students with Science degrees were moving on to Business Studies. The analyses were therefore re-run, with students reclassified according to their past field of study. There were few changes in the results. For the Whole Sample PST and LMS groups there was no longer a significant difference between the difficulty levels of tests T(PST) and T(LMS), and for the BSS and LMS groups there was no longer a significant difference between their scores. In no case was there a significant subject area effect.

Table 5.6

Tables of Means (%) – BSS and LMS Groups

Whole Sample

N	Group	T(BSS)	T(LMS)	Total
85	BSS	45.4 [1]	53.0 [2]	49.2
11	LMS	35.4 [3]	38.5 [4]	37.0
96	All	44.2	51.4	47.8
		51.0	Diag. Mean	44.5

Indonesians

N	Group	T(BSS)	T(LMS)	Total
14	BSS	37.8 [1]	37.9 [2]	37.9
8	LMS	36.8 [3]	42.3 [4]	39.6
22	All	37.4	40.0	38.7
		37.5	Diag. Mean	39.4

Differences between Means – Significance of Results

Whole Sample

	Means		F ratio	P values
Groups	(BSS) 49.2	(LMS) 37.0	6.85	.01
Tests	T(BSS) 44.2	T(LMS)51.4	6.21	.05
Subj. Area Effect	51.0	44.5	1.15	NS

Indonesians

	Means		F ratio	P values
Groups	(BSS) 37.9	(LMS) 39.5	.18	NS
Tests	T(BSS) 37.4	T(LMS)40.0	.75	NS
Subj. Area Effect	37.5	39.4	.68	NS

Since in the live test students' raw scores are adjusted for test difficulty, and are reported in the form of a nine-point band scale, and since it is these band levels rather than raw scores which affect students' futures, the analyses were also run again using band levels rather than percentages. Again there were no marked changes. The only major change was that, as might be expected, there was no longer a significant difference between the T(LMS) and T(PST) tests. There was, however, still a significant difference between T(BSS) and T(LMS) and between T(BSS) and T(PST). The p values were <.001 and .05 respectively, and in the case of the former, this implied a difference of half a band. This difference may be of no importance, since the study sample was not typical of the trial population, but it emphasises the fact that the comparative difficulty levels of the live academic modules need to be monitored by the IELTS examiners.

Discussion

It is not possible from this study to make any comments on the effect of background knowledge in general on test scores, since so little is known about these students' backgrounds. However, it is possible to draw some conclusions about the effect of academic background, though it must be emphasised that these are necessarily tentative because of the small number of cases involved.

From this study there is almost no evidence that students were disadvantaged if they took an unsuitable subject module. In all but one of the analyses, subject area had no significant effect on scores, and in two out of the three Whole Sample investigations, students appeared, if anything, to do slightly better at the tests outside their field of study. In the Indonesian sample, there was a significant subject area effect for the students who took the two Science papers. In this case students did better at the test in their own field. However, this is a difficult result to credit since there was no apparent subject area effect for the two groups who took T(BSS) and a Science paper. What this Indonesian result shows is the advantages of having a homogeneous group. With students sharing the same first language, culture and level of previous education, the variability of the results was much reduced. None of the three groups of students were significantly different from each other, and in only one case – tests T(BSS) and T(PST) – were the tests significantly different. However, the number of cases was small; any future study should be based on a larger sample.

Conclusions that might cautiously be drawn from this study are, firstly that the evidence does not show the need for three academic subject modules in the test battery, and secondly that if students are given academic modules outside their subject areas they are not placed at a disadvantage.

However, a study based on a different set of texts might produce different results because the findings inevitably depend on the subject specificity of the reading passages. If one or more of the reading passages within a module are not suitably subject specific, then any subject area effect will naturally be reduced. The texts in this study came from academic or other papers written for the relevant subject areas and were considered appropriate for students in those disciplines . However, some might be more specific to a particular subject area than are others. For example, some of the passages, regardless of their discipline, are written in such a way that readers in other disciplines could easily follow them, whereas others seem to depend on specialist concepts and vocabulary which no one outside the subject area could be expected to understand.

This problem with subject specificity is compounded by the fact that the three IELTS subject areas cover so wide a range of academic fields that they can only loosely be thought of as discipline specific. Any reading passage that has to be appropriate for students across such a range of disciplines must be very broadly based. In addition this breadth of coverage sometimes means that a passage which is appropriate for students within the designated subject area is also appropriate for those in at least one of the other two subject areas. (For an example of this, see Chapter 8, where a PST reading passage in the main study is shown to be appropriate not only for chemists and physicists but also for life scientists.)

The Main Study

Because the question of subject specificity is so important, it was clear that in the main study the appropriacy of each subtest would have to be assessed. A content analysis of the texts and items would have to be carried out, and each passage would be classified according to its level of specificity.

In this pilot study, no account was taken of the students' background knowledge as distinct from field of study, and as Koh (1985) showed, this can lead to misleading results. The scope of the next stage of the research was therefore broadened to include the wider question of the effect of background knowledge on test performance. Students were given a questionnaire which asked not only about their present and future areas of study, but also about subjects studied at school and afterwards. In the short time allowed for the questionnaire it was impossible to discover everything about a student's interests and areas of expertise, but questions were asked about what they read for work and for pleasure, and how familiar they were with the subject matter of the reading test passages.

In the main study, too, there was a large enough sample of students at different stages of language proficiency for an investigation to be carried out into the effect of the interaction between language level and subject area on test performance. Students therefore all took the same test of language proficiency so that they could be placed on one common scale of ability.

Chapter 6 lists the research questions which were asked in the main study, and describes the tests and the questionnaire on which the research was based.

6 The Main Study – Research Questions and Data Collection

Research Questions

From the conclusion to Chapter 5, it can be seen that this main study has two chief aims. The first is to explore the value of including subject specific reading modules in English proficiency tests and the second is to discover more about the effect of background knowledge on reading comprehension.

Academic Field of Study

The research question which was posed in the pilot study ('Do students score significantly higher in a reading test within their academic field of study than in one outside it?') is very general, and any answers to it depend at least partly on which academic fields are targeted. If the fields are restricted to comparatively narrow areas such as civil law, marine biology, or nuclear physics, the study is likely to produce different results from a study based on broad subject areas of the IELTS type. It might be that the three IELTS reading modules cover such wide subject areas that the very breadth of their coverage reduces any subject area effect to almost nothing. The first research question, therefore, which is more specific than the pilot study question, is directly related to the design of the present IELTS:

> **Research Question 1:** *Do students in the three broad subject areas of BSS, LMS and PST achieve significantly higher scores in a reading test within their own subject area than they do in a test in one of the other two subject areas?*

In the pilot study the students were allocated to subject areas according to the discipline in which they would be studying after they had completed their pre-sessional English course. This discipline is sometimes different from that which they had studied previously. It is quite common, for example, for scientists to transfer to business studies. This means that at the time of taking the IELTS reading modules some students may know nothing about their new discipline. Analysing the subject area effect on students grouped according to their future discipline is useful for studies into the validity of IELTS, since in live administrations of the test battery students take the academic module which is closest to their future area of study; however, such a classification is not useful for a study of the effect of background knowledge on test performance. In this main study, therefore, students will be primarily grouped according to their past or present field of study. That is, students attending an English course before progressing to university will be classified according to the subject area in which they previously studied. Students already attending university will be classified according to their present field of study. However, since the question of the effect of future area of study is important for IELTS examiners, for one analysis the students will be classified according to their future discipline.

Research Question 2 is as follows:

> **Research Question 2:** *Do students in the three broad subject areas of BSS, LMS and PST achieve significantly higher scores in a reading test within their future subject area than they do in a test in one of the other two subject areas?*

Level of Academic Studies

One of the questions which the ELTS Revision Project Team spent some time considering was whether there should be separate Academic modules for undergraduates and postgraduates. It was eventually decided that the tests should be aimed at both levels, but it was accepted that the two groups were likely to have differing amounts of subject knowledge at their disposal, since many of the future undergraduates would not yet have received any academic training in their chosen subject area. If this is indeed the case, the fact that the two levels of students are grouped together for the analyses may obscure important differences in subject area effect between the two groups. A further research question therefore is:

> **Research Question 3:** *Do either undergraduates or postgraduates in the three broad subject areas of BSS, LMS and PST achieve significantly higher scores in a reading test within their own subject area than they do in a test in one of the other two subject areas? If there is a significant difference, on which of the two groups is the subject effect the stronger?*

Research Questions 1 to 3 are addressed in Chapter 7.

Subject Specificity of the Reading Passages

The findings from the above Research Questions depend upon how appropriate each of the reading passages within a module is for its specified subject area. Although all the texts were chosen according to the IELTS specifications for the relevant reading module, it may be the case that some are more specific to a particular subject area than are others. For example, as I pointed out in Chapter 5, some passages, regardless of their subject matter, may be written in such a way that readers in other disciplines can easily follow them, whereas others may depend on specialist concepts and vocabulary which no one outside the subject area could be expected to understand. Since the specificity of these texts will affect the results of the study, Research Question 4 asks:

> **Research Question 4:** *Are the reading passages in the three reading modules specific to the appropriate subject area?*

If the reading passages turn out to vary in their specificity, the following research question will be asked:

> **Research Question 5:** *Is it possible to identify some characteristics of the reading passages which make them either more or less specific to their chosen subject areas?*

Research Questions 4 and 5 are addressed in Chapters 7, 8 and 9.

Background Knowledge

All the above questions relate most directly to questions which EAP test writers need to consider before they make decisions about the inclusion and choice of academic reading modules in their test batteries. However, by itself such research is unsatisfactory, since the schemata that students bring to a language test are affected not only by their academic studies but also by many other sources of knowledge. For the next research question, therefore, the scope of the research is broadened to include the effect of background knowledge on test performance.

> **Research Question 6**: *Do students with some familiarity with the subject area of an academic reading test, whether or not it is obtained from formal study in that area, achieve significantly higher scores on the test than students who do not have that knowledge?*

Research Question 6 is addressed in Chapter 10.

Language Proficiency

Of the many variables that play a part in the effect of subject area on test performance it seems likely that level of language proficiency is one of the more important (see Chapter 1). Alderson and Urquhart (1985b), Koh (1985) and Tan (1990) all found that although academic subject area tended to have an effect on reading test performance, level of language proficiency was equally or more important. Research Question 7 is:

> **Research Question 7**: *Which contributes more to EAP students' reading proficiency scores – background knowledge or level of L2 reading proficiency?*

In addition, Coady (1979), Clarke (1980), Cziko (1980) and McLeod and McLaughlin (1986), among others, all reported that the ability of students to make use of top-down reading skills depended on their level of L2 language proficiency. It is likely, therefore, that the effect of subject area on test performance is not the same for students at different levels of L2 language ability. The final Research Question, therefore, is:

> **Research Question 8**: *Does the effect of background knowledge on L2 reading comprehension vary according to the level of L2 proficiency of the reader?*

Research Questions 7 and 8 are addressed in Chapter 10.

Research Methods

Details about the methods used to seek answers to each of the research questions will be given in the appropriate chapters, and the findings will be summarised in Chapter 11.

Research Instruments

In this main study, it is necessary to have a large enough sample of students at different stages of language proficiency for us to see what effect the interaction of language level and subject area has on test performance. For this, students have to be ranked according to some common measure of proficiency. One way of providing this (and it would have the advantage of providing more information about how students performed in subject areas other than their own) would be to give each person all three of the reading modules. Unfortunately, however, the examinees were mostly attending pre-sessional English courses, and not only would few language centres be able to set aside more than half a day for such testing, but students would not have the stamina or the inclination to take three 55-minute tests in one session, especially when two of the three would contain material outside their own subject areas. It would, of course, be possible to reduce the length of the tests, but test reliability would then drop, and the tests would not include the variety of text types which is an important part of the research design. For this stage of the research, therefore, a compromise was adopted: in addition to the two reading modules, all students would take a 'general' grammar test which, although equalling the reading tests in number of items, would be quicker to administer. The students would also complete a questionnaire.

The Reading Tests

For the main study it was not possible to use the same tests as for the pilot study as these formed part of the standardised test battery which was being used round the world. The Reading modules used in the main study, therefore, are parallel versions of the pilot tests, and were prepared for the IELTS Specimen Materials pack. They will be referred to as M(BSS), M(LMS) and M(PST).

The three Reading tests were constructed by teams of two or three item writers, according to the same specifications as those in the Pilot Study. They were not subjected to the same detailed content analysis as the pilot versions as it was presumed that if item writers followed the revised test specifications the tests would be appropriate for their designated students. As a check on this, however, the IELTS Project Steering Committee members studied all the reading passages before the item writers started writing test questions, and where necessary asked for changes to be made. One of the original passages for the PST module was rejected to make way for an engineering text (the original three had all been in the physical sciences), and one of the BSS texts had to be altered so that it was possible to write unambiguous test items on it. The tests were piloted in Australia, Britain and other countries, and were then substantially revised.

The Reading Passages

Each test contains three or four reading passages from different disciplines within the module's overall subject area. Table 6.1 lists the passages and the abbreviations by which they will be known. The complete tests are provided in Appendix 6.1.

In order to check whether there are any gross differences in the linguistic complexity of the texts which might confound the findings of this study, two readability ratings and a simple estimate of clausal complexity have been calculated for each passage.

Readability indices were devised to measure the readability of texts for L1 readers, and are mostly related to vocabulary complexity and sentence length. They have been shown to have high correlations with L1 reading scores and teachers' judgements, but only when the students in the trials have a wide range of ability, and when a large number of texts of varying content and levels of difficulty are included in the study (Carrell 1987b:25). Carrell points out that since individual texts differ so widely in their content and style, the assessment of the readability of single passages is unreliable. The problem is that readability formulae fail to take account of the structural and rhetorical features of a text, and also ignore such factors as the skills and interests that the reader brings to the text. However, for a simple comparison of vocabulary and sentence length they have their uses. Table 6.2, therefore, lists the Flesch Reading Ease and the Gunning Fog indices for each of the ten texts. These indices have been chosen because they are two of the most common (see Klare 1974). The Flesch Reading Ease

Table 6.1

The Reading Passages

Business Studies and Social Sciences (BSS)

1. (Qual)	'Quality Circles' (from *Study Document 342*, Incomes Data Services Ltd. Great Portland Street, London. No date)
2. (Educ)	'The Purposes of Continuing Education' (adapted from *Adult Learning: Issues and Innovations*, edited by Robert M. Smith. ERIC Clearinghouse in Career Education. 1976)
3. (High)	'Access to Higher Education' (from a Department of Education and Science White Paper on Higher Education, London. 1988)

Of the three reading passages, one is in the field of Business Studies (Quality Circles) and the other two are in Education.

Life and Medical Sciences (LMS)

1. (Tooth)	'The Mystery of Declining Tooth Decay' (from *Nature*, Vol. 322, 1986)
2. (Child)	'Our Children's Teeth' (from *The British Medical Journal*, Volume 298, February 1989)
3. (Genes)	'Three Ways to Make a Transgenic Beast' (from the *New Scientist*, 7 July 1988)
4. (Nitro)	'Nitrogen Fixation' (from 'Agricultural Microbiology', *Scientific American*, 1981)

Passages 1 and 2 are in Medicine, and Passages 3 and 4 are in the Life Sciences.

Physical Science and Technology (PST)

1. (Sun)	'Life without a Sunscreen' (from the *New Scientist*, 10 December. 1988)
2. (Fuel)	'Energy from Fuels' (from a textbook written for young non-scientists; source unknown)
3. (Ship)	'The Recovery of the Mary Rose' (from *The Structural Engineer*, Vol. 62a, No. 2, 1984)

Passages 1 and 2 are in the Physical Sciences, and Passage 3 is Engineering.

Table 6.2

Readability Indices

Text	N. Words	Flesch Index	Fog Index
BSS			
Qual	724	39.6	13.9
Educ	721	45.3	14.3
High	1,000	24.5	19.3
LMS			
Tooth	493	26.9	17.7
Child	459	50.8	11.6
Genes	767	45.6	14.5
Nitro	196	19.4	20.2
PST			
Sun	775	44.1	15.8
Fuel	698	42.0	15.8
Ship	1,397	47.8	15.3

index is based on the number of words per sentence and the number of syllables per 100 words, and this version of the Fog index is based on sentence length and the percentage of words with more than two syllables. The lower the Flesch index, or the higher the Fog index, the more difficult the passage is deemed to be. The indices were calculated by *Wordfor Windows* (see Microsoft 1991:291). Not surprisingly, since both formulae are based on sentence and word length, the Flesch and the Fog indices agree fairly closely: in each case they rate (LMS)Child as the easiest passage to read, and (LMS)Nitro as the most difficult. According to Flesch (1962), a Reading Ease Score of 0–30 describes a very difficult text, suitable for college graduates, and a score of 30–50 a difficult text suitable for students in Grades 13–16. A score of 100 means that the text is very easy for any literate person. It can be seen therefore that the Reading Ease Score is not finely tuned for university level texts, and can only show (within its limitations) that with the possible exception of (LMS)Child, all the texts are of a suitable level of difficulty for university students, and that (BSS)High, (LMS)Tooth and (LMS)Nitro are more appropriate for postgraduates.

For the grammatical complexity of the texts two measures were used: firstly an analysis of the percentage of passive rather than active verbs in each passage, since a heavy use of passivisation is sometimes considered to add to the 'malreadability' of a passage (Namukwai and Williams 1988), and secondly a simplified version of the clausal analysis method used at the Laboratory of Applied Cognitive Science at McGill University. This method of clausal analysis was adopted by Turner (1988) for a comparison of reading texts, and is based on Winograd's (1972, 1983) clausal grammar. In the McGill system, sentences are described according to their constituent clauses. The clauses are classified as major or minor, the former being divided into declarative, imperative or interrogative, and the latter into classes such as adjunct, qualifier and noun group, which in their turn are broken down into further sub-classes (see Turner 1988:186). Although Turner initially analysed all her sentences in this manner, when she came to compare the texts she only looked at the number of major and minor clauses per sentence, and the level of embedding, presumably because that was all she required for comparative purposes (see Turner 1988:50). In my study, I simply broke each sentence in the reading passages down into major and minor clauses, and kept a record of the levels of embedding. (See Appendix 6.2 for a graphical display of the mean number of major and minor clauses, levels of embeddings, and the percentage of passives used in each text.)

On the whole the readability indices and the clausal analyses did not reveal any major differences in the sentence level complexity of the different reading texts, although (LMS)Child, which is the most easily readable text according to the two readability indices, also has the lowest mean number of minor clauses. (LMS)Nitro, which has the lowest Flesch Reading Ease Score, does not have a more complex clausal structure than the other reading passages, but its proportion of passives is high.

Although the differences between the texts are not shown to be great, it is perhaps useful to show how they rank in order of difficulty according to the two readability indices, the number of main clauses per sentence, and the number of passives per sentence. Table 6.3 lists these rankings, the easiest passage being assigned a '1' in each case.

Because of their limitations, the above indices cannot tell us much about the comparative readability of the passages. However, they do give us simple information about sentence length, and complexity of vocabulary and syntax, and it is perhaps useful to see that according to these simple measures the passages do not show wide differences in difficulty level.

Table 6.3

Difficulty of the Reading Texts According to Different Indices

Text	Flesch Index	Fog Index	Main Clauses	Passives
BSS				
Qual	7	2	3	6
Educ	4	3	2	4
High	9	9	8	3
LMS				
Tooth	8	8	5	9
Child	1	1	1	7
Genes	3	4	6	1
Nitro	10	10	4	8
PST				
Sun	5	6	9	5
Fuel	6	6	7	2
Ship	2	5	10	10

The Test Items

Item Types

Table 6.4 lists the kinds of test items included in each of the three reading modules. As far as possible the names of the item types follow those used in the test specifications. It can be seen from this list that the modules include a wide range of item types and between them cover almost all the item types listed in the specifications. Three of the item types are not listed in the specifications, but these all come under the general heading 'matching', which is. The only item type which appears in all three modules is summary completion, and even then the method of answering varies across modules: in M(BSS) and M(PST), students use an item bank to select appropriate answers, and in M(LMS) they choose words from the text. In M(PST) there are only four kinds of question, ranging from eight which are information transfer, to thirteen which are summary completion. However, in M(LMS), eighteen items, that is, almost half the test, are devoted to the summary completion task, whereas the other six item types in the module are poorly represented with, for example, only one ordering question and three short answer ones. The item types are therefore ill-balanced, and the strong emphasis on summary completion might penalise students who are weak at such gap-filling tasks. If such students tend to come from only one of the three subject areas under consideration, this might lead to a marked subject area effect for this module. We shall be looking at bias and subject area effect in Chapter 7.

Table 6.4

The Item Types

Module	Item Type		N. of Items
BSS	Choosing from a 'heading bank'		7
	Copying words from the text		11
	Matching diagrams to paragraphs*		4
	Multiple choice		4
	Sorting names/objects into sets		1
	Summary completion (selecting from a bank)		8
		Total	35
LMS	Completing a diagram		4
	Information transfer		4
	Matching examples to statements*		4
	Multiple choice		6
	Sorting events into order		1
	Short answer questions		3
	Summary completion		18
		Total	40
PST	Information transfer		8
	Matching stages in a process*		9
	Short answer questions		10
	Summary completion (selecting from a bank)		13
		Total	40

* Not listed in specifications

It was not always easy to categorise the item types. For example, the M(LMS) Items 29–32 are labelled 'information transfer', but they could also be described as 'copying words from the text'. However, this categorisation was straightforward compared to the identification of the *skills* being tested.

Reading Skills

It is standard practice for reading test specifications to list the kinds of skills which must be tested. However, it is difficult for item writers to follow such requirements, because it is not easy to know what an item is testing. Most reading items test a variety of skills, and different readers bring different processes or skills to bear on a text. Although many taxonomies of reading skills have been drawn up over the years, some of which, such as Bloom's taxonomy of educational objectives (Bloom 1956) and Munby's taxonomy of language skills (Munby 1978), have proved influential in the selection of tasks for teaching and testing reading, it is not clear whether experts can agree on what these skills are. Alderson, for example (Alderson and Lukmani 1989; Alderson 1990a), found that there was often little agreement between applied linguists who were asked to select the skills tested by the items in a reading test. In a follow-up study, Alderson (1990b) also found that when students introspected on their test-taking methods they sometimes appeared to use different processes from each other, and yet reached the same answer.

This is not to say that it is impossible for judges to reach agreement. They can be trained to identify the skills tested in a reading item, and a high level of inter-rater reliability can be achieved (Lumley 1993), especially if the list of skills is devised to match the items under investigation. However, the danger is that this training, which makes judges concentrate on certain agreed skills, will lead to a somewhat narrow and rigid method of assessing items which will obscure valid individual views.

Since a discussion of the comparative performance of the IELTS subtests requires some description of the individual items, I have, in spite of my reservations, attempted to identify the main skill that each item is testing. However I must emphasise that since this list is based on my own views of how each question might be answered, it is likely to be idiosyncratic. In addition, my selection of skills is inevitably based on my own introspection of how I answered the items, and this may itself be faulty. As Matthews says:

> *It is doubtful whether judges ... can retrace the particular route taken by themselves with any degree of reliability.*
> (Matthews 1990)

The fact that I was involved in the original editing of these items, and had marked all the students' test papers may also mean that my view of how the items would be answered is different from that of someone seeing them for the first time. In addition, my approach to the questions will be very different from that of a non-native speaker of English.

As my analysis of the test items can serve not only to compare the content of the three modules, but also to evaluate one aspect of their content validity, I based my skills identification on the IELTS list of academic reading tasks. The list is somewhat limited for my purposes because it does not include any reference to understanding vocabulary or recognising structural relations within the text, both of which play an important part in the comprehension of these items. There are other taxonomies which might be equally or more appropriate for the task, for example, the Barrett Taxonomy (Clymer 1972), Munby's list of Language Skills (Munby 1978), or the Iowa Basic Skills (Hieronymus, Hoover and Lindquist 1986), but none of these precisely cover all the skills that I think are tested by the IELTS reading modules. In many ways, the TEEP list of enabling skills (Weir 1983) would be the most useful, since, as well as covering most of the skills listed in the IELTS specifications, it also includes:

> *deducing the meaning and use of unfamiliar lexical items through understanding word formation and contextual clues*

understanding relations within the sentence
understanding relations between parts of text through cohesion devices especially grammatical cohesion, e.g. pronoun reference.
(Weir 1983:335)

However, because of the need to validate the content of the modules, the IELTS specifications are used in spite of their limitations. Table 6.5 condenses the academic skills into tabular form.

In the same way as M(PST) has a more limited range of item types than the other two modules, see above, it also seems to be testing a more limited range of skills. It has no items asking students to identify the main ideas or the underlying concepts or to identify and compare facts.

None of the modules contains items which require students to read widely in the text and to form their own conclusions. Academic tasks (vii)–(ix) in the IELTS list (see Appendix 4.5), which ask students to evaluate and challenge evidence, formulate a hypothesis and reach a conclusion, are not tested at all. In Chapter 9, I report on three judges' reactions to various aspects of the test item characteristics, and it will be seen that all three judges considered that almost no items asked students to read beyond a single phrase or sentence.

Table 6.5

Academic Skills Being Tested in Each Module

Module	Academic Skill	Number of Items
BSS	Identifying content	8
	Identifying sequence of events	4
	Finding main ideas	11
	Identifying the underlying concept	5
	Identifying cause and effect	3
	Identifying and comparing facts	2
	Drawing logical inferences	2
LMS	Identifying content	15
	Identifying a sequence	5
	Finding main ideas	3
	Identifying relationships	1
	Identifying cause and effect	6
	Identifying and comparing facts	3
	Identifying evidence	3
	Identifying definitions	2
	Drawing logical inferences	2
PST	Identifying content	12
	Identifying a sequence of events	15
	Identifying cause and effect	8
	Identifying a definition	3
	Drawing logical inferences	2

It seems from my analysis that although the range of item types, with the possible exception of M(PST), matches the test specifications, the range of academic skills tested does not: the tests are not testing a sufficiently wide range of skills. This means that in these specimen versions of the IELTS reading modules the students may not be being fully stretched in some important aspects of reading comprehension. It is possible, therefore, that some differences in the specificity of the three modules will not be exposed.

The Grammar Test

The Grammar test was originally part of the General Component of the IELTS test battery. It is designed to assess vocabulary, syntax, reference and cohesion, and, like the Reading tests, is clerically markable. Some of the test items are multiple choice, and the others are open ended, gap filling and ordering questions. (See Appendix 6.3 for a copy of the test paper.)

The Grammar test underwent the same processes of preparation, trialling and editing as the Reading modules, but was later dropped from the test battery. One of the reasons for its exclusion was that it appeared to be testing skills similar to those tested in the reading tests (see Alderson 1993a). The reasons for choosing this test as a measure of the students' language proficiency are given in Chapter 10.

The Questionnaire

The questionnaire has four sections. Questions 1–6 cover basic information about each student such as age, sex and first language. The responses to these questions are reported below. Questions 7–16 cover the students' previous education and future educational plans, and provide the information necessary for Research Question 1. The responses to these questions are reported in Chapter 7. Question 17 asks students about their reading habits and is used for the additional research into background knowledge (Research Question 3). Questions 18–63 ask students for their comments on the reading passages they have just read (Research Question 3), and the responses are reported in Chapter 8. A copy of the questionnaire is provided in Appendix 6.4.

The questionnaire went through several drafts, during which stages the questions were considered by students doing the Questionnaire Design option in the Lancaster University MA in Applied Linguistics and Language Teaching course, and it was finally piloted on four students in the Lancaster University Institute for English Language Education.

The Students

The students in the study were 842 non-native English speakers, most of whom were about to start undergraduate or postgraduate studies at English medium universities. 536 were taking pre-sessional English classes in the UK, and 58 in New Zealand. 101 were attending general English classes in Cyprus, and 147 were science undergraduates in South Africa, Israel and Germany. There was no attempt to make the sample representative of the IELTS examination population, since the research questions are related not just to IELTS but to EAP tests in general, and since it is difficult to find suitable candidates. However, the sample was intended to include students with a wide range of language proficiency. From Tables 6.6 and 6.9 below it will be seen that the examinees formed a heterogeneous sample: they ranged from pre-university students to postgraduates, and spoke 77 different first languages.

Academic Level and Subject Area

Table 6.6 shows the numbers of students intending to follow undergraduate or postgraduate studies, tabulated according to their broad academic subject area. These are the students' past or present subject areas (past for those students who were attending pre-sessional language classes at the time of the test administration, present for those already attending academic courses). As with the pilot study, students' disciplines are classified according to the subject areas listed in the IELTS Administrators' Manual. Students in the column headed 'Combined', were studying in two or three of the subject areas at the same time. For example, some students were studying biology and chemistry which are in the LMS and PST subject areas respectively. (Note that because of missing data, the total number of students varies from table to table.)

Table 6.6

Future Academic Level and Present or Past Field of Study

Level	BSS	LMS	PST	Combined	Nothing	Total
Pre-Undergraduate	2				1	3
Undergraduate	92	68	63	41	48	312
Postgraduate	213	78	108	20	2	421
Technical Training	8	1	5	2	2	18
Teacher Training	21	1		2	3	27
Not Sure	6					6
Total	342	148	176	65	56	787

The above table shows that there was a preponderance of BSS students. This may partly be due to chance as different pre-sessional courses in Britain have varying proportions of BSS, LMS and PST students. At the time of the trials, for example, Exeter and Lancaster had almost exclusively social science students, whereas in Birmingham and Newcastle the numbers of social, life and physical scientists were more evenly balanced. However, it may also be due to changes in the study patterns of students coming to Britain. In 1986, when the revision of ELTS was mooted, there were approximately the same numbers of ELTS takers in the three subject areas. However, by October 1991 the picture was very different (see Table 6.7).

Table 6.7

Total Numbers Taking IELTS Academic Modules – October 1989 to October 1991

BSS	17,868
LMS	5,654
PST	6,372

Source: UCLES, October 1991

Table 6.8 lists students' academic levels according to future subject area, as some of the students in the sample were in the process of changing from one field of study to another. Few were changing from BSS to the sciences, or from one science to the other (three or fewer cases in each instance), but 22 students were moving from LMS to BSS, and 22 from PST to BSS. Most of these were changing to business studies or accounting.

Table 6.8

Future Academic Level and Future Field of Study

Level	BSS	LMS	PST	English Course	Total
Pre-Undergraduate	2		1		3
Undergraduate	121	76	83	29	309
Postgraduate	270	61	95	1	427
Technical Training	9	1	7	1	18
Teacher Training	26		2	1	29
Not Sure	2				2
Total	430	138	188	32	788

Language

The examinees spoke 77 different first languages. Table 6.9 lists those spoken by 20 or more students in the sample. The other 203 are included under 'Other'.

Table 6.9

Examinees' First Languages

Language	N. of speakers	Language	N. of speakers
Arabic	64	Japanese	80
Bengali	20	Spanish	50
Chinese	48	Thai	41
German	97	Turkish	21
Greek	121	Zulu	20
Hebrew	41	Other	203
Indonesian	36	Total	842

Test Administration

In some British centres the tests were compulsory for whole groups of pre-sessional students and the marks were used either for an end-of-course assessment or as part of a report to future academic tutors. In other cases the tests were optional, and students were told that the tests would give them an idea of how their English was progressing. All the students in the British trials appeared to take the tests very seriously, regardless of whether the test was optional or compulsory. The students could be seen to be working hard, and according to their supervisors were keen to know their results. In the overseas trials local course supervisors gave the tests, and in all cases the tests were compulsory.

For all the trials there was a set pattern of administration. All students took two reading modules, one in their own subject area, and one in another, and to avoid test order effect, half the students took their own subject module first, and half took it second. To cut down the possibility of cheating, different modules were handed to alternate students so that people sitting next to each other were not taking the same module at the same time. When administrators collected the first batch of papers, they wrote a '1' on each of the papers to record which test had been administered first. The order of the tests was as follows:

1 First Reading Module
2 Second Reading Module
3 The Questionnaire
4 The Grammar Test

The questionnaire came directly after the Reading tests because it included questions about the reading passages, and it was important that the students answered these questions while they still remembered the texts. Administration of the questionnaire before the end also ensured that it was filled in by everybody, as no one wanted to leave before the Grammar test.

Test administrators were on hand to explain the questionnaire to students, and on the whole students appeared to find it easy, though time consuming, to answer. There was some difficulty with Question 7, 'School Education', as 'school' has many different meanings throughout the world, but this was anticipated, and administrators explained to the students what was wanted . The main problem was that some students only answered questions relating to one of the two Reading modules they had taken. Administrators were asked to check each questionnaire for this as it was handed in, but inevitably some incomplete questionnaires were not identified in time.

Two aspects of the administration varied from centre to centre. Firstly, at some centres the test was given in one large auditorium, whereas in others it was given in smaller classrooms. When a large auditorium was used the test administrator was helped by one or more teachers. Secondly, the positioning of the break was not fixed. At some centres the time fixed for a break was unalterable, and the trials had to fit in with it. This meant that although the break generally came after the second reading module, it occasionally came after the first. There was some concern that in these instances students might discover the contents of their second module by asking students who had already taken it. However, since students did not know what test they would be taking next, there was no cause for them to do this.

Marking

Many of the open-ended test items had more than one possible answer, so I marked them all myself in order to make the marking as consistent as possible. Most of the acceptable answers had been decided during the trial stages of the tests, but some further possibilities emerged as the scripts were marked. These were progressively added to the answer key. The final answer keys are provided in Appendix 6.1.

Test Results

Descriptive Statistics

Table 6.10 shows the distribution of scores for each of the four tests.

Table 6.10

Distribution Statistics for the Whole Sample

	M(BSS)	M(LMS)	M(PST)	Grammar
Number of Students	634	513	527	833
Number of Items	35	40	40	38
Mean Raw Score	16.87	18.59	18.25	26.24
Raw Score S.D.	6.60	7.91	7.50	6.24
Mean Score as a %	48%	46%	46%	69%
Percentage Score S.D.	18.86	19.78	18.75	16.42
Mode as a %	43%	30%	33%	74%
Median as a %	46%	45%	43%	71%
Max/Min as %	97%/6%	98%/3%	100%/5%	100%/13%

Percentages are rounded up to the nearest whole number.

The standard deviations and ranges show that there is a wide spread of scores in this sample, with high and low level students well represented. Scores in M(PST), for example, range from 100% to 5%, and all the standard deviations except for that of Grammar are high. (The Grammar test was intended to be easier, and was expected to bunch the students more closely together.) However, all three distributions show positive skews M(BSS) .22, M(LMS) .29, M(PST) .51. This can be explained by the fact that comparatively few high proficiency students attend pre-sessional language courses. However, some students, as can be seen from the score distribution tables in Appendix 6.5, did achieve very high scores.

The three reading tests seem to be of a similar level of difficulty, and this is borne out by the analyses of variance described in Chapter 7, which show that there are no significant differences between any of the test score means. Although the Grammar test is easier than the Reading modules, it still has a wide enough spread of scores to be useful as a measure of the students' language proficiency.

The above figures show the results of the whole sample. They include the scores of those students who took modules which were outside their own subject area. The distribution of scores of those who took the module in their own subject area are shown in Table 6.11.

Table 6.11

Distribution Statistics for Students Taking Reading Modules in Their Own Subject Area

	M(BSS)	M(LMS)	M(PST)
Number of Students	341	161	200
Number of Items	35	40	40
Mean Raw Score	17.81	20.70	18.49
Raw Score S.D.	6.85	8.82	7.18
Mean Score as a %	51%	52%	46%
Percentage Score S.D.	19.57	22.05	17.95
Mode as a %	43%	38%	48%
Median as a %	49%	50%	45%
Max/Min as %	97%/9%	98%/8%	100%/10%

The M(BSS) and M(LMS) distributions are both different from those in Table 6.10 above. Both no longer have skewed distributions (see the score distribution tables in Appendix 6.5), and their means are significantly higher*: whole sample: M(BSS) 48%, M(PST) 47%; subject specific sample: M(BSS) 51%, M(LMS) 56%. In addition the standard deviations are higher. At first glance it looks as if these tests are more appropriate for students in the relevant subject areas than for students in the other subject areas, since we might expect students to do better in their own subject area. However, the distribution of scores for PST students who took M(PST) does not show a similar difference from the whole sample results. The results of both the whole and the select sample have positive skews, and the mean and standard deviation are little changed (whole sample: mean = 46%, S.D. = 18.74; PST students: mean = 46%, S.D. = 17.95). The similarities or differences may of course be caused by variations in the language abilities of the two groups. It could be that the PST students were of a lower level of English proficiency. I shall discuss this further in Chapter 7.

Item Analysis and Reliability

On the whole the items worked well: the mean point biserial correlations ranged from .40 to .47, there were no negative discriminations and few of lower than .2 (pt. biserial). (For a complete set of the item statistics, see Clapham 1994.) What is most striking, however, is that many items have a high proportion of unattempted answers, and that these often contribute to the high discrimination indices. For example, Item 40 in M(PST) has a point biserial correlation index of .49, which is satisfactory although the item is difficult (percentage correct [F.V.] = 18%) and only 37% of the students answered it. If an item is to discriminate well, any wrong answer to that item should correlate negatively

* M(BSS) $t = 2.54$ $df = 340$ $p = <.01$ M(LMS) $t = 3.04$ $df = 160$ $p = <.01$

with the total scores. In this case the wrong-answer discrimination is little different from zero (pt. biserial = -.03) and therefore it is not contributing to the discrimination. However, the discrimination for students not answering the question is -.36.

In the main it seems clear that the blanks were due to shortage of time. The number of blanks increases as the tests progress, and many students never attempted the last subtest at all. This is particularly true of M(PST), where 21% of the students did not attempt the final subtest of 13 items. Some students did not complete earlier subtests. They had been advised to move on from sections which they found difficult, and to return to them later if they had time, and I observed during the test administration that some students followed this advice and then did not have time to return to earlier sections. These unfinished test papers are discussed in more detail in Chapter 7.

The fact that so many students did not finish the reading tests is not surprising. The tests were specifically designed to force students to read quickly, and in some cases to skim and scan, since those were considered essential skills in reading for academic purposes. It was expected that students with higher scores, would, therefore, have the ability to get through the texts more quickly. However, this poses a certain problem for the analysis of the tests' reliability, since it should preferably be assessed on tests in which candidates have been given time to finish every item. However, if one of the aims of a test is to test speed, then it seems unrealistic to leave that aspect out when calculating reliability indices. Crocker and Algina (1986) recommend that under these circumstances it is best to use the test-retest, or equivalent forms method of assessing reliability. However, that would have been impossible during the present study because of time constraints. Crocker and Algina do not rule out the use of Kuder Richardson (20) under such circumstances, but say that the ensuing reliability estimates must be 'interpreted with caution' (Crocker and Algina, 1986:145). Table 6.12, therefore, gives the Kuder Richardson (20) reliability indices, but it must be remembered that these may be somewhat inflated because the tests were speeded.

Table 6.12

Reliability (K.R. 20) for the Overall Group and the Appropriate Subgroup

	M(BSS)	M(LMS)	M(PST)	Grammar
Number of Items	35	40	40	38
Whole Sample	.85 (n=634)	.88 (n=513)	.88 (n=527)	.86 (n=833)
Appropriate Subgroup	.86 (n=341)	.91 (n=161)	.88 (n=200)	-

The four tests seem to be adequately reliable, especially as none of them has more than 40 items. They also seem to be equally reliable whether they are taken by all candidates, or by those in the appropriate subject groups. Although there are marginal increases in the reliability of the M(BSS) and M(LMS) tests when they were taken by the appropriate students only, the M(PST) reliability index did not change.

Intercorrelations

Table 6.13 gives the Pearson Product Moment intercorrelations between the three reading modules and the Grammar test. (Since there were few differences in the results of the whole sample and the subject specific groups, these correlations are based on all students who took the tests.) The reading modules all correlate with each other at about the same level. The correlation coefficients with Grammar are all slightly lower - ranging from .51 to .59, rather than from .63 to .73 – but the differences are small. It is not surprising that the Grammar correlations are lower, as Grammar is intended to test grammatical proficiency rather than reading, and requires productive as well as receptive skills. However the differences are not great: the correlation between Grammar and M(BSS) is .59, whereas that between M(BSS) and M(PST) is .63.

Table 6.13

Intercorrelations Between the Tests

	M(BSS)	**M(LMS)**	**M(PST)**	**Grammar**
M(BSS)	-	.73 (310)	.63 (328)	.59 (634)
M(LMS)	.73 (310)	-	.71 (204)	.57 (506)
M(PST)	.63 (328)	.71 (204)	-	.51 (526)
Grammar	.59 (634)	.57 (506)	.51 (526)	-

N. of subjects are shown in parentheses.

Concurrent Validity

It was difficult to run concurrent validation studies on the reading modules because there were no validated tests to which they could be compared. Some of the language centres at which the tests were given did have their own placement tests, but these had not been validated. A few students had taken the complete IELTS battery, or the TOEFL or Cambridge First Certificate (FCE), and these scores were collected in the students' questionnaire, but these scores were for the whole test batteries, rather than the reading components.

Table 6.14 gives the correlations between students' scores on the reading modules, the grammar test and the three external exams.

Table 6.14

Correlations with FCE, IELTS and the TOEFL

	FCE	IELTS	TOEFL
M(BSS)	.09* (74)	.60 (190)	.48 (103)
M(LMS)	.32 (51)	.47 (131)	.33 (61)
M(PST)	.17* (71)	.34 (153)	.39 (68)
Grammar	.35 (98)	.50 (236)	.49 (116)

*Not significant at .05
N. of students are shown in parentheses.

On the whole the above correlations are low, which is to be expected, since not only were the reading tests being compared to overall proficiency tests, but the students listed their IELTS and TOEFL test scores themselves (most of the First Certificate scores were provided by a Language Centre), and, although there is no evidence that students report their scores dishonestly, it is likely that some will not remember them accurately. In addition the students had not taken the other test battery directly before they participated in this study. The questionnaire had asked them to give the results of any tests taken over the previous year. They might, therefore, have taken them up to a year before the study, and during that year their English might have improved or deteriorated. Under the circumstances it is perhaps surprising that some of the correlations are as high as they are. The correlations with the FCE grades are the lowest, which is to be expected as the FCE is not specifically written for academic students. It is also not surprising that the correlations between the test batteries and Grammar are marginally the highest. The Grammar test is testing productive as well as receptive skills, and is therefore more likely than the reading tests to be testing general linguistic proficiency. The highest correlation of all is between M(BSS) and IELTS. This correlation is almost as high as the correlation between M(BSS) and M(PST) in Table 6.13 above. Since IELTS includes a reading module similar to the one the students took in the study this is perhaps not surprising, but since the correlations between the total IELTS score and the other two reading modules are not so high, the M(BSS) correlation could be due to chance.

Conclusion

In this chapter I have outlined the research questions for the main study, introduced the research instruments, and described the student sample. One of the important points to emerge was that the sample was heavily weighted towards BSS students, but that there were enough LMS and PST students for all the main analyses to be run. The three reading modules proved to be of comparable difficulty, and all three had reliability indices of .85 or above.

Chapter 7 starts with a replication of the repeated measures analyses of variance that were carried out in the pilot study, and then looks at the subject specificity of the individual subtests.

7 The Effect of Academic Subject Area on Test Performance

We now come to the results of the main study. This chapter describes a replication of the analyses run on the pilot data (see Chapter 5), and attempts to account for the differences between the findings of the two studies. Chapters 8 and 9 look at the choice of reading texts and the content of the passages and test items. Chapter 10 widens the scope of the enquiry from the question of whether students perform better on reading tests in their own academic areas, to the broader issue of the effect of background knowledge on reading comprehension.

In this present chapter, Section 1 reports on the effect of subject area on students' performance on the complete academic modules. Section 2 studies the same effect on performance on the constituent subtests, Section 3 describes the results of tests of item bias which were used to see whether students were disadvantaged if they took subject modules outside their own field of study, and Section 4 looks at unfinished test papers to see whether students appear to complete their own subject module more quickly than the other one.

Section 1: Analysis of Variance – the Complete Tests

Past Subject Area

As in the pilot study, students' test scores were compared to see whether they were higher on the test in their own subject area than on the test outside it. For this analysis the students were grouped according to their present or past subject area, so that, for example, PST students who were about to take a BSS subject were listed under PST. Three repeated measures analyses of variance were carried out, one on each pair of tests: (M)BSS and (M)LMS, (M)BSS and (M)PST, (M)LMS and (M)PST.

Table 7.1 gives the table of means of the first pair, showing how the BSS and LMS groups fared on each test singly and on the two tests together. The BSS students had a mean of 52.2% for the BSS module, and 46.5% for (M)LMS; their mean score for the two tests was 49.4%. This was little different from the LMS students' overall mean of 47.5%. The F value is 0.57 which is not significant – the two groups did equally well on the two tests. Similarly the overall means of the two tests, 50.4% and 47.2% for M(BSS) and M(LMS) respectively, are not significantly different ($F = 3.12$).

Table 7.1

Table of Means (%) – Groups BSS and LMS

N	Group	M(BSS)	M(LMS)	Total
175	BSS	52.2	46.5	49.4
78	LMS	46.3	48.7	47.5
253	All	50.4	47.2	48.8
	46.4	Diag. mean		51.2

Significance of Results

	F	P
Between Groups	.57	NS
Between Tests	3.12	NS
Interaction (Subject Area Effect)	19.48	<.001

The effect of subject area on these test scores is shown in the diagonals within the table of means. These show the interaction between the group and the test effect. Since students might be expected to be at an advantage if they took a test in their own subject area, we should expect the BSS students to have higher scores in (M)BSS, and the LMS students in (M)LMS. The mean of these two (51.2%), the mean of advantage, is shown at the bottom right of the table of means, and the mean of disadvantage, 46.4%, at the bottom left. In this case the mean of advantage is appreciably the higher of the two, with the difference significant at <.001. There is therefore a highly significant subject area effect. The students did better at their own subject module than they did at the other. Since this is what might be expected, it would not be a surprising result if it were not for the fact that the pilot study result was so very different. Admittedly the number of cases in that was much smaller. There were 85 BSS students and 11 LMS, and the smaller the number of cases the harder it is to achieve a significant result. However, that may not be the only reason. No analysis was made of the pilot study reading passages. Because they came out of journals and magazines in the appropriate subject areas, they were expected to be appropriate and 'specific' to the relevant reading modules, and therefore by implication to be unsuitable for, or unfamiliar to, students in other disciplines. However, this was not necessarily so. Academic papers, let alone magazine articles, vary in their level of subject specificity. In linguistics journals, for example, academic papers range from ones comprehensible to readers outside the social sciences to highly technical ones which can only be read by experts in some limited sub-field. It may well be that the reading passages in the pilot tests vary in just this way. I shall discuss the specificity of the individual subtests below.

Table 7.2 gives the tables of means and the F values for the other two test pairings, (M)BSS/(M)PST and (M)LMS/(M)PST.

Table 7.2

Tables of Means (%) – BSS/PST and LMS/PST

Groups BSS and PST

N	Group		M(BSS)	M(PST)	Total
167	BSS		48.4	42.8	45.6
105	PST		44.4	45.6	45.0
272	All		46.8	43.9	45.4
		43.4	Diag. mean		47.32

Significance of Results

	F	P
Between Groups	.09	NS
Between Tests	5.78	NS
Interaction		
(Subject Area Effect)	13.11	<.001

Groups LMS and PST

N	Group		M(LMS)	M(PST)	Total
70	LMS		55.9	54.0	54.9
89	PST		43.3	46.3	44.8
159	All		48.8	49.7	49.3
		48.0	Diag. mean		50.6

Significance of Results

	F	P
Between Groups	11.87	.001
Between Tests	.24	NS
Interaction		
(Subject Area Effect)	4.39	.05

Once again, unlike the pilot study, the subject area effect is significant for both groups, with p = <.001 for BSS/PST and .05 for LMS/PST. It is not surprising that the level of significance for LMS/PST is lower than that for the BSS/LMS and BSS/PST ones. Many, indeed perhaps most, students in the Life and Medical Sciences have to study physics and chemistry before they embark on university medical or biological studies and they might therefore be expected to do just as well as the PST students on the PST test. The PST students, however, may not have studied any of the life sciences except elementary biology. The LMS students, therefore, might do as well as the PST students on M(PST), but the PST students might not do equally well on M(LMS). In the present study the LMS students have done better than the PST students on both tests. The LMS group mean of 54.9% is significantly higher than that of the PST students (44.8%). This is at least partly due to the fact that the LMS students in this sample seem to have a higher level of language proficiency than the PST students. (Although there is no significant difference between the level of difficulty of the two tests, the LMS students had scores that were approximately 10% higher on each one.) The significant group effect may also be caused by a one-sided LMS/PST subject effect, with PST students doing better on their own subject module than on M(LMS), but LMS students doing equally well on both. If this does turn out to be the case it naturally has implications for the IELTS test battery. The examination board could save itself some time and effort by providing a PST-based module for all science students.

Future Subject Area

As I mentioned in Chapter 5, IELTS candidates are not given the module which is closest to their past field of study but are given the one which is related to their future academic subject area. The reason for this is that as the test is a proficiency test, which is designed to be predictive and to show how well students will cope in their future disciplines, students are given the academic module nearest to that future subject area. Since, as we saw in Chapter 6, 44 students in the main study were about to change from science to BSS subject areas, 22 from LMS and 22 from PST, repeated measures analyses of variance were run to see whether the subject area effect was different if students classified according to their future rather than their past subject area. Table 7.3 gives the mean totals, and F. and p values.

Table 7.3

Repeated Measures Analysis of Variance – Future Area of Study

		Means (%)			F	P
BSS/LMS (n = 218/63)						
Groups	(BSS)	48.9	(LMS) 48.2	Between Groups:	.07	NS
Tests	M(BSS)	50.6	M(LMS)46.9	Between Tests:	2.29	NS
Diag. means:						
Disadvant.		46.4	Advant. 51.1	Subject Area Effect:	14.29	<.001
BSS/PST (n = 199/76)						
Groups	(BSS)	46.4	(PST) 42.6	Between Groups:	2.87	NS
Tests	M(BSS)	47.2	M(PST) 43.5	Between Tests:	5.96	.05
Diag. means:						
Disadvant.		43.4	Advant. 47.3	Subject Area Effect:	8.41	.01
LMS/PST (n = 74/100)						
Groups	(LMS)	55.1	(PST) 44.2	Between Groups:	16.89	<.001
Tests	M(LMS)	48.2	M(PST) 49.5	Between Tests:	.63	NS
Diag. means:						
Disadvant.		47.5	Advant. 50.1	Subject Area Effect:	4.37	.05

Although there is little difference between these results and those presented in Tables 7.1 and 7.2 above, there is enough to cause one to query the IELTS examiners' policy. The significance of the subject area effect between BSS and PST has dropped from <.001 to .01, and although the subject effect between BSS and LMS is still highly significant (p = <.001.) the F. ratio has dropped from 19.48 to 14.29. Not surprisingly, these students are not doing significantly better in their supposed subject area (BSS), because this subject area is actually new to them. This drop in F. ratios and significance levels, although small, supports the hypothesis that students do better in the subject area with which they are more familiar, and it therefore seems unreasonable to expect students who are changing their study area to take a module which is outside their present field of study. Such candidates who are changing disciplines will not be able to make use of their subject area knowledge in the same way as the other students. They are therefore being put at a disadvantage, and their test results may not accurately reflect their academic language proficiency. It is true that they will have to study in an unfamiliar area, but it would probably be more fair to them to give them a reading module in their own subject area, and to expect them to acquire the relevant skills for the new subject area once they have started their course, in the same way that native speakers do. However, this is a controversial area, which I will be in a better position to discuss once I have presented more of my results. I shall therefore return to this point in my concluding chapter.

Academic Level

As we saw in Chapter 4, one of the reasons why the IELTS Project Steering Committee agreed that undergraduates and postgraduates should take the same IELTS modules was because there was little evidence of major differences in the study skills requirements of the two groups. It might be presumed, however, that undergraduates would have less need of subject specific reading modules than postgraduates because their previous studies, at school, would have usually been more general. In the questionnaire students were asked for information about the subjects studied during their last two years of school and during their subsequent undergraduate careers. A comparison of the responses to these two questions is presented in Table 7.4.

Table 7.4

Percentage of Students Studying BSS, LMS and PST at School and as Undergraduates

	Last 2 years of School	Undergraduate
BSS	13.1%	43%
LMS	.4%	27%
PST	1.8%	30%
BSS/LMS	5.8%	-
BSS/ PST	15.6%	-
LMS/PST	5.6%	-
BSS/LMS/PST	53.8%	-
Missing	3.9%	-
Total	100.0%	100.0%

From this table it will be seen that during their last two years of school more than half the students were studying subjects in all three subject areas, whereas only 0.4% were studying just LMS, and only 1.8% just PST. During their undergraduate studies, however, no students studied in more than one of the subject areas.

As students become more specialised in their studies more disciplines are excluded from their area of study. It might be expected, therefore, that postgraduates would be more affected by the subject area of reading passages than undergraduates. On the other hand, as they become older, students' general knowledge is likely to become greater and may offset the effects of this specialisation. To see whether undergraduates or postgraduates were most affected by the subject areas of the reading tests, repeated measures analyses of variance were calculated separately for future under- and post-graduates. Table 7.5 gives the results.

Table 7.5

Repeated Measures Analysis of Variance – Academic Level

Undergraduates

BSS/LMS (n = 48/43)		**Means (%)**		**F**	**P**
Groups	(BSS) 50.3	(LMS) 50.9	Between Groups:	.02	NS
Tests	M(BSS) 51.3	M(LMS)50.0	Between Tests:	.53	NS
Diag. means:					
Disadvant.	48.4	Advant. 52.8	Subject Area Effect:	8.44	.01
BSS/PST (n=40/49)					
Groups	(BSS) 47.3	(PST) 45.2	Between Groups:	.33	NS
Tests	M(BSS) 46.9	M(PST) 45.4	Between Tests:	.89	NS
Diag. means:					
Disadvant.	45.8	Advant. 46.5	Subject Area Effect:	.27	NS
LMS/PST (n = 27/39)					
Groups	(LMS) 68.8	(PST) 44.6	Between Groups:	29.48	<.001
Tests	M(LMS)53.8	M(PST) 55.2	Between Tests:	.33	NS
Diag. means:					
Disadvant.	53.6	Advant. 55.3	Subject Area Effect:	.69	NS

Postgraduates

BSS/LMS (n = 120/35)		**Means (%)**		**F**	**P**
Groups	(BSS) 49.2	(LMS) 43.3	Between Groups:	3.12	NS
Tests	M(BSS) 50.2	M(LMS)45.6	Between Tests:	3.67	NS
Diag. means:					
Disadvant.	45.2	Advant. 50.5	Subject Area Effect:	9.37	.01
BSS/PST (n=121/56)					
Groups	(BSS) 45.0	(PST) 44.8	Between Groups:	.00	NS
Tests	M(BSS) 46.8	M(PST) 43.1	Between Tests:	2.85	NS
Diag. means:					
Disadvant.	42.2	Advant. 47.7	Subject Area Effect:	15.89	<.001
LMS/PST (n = 43/50)					
Groups	(LMS) 46.2	(PST) 45.0	Between Groups:	.12	NS
Tests	M(LMS)45.3	M(PST) 45.8	Between Tests:	.04	NS
Diag. means:					
Disadvant.	44.0	Advant. 47.2	Subject Area Effect:	3.95	.05

The first thing to notice about these results is that the significance levels are mostly lower than those for all students together (see Tables 7.1 and 7.2). This is probably due to the smaller sample sizes. On the whole the suggestion that postgraduates might be more affected by subject area than undergraduates is supported. Although there is no difference in the significance levels of the subject area effects for the BSS/LMS under- and post-graduates (.01 each time), for the other two pairings there is. For the undergraduate LMS/PST and BSS/PST students there are no significant subject area effects, whereas for the postgraduates the p values are .05 and <.001 respectively. This BSS/PST change from NS to <.001 is striking and cannot be accounted for simply by the increased size of the sample.

Section 2: The Reading Subtests

The above analyses of variance have shown that students did better at the module in their own broad subject area, and therefore that these three modules seem to be appropriate for students in the relevant fields of study. However, it may be the case that within these modules the constituent subtests vary in their 'specificity'. A 'specific' test or subtest is one which is based on content which is appropriate for students in the relevant subject area, and is not appropriate for students in the other subject areas. Even a cursory glance at the reading passages reveals that the texts appear to vary in their level of specificity. One of the PST passages, for example, on fuel resources, seems to be the sort of general informative text that could be read by anyone in the BSS field, whereas two of the LMS texts seem more subject specific as they depend on at least some basic medical or biological knowledge. Appearances can, of course, be deceptive, but it is worth seeing whether these texts are at the appropriate level of specificity for these modules.

In Chapters 8 and 9 I shall be discussing the content of the reading passages in some detail. In this chapter, therefore, I shall only give as much information as is necessary for an understanding of the ensuing discussion of the analysis of variance results.

Table 7.6 lists the reading passages, together with the titles by which I will refer to them, and my own impressionistic view of whether each text is specific or general. We shall see later how well my impressions stand up to scrutiny.

Table 7.6

Subject Areas of the Reading Passages – Specificity Table 1

Title	Abbreviation	Specificity	Subject Area
BSS Module			
Quality Circles	(Qual)	Specific	Business Studies
The Purposes of Continuing Education	(Educ)	Specific	Education
Access to Higher Education	(High)	Specific	Education
LMS Module			
The Mystery of Declining Tooth Decay	(Tooth)	General	Dental Health
Our Children's Teeth	(Child)	General	Dental Health
How to Make a Transgenic Beast	(Genes)	Specific	Genetics
Nitrogen Fixation	(Nitro)	Specific	Plant Biology
PST Module			
Life Without a Sunscreen	(Sun)	Specific	Physics
Energy from Fuels	(Fuel)	General	Physics
The Recovery of the Mary Rose	(Ship)	Specific	Engineering

Table 7.7 lists the descriptive statistics for each subtest. Since there appeared to be no major differences between the distribution of scores of the total sample, and those of the samples of students taking their appropriate modules (see Chapter 6), the results here are based on the total sample.

Table 7.7

Descriptive Statistics for the Reading Subtests

	Text	N. of Items	Mean	S. D.	Reliability (KR 20)
BSS Module	Qual	12	54%	26.34	.81
n = 634	Educ	15	49%	23.93	.78
	High	8	38%	24.16	.64
LMS Module	Tooth	18	48%	27.61	.89
n = 513	Child	9	60%	21.76	.53
	Genes	9	37%	27.68	.78
	Nitro	4	30%	31.55	.67
PST Module	Sun	11	61%	24.98	.77
n = 527	Fuel	7	70%	29.40	.79
	Ship	22	30%	22.78	.87

One of the tests, Nitro, is very short, with only four items, but it seemed worth including in the survey as its reliability index is surprisingly high.

It will be noticed that the last subtest in each module is always the most difficult. This may partly be because the specifications require item writers to try and make the reading passages within a module progressively more difficult, but it is also partly because many students did not complete the last subtest because they ran out of time.

Repeated measures analysis of variance was carried out on all the subtests, following the procedure described earlier. In each case the analyses were carried out in pairs, so that, for example, the performance of BSS and LMS students on (BSS)Qual and (LMS)Tooth was compared, and similarly that of BSS and PST students on (BSS)Qual and (PST)Sun, and LMS and PST students on (LMS)Tooth and (PST)Sun. In this way all possible pairs of subtests were tested for subject area effect. Table 7.8 shows the difference in the effect of academic subject area for each pair of subtests. Table 7.8.1 shows the subject area effect when BSS and LMS students took the BSS and LMS subtests, and when BSS and PST students took the BSS and PST ones. Similarly, Tables 7.8.2 and 7.8.3 show the pairings with LMS students and PST students respectively. It will be seen from this chart how the subtests vary in their 'specificity', that is in the extent to which they appear to have significant subject area effects when paired with subtests from other modules. The LMS subtest Nitro, for example (Table 7.8.2), has significant subject area effects when paired with all the subtests from the other modules ($p = .01$, $<.001$ and $.01$ with the PST subtests Sun, Fuel and Ship respectively, and $.001$, $<.001$, and $<.001$ with each of the three BSS subtests). On the other hand, (PST)Ship (Table 7.8.3) has no significant subject area effects with any of the BSS subtests and only with one LMS subtest – Nitro. When BSS and PST students' scores are compared on (BSS)Qual and (PST)Sun (Table 7.8.1) there is a highly significant subject area effect (p = <.001), but when those same students' performances are compared on (BSS)Qual and (PST)Fuel there is no significant difference. Of course any interpretation of these results is confounded by the fact that there are always two subtests contributing to an effect, and one of these may be subject specific and the other not. However, some subtests have consistently higher significant effects than others. For example, (LMS)Nitro, as shown above, has significant effects with every one of the subtests, whereas (LMS)Child has none.

When a subtest has highly significant subject area effects when paired with all or almost all the other subtests it can be presumed that the subtest is highly specific to students in that subject area. Such subtests will therefore be labelled 'highly specific'. So (LMS)Nitro is a highly specific subtest. Subtests such as (BSS)Qual and (PST)Sun, which have highly significant effects with some but not all the other subtests, will be called 'specific', and subtests which have few or no significant effects will be called 'general'. (LMS)Tooth, (LMS)Child, (PST)Fuel and (PST)Ship are all 'general' according to this definition. One intriguing finding is that in all three subject area interactions between (LMS)Child and the three BSS subtests, and between (LMS)Child and (PST)Ship the mean of disadvantage is higher than the mean of advantage. Since the differences are not significant one such result by itself would not be important, but when all three of the BSS/LMS results are the same it looks as if the results may not be due to chance alone. It may be that the subtest is marginally more appropriate for BSS than for LMS students. The same occurs between (PST)Fuel and the three BSS subtests. In all three cases the mean of disadvantage is the higher. In this case this result is not surprising. The Fuel reading passage was taken from a textbook explaining scientific facts to Arts and Social Science students. It may well be, therefore, that although the passage contains material in a suitable subject area, it is not presented in a scientific manner or genre.

In all but one case my intuitions about the specificity of the subtests (see Table 7.6) seem to have been borne out. However, in the case of the PST text, Ship, it was not. This text seems to be a maverick: in only one case does it show a significant subject area effect (.01 with Nitro, the Highly Specific text) and there are no significant effects between it and the BSS passages. This is particularly surprising as, on the whole, the chart shows strong background effects between BSS and LMS texts, and somewhat lower ones between LMS and PST ones. However, as I pointed out earlier, the background effect is caused by the interaction of two texts, one of which may be more specific than the other. What happens, for example, to the level of significance if one subtest is specific and the other general? To answer this we need something more trustworthy than my intuitive views to show which texts are specific and which general. Chapter 8 will look at students' and academic lecturers' views on the appropriacy of the reading passages, and Chapter 9 will consider the content of the reading passages and the test questions.

Table 7.8

Level of Significance of Subject Area Effect for Pairs of Subtests (Repeated Measures Analysis of Variance)

7.8.1		LMS Module				PST Module		
		Tooth	Child	Genes	Nitro	Sun	Fuel	Ship
BSS Module	Qual	NS	NS	.01	.001	<.001	NS	NS
	Educ	NS	NS	<.001	<.001	<.001	NS	NS
	High	NS	NS	<.001	<.001	<.001	NS	NS
		PST and LMS students				PST and BSS students		

7.8.2		PST Module			BSS Module		
		Sun	Fuel	Ship	Qual	Educ	High
LMS Module	Tooth	NS	.05	NS	NS	NS	NS
	Child	NS	NS	NS	NS	NS	NS
	Genes	NS	.05	NS	.01	<.001	<.001
	Nitro	.01	<.001	.01	.001	<.001	<.001
		LMS and PST students			LMS and BSS students		

7.8.3		LMS Module				BSS Module		
		Tooth	Child	Genes	Nitro	Qual	Educ	High
PST Module	Sun	NS	NS	NS	.01	<.001	<.001	<.001
	Fuel	.05	NS	.05	<.001	NS	NS	NS
	Ship	NS	NS	NS	.01	NS	NS	NS
		PST and LMS students				PST and BSS students		

Before we leave this chapter, however, there are two other areas I should like to explore. The first relates to test bias, and the second to the effect of academic subject area on test completion.

Section 3: Item Bias

Tests of Item Bias

So far in this study we have looked at the performance of the complete reading modules and their subtests, but not at the performance of the individual test items. I have therefore supplemented my use of analysis of variance with two of the many statistical methods that are used to identify item bias. Such methods are mainly used to check that test items are not biased against subgroups of candidates such as those with particular linguistic or ethnic backgrounds, but some researchers have used them to see how test performance varies according to other criteria. For example, Tatsuoka *et al.* (1988) used measures of test bias to compare the mathematical performance of three groups of students following different types of instruction, and O'Neill, Steffen and Broch (1994) used test bias techniques to assess the effect of field of study on reading comprehension. Since the assessment of test bias focuses on items, it seems appropriate to use it to compare the performance of the BSS, LMS and PST students on the items in the three reading modules. If in any one module there are several items which seem to be biased towards students in the appropriate subject area, then those items may be considered to contribute to the subject specificity of that module.

Tatsuoka *et al.* (1988) recommend the Mantel-Haenszel statistical technique which is based on chi squared. This compares the odds of two different groups answering an item correctly when differences in ability between the two groups are taken into account. Educational Testing Service (ETS) researchers also use the Mantel-Haenszel technique in their studies of bias, but they standardise the results using the Mantel-Haenszel Delta Difference (MH D-DIF) statistic to produce their own index of Differential Item Difficulty (DIF) (see O'Neill, Steffen and Broch 1994).

Another, simpler, way of identifying item bias is to compare the performance of the different focal groups by carrying out separate classical item analysis for each group, and by comparing an item's facility values and discrimination indices across the groups. The chief disadvantage with this method is that the item statistics from classical methods of analysis vary according to the level and spread of ability of the students in a sample. It is impossible, therefore, to know how much the variations in the item statistics are due simply to sampling differences and how much, if at all, to item bias.

If, however, the items are compared using Item Response Theory (IRT) analysis, this problem should be avoided because the item statistics are claimed to be sample free, that is, independent of the distributional characteristics of the sample (see Wright and Masters 1982). As Miller and Linn (1988:205) put it, 'The item parameters, and therefore the item characteristic functions, are assumed to be invariant across subpopulations.' This being the case, it is possible to use IRT to compare the item statistics for different groups of students, and to presume that any differences in the resulting statistics are not due to variations in the ability levels and range of the students in the different samples. Many of the most popular methods of assessing item bias are based on IRT, and I have used two closely related IRT methods in my study.

Possibly the most usual method of using IRT to check for item bias is to use the three parameter model, and to compare the item characteristic curves (ICC) for each item. If the test is given to two different samples, and if an item's ICC from each sample differs according to the item's discrimination and difficulty level, then 'the presence of bias can be inferred' (Adams and Rowe 1988). Adams and Rowe claim that the three parameter model is generally preferred for comparisons of this kind because it most closely describes the most common kinds of test and test data. However, three parameter models require sample sizes of at least 1,000 subjects, and are therefore beyond the scope of this study. Since the majority of the IELTS reading module items are open ended, and cannot be answered by guessing, a two parameter model taking item discrimination into account might be equally suitable, but again such models require at least 200 cases per sample. One parameter model ICCs, which only require a minimum of 100 cases, can be compared in the same way, but are not so informative since they only take item difficulty/person ability into account. (See Lord 1980, and Hambleton, Swaminathan and Rogers 1991, but see also see Andrich 1988:66, who shows how the shape of a Rasch model ICC curve gives some idea of the items' discrimination.)

Rasch Analysis

A simple way of using the Rasch model to check for item bias is to use the Fischer-Scheiblechner statistic, which was introduced in 1970 (see van den Wollenberg 1982) and is described in Wright and Masters (1982:115). Wright and Masters recommend the use of this statistic for checking whether test item measures are invariate over different trials. They suggest that items should be tested on two or more samples of varying levels of ability, and that each item's delta statistic, that is, the item's difficulty measurement, centred to remove the effect of the sample means (see Wright and Masters, 1982:63), should be compared. If more than 5 per cent of the items show significant differences in their delta values, then the item calibrations are not showing the required stability, and the offending items should be reconsidered.

The difference between an item's delta scores from different administrations can be converted into a z score, using (if the test is tried out on two samples) the following formula:

$$z = (d_1 - d_2)/(s_1^2 + s_2^2)^{\frac{1}{2}}$$

d_1 = *the item delta in the first trial*
d_2 = *the item delta in the second trial*
s_1 = *the* d_1 *standard error*
s_2 = *the* d_2 *standard error*

(Wright and Masters 1982:115)

The resulting z score is a standard score with an expectation of zero and a variance of one. Each z score can be compared to areas of the normal curve in the usual way, using a table of standard scores. If a z score is more extreme than ±1.96, then there is a significant difference in the performance of the two samples on this item (see Cushing Weigle 1994).

This method of assessing item bias seems to be appropriate for my purposes, although the limitations of the Rasch model outlined above need to be held in mind. However, before running the analysis it is important to see how closely the data set meets the underlying assumptions of the Rasch model, which are that the data is unidimensional, that each item is independent, that all the items discriminate at approximately the same level (see Hambleton, Swaminathan and Rogers 1991:56), and that candidates have enough time to finish the test.

A unidimensional test is one that tests a single underlying trait. This seems an almost impossible requirement for a language test since even a discrete point item which is supposed to assess, say, one grammatical feature, will be testing more than just grammar. Indeed, as Reckase, Ackerman and Carlson (1988) point out, a relatively simple construct like vocabulary ability can be described as multidimensional if analysed in enough detail. However, Reckase, Ackerman and Carlson show that many studies have demonstrated how robust the IRT procedures are in the face of violations of this assumption, and they say that the unidimensional assumption requires only that the items in a test measure the same composite of abilities, rather than one single ability.

The three IELTS modules, M(BSS), M(LMS) and M(PST), are all intended to test the composite ability of reading comprehension, and an initial Rasch item analysis using the Quest program (Adams and Khoo 1993) reveals that in all the six analyses of the data (two groups of students by three modules) there is only one item which has an infit value of over 1.3. It can be taken, therefore, that for the purposes of Rasch analysis the data is unidimensional. (For more about the assumptions of unidimensionality and independence see Baker 1995.)

Classical and IRT test analyses are both based on the assumption of independence – that is, that each test item stands alone and is not affected by the candidate's answers to other items. However, many tests, for example cloze tests, contain items which are not independent of preceding or succeeding items, but their performance is routinely, and usefully, assessed using classical item analysis.

In the case of the three IELTS reading modules each item is designed to be independent, but some, particularly those in the summary completion gap filling tasks, are probably not. Although the Rasch model may be as robust as classical methods, this possible lack of independence should be held in mind when the results of the analyses are considered.

If the assumption that items have similar levels of discrimination was strictly adhered to it would rule out the use of Rasch analysis with most norm-referenced language tests, since such tests generally contain items with a wide range of discrimination indices.

The final assumption – that students are given ample time to finish a test – gives more cause for concern. If a test constructor wants to study the difficulty levels of different items, it is clear that students should have time to attempt all items; if they do not we cannot know how difficult the unattempted items are. In addition, some of the items near the end of a speeded test may have been answered in such a hurry that their true levels of difficulty are concealed. However, if a test is specifically designed so that students of lower levels of ability may not have time to finish, as is the case with tests such as the IELTS Reading modules, it is difficult to know how to achieve an accurate estimate of an item's difficulty without changing the purpose of the test. Since it would be unreasonable to give students as much time as they wished to answer these modules, the best we can do is to omit students' unanswered items from the Rasch analysis, and to hold the problem in mind when considering the results. The effect of the tests' speededness will be considered again briefly below.

The Results

Since a Rasch analysis requires a minimum of 100 students in a sample, there are not enough LMS students in my data set for their performance on each of the three modules to be analysed separately. LMS and PST students, therefore, who both have a background in the physical sciences, have been reallocated to one group, Science, and their combined performance is compared with that of the BSS students on each item in each of the three modules. Of course this aggregation may conceal differences in performances on the LMS module, because although all LMS students have studied physical science, some of the PST students have not studied life sciences. Some items in the LMS module, therefore, may fail to show significant differences which might have emerged if the three groups' results had been analysed separately.

For each item the delta statistic (see Wright and Masters, 1982:62) and its accompanying error of measurement were computed using the Quest program. For the two groups' deltas, errors and z scores for the three reading modules, see Clapham (1994).

As can be seen, very few z scores are greater than ±1.96. In the case of the PST module, for example, only three – Items 1, 4 and 27 – show significant differences. Of these three, Items 1 and 4 proved easier for the scientists, and Item 27 for the BSS students. An initial inspection of the items does not reveal any striking reasons why they should stand out in this way, and since the module has 40 items, it is probable that these three differences are due to chance. The fact that overall the items are invariant over the two samples may serve to confirm the earlier findings that a great part of the PST test is not subject specific.

The results of the BSS module are similar. Only two items – 14 and 21 – show significant differences, and of these, one was easier for the BSS and one for the Science students. However, in this case, the invariance cannot be ascribed to the non-specificity of any subtests (all three subtests showed evidence of being subject specific). It may, therefore, be the case that Science students are not disadvantaged when they read social science texts because such texts are familiar in their everyday reading. If this is the case it would support O'Neill, Steffen and

Broch (1994), who found that scientists were not disadvantaged if they read social science texts, but that social science and humanities students were disadvantaged if they read scientific texts.

In the case of the LMS module, there are five items which show significant differences, and although two of these – Items 15 and 28 – were easier for the BSS students, the other three – Items 30, 38 and 39 – were easier for Science students. These three all come from the two final LMS subtests which were shown in Section 7.2. to be the most highly specific subtests of the three modules. It looks as if these items, therefore, may be genuinely biased against BSS students.

The above method of identifying item bias, in which items are looked at separately, and are only considered if they show significant differences, may obscure overall differences among groups of items, and indeed Tatsuoka *et al.* (1988) say that:

> *item-by item comparisons are rather simple minded and statistically questionable when the number of items is as large as 40*
> (1988:306)

I have therefore also looked for bias in groups of items, by using a variant of the Fischer-Scheiblechner method in which items are grouped and each z score in the group of items is squared and summed to produce a chi squared statistic (see Andrich 1988).

Since some of the subtests in my study include two sets of questions containing different item types (see, for example, (PST)Sun in Appendix 6.1), and since it is possible that students in different subject areas might be disadvantaged by different kinds of test item, the chi squared statistics were produced for subtests and, where appropriate, their component parts. Table 7.9 gives the results of the chi squared tests.

It will be seen from this table that most of the differences between the two groups on the various subtests are not significant, and do not, therefore, show evidence of any consistent bias against one of the two groups. Even where the results are significant, the subtests are not necessarily biased against one group. For example, the differences between Items 28–32 in the LMS Module are significant at the .01 level. However, if the item differences are looked at individually (see Clapham 1994), it can be seen that two of the items are easier for the Science students, and three for the BSS. So although this group of items is behaving differently for the two groups, it is not doing so consistently.

There is also a BSS set of items, Items 13–19, which shows a significant difference. In this case all seven items are easier for one group, but, surprisingly, this is not the BSS but the Science group. Since this subtest as a whole, (BSS)Educ, has already been shown to be subject specific, it looks as if the item type in this subsection, which involves identifying the 'aims' in a passage and matching them to a list of such aims, may be more suited to Science than to BSS students, and is therefore biased against BSS students. These items will be discussed further in Chapter 9.

Table 7.9

Chi Squared Tests of Bias

	Items	Chi squared	df	P
	1–3	8.81	2	.05
	4–11	7.74	7	NS
	1–11	16.55	10	NS
M(PST)	12–14	1.26	2	NS
	15–18	.25	3	NS
	12–18	1.51	6	NS
	19–27	13.12	8	NS
	28–40	6.16	12	NS
	1–18	18.92	17	NS
	19–23	3.72	4	NS
	24–27	1.54	3	NS
M(LMS)	28–32	16.33	4	.01
	33–36	4.36	3	NS
	28–36	20.69	8	.05
	37–40	18.60	3	.001
	1–12	9.54	11	NS
	13–19	17.73	6	.01
	20–27	10.30	7	NS
M(BSS)	13–27	28.03	14	.05
	28–31	4.15	4	NS
	32–35	8.30	3	.05
	28–35	12.45	7	NS

The only two groups of items which show evidence of bias in the expected direction are the last two sections of the LMS module, which include the items with significant delta differences that I discussed above. Here the differences are significant at .05 and .001 respectively, and in each case it was the Science students who found the items easier. These items are based on the (LMS)Genes and (LMS)Nitro texts which were earlier shown to be highly specific to students in the Life and Medical Sciences, and the bias results therefore reinforce these earlier findings. However, it may not simply be the subject specificity of the texts which has caused the difference, because the section containing Items 33–36, which forms the second part of the (LMS)Genes subtest, shows no significant difference between BSS and Science students. It may be that there is also an item type effect, or it may be that the aggregation of the LMS and PST students has obscured other possibly significant results.

What this present study shows is that, on the whole, students in the two major subject areas are not disadvantaged by the test items, except in the case of one M(BSS) section which appears to disadvantage students in its own subject area, and two M(LMS) sections which disadvantage BSS students. The expectation that the items would be biased against students who were not in the appropriate subject area has not been fulfilled. It is only when the subtests appear to be highly specific that the items show consistent evidence of bias against students in the other subject area.

Section 4: Unanswered Items

Since between 30% and 50% of the students failed to finish the reading modules, and since it may be the case that students read more quickly in their own fields of study than in other fields, I shall now look briefly at the distribution of the unanswered items. This is a complex area which requires a study of its own, but here I shall simply look at the uncompleted papers to see whether students finished more items in the test within their own subject module than in the one outside it.

The number of blank answers at the end of each student's two reading tests were counted, and the results were classified under three headings: a) test completed, b) 1–4 items unanswered (for the PST module it was 1–6 items) and c) more than 4 items unanswered. In the case of the BSS and LMS modules the last four items comprised a complete subtest section, and seemed to provide a natural break for students who did not have time to finish the test. In the case of the PST module the six items formed half the last section, and again proved to be a frequent stopping point for students who had run out of time. Unanswered items earlier in the students' answer sheets were ignored as it was impossible to know whether they were unanswered because of time constraints or because the items were too difficult.

The resulting frequencies in the three subject areas were entered into contingency tables, and chi squared statistics were calculated. Table 7.10 shows these tables and the chi squared significance levels.

It will be seen from these figures that for the BSS module there was no significant effect; students in all three subject areas had the same number of unfinished items. However, for the LMS and PST modules, the results were significant at the .01 level. It looks at first glance, therefore, as if students in all three subject areas had an equal chance of completing the BSS paper, but not the two science papers. However, the two significant results have different causes. The LMS results are in the expected direction. The boxes with the greatest differences between the observed and the expected frequencies are 1,3, 4 and 6, with the BSS students completing fewer of the LMS test items than would have been the case if the groups had been randomly selected. These students also had a greater than expected number of papers with over four unanswered items. At the same time the LMS students completed more papers than expected, and left fewer items unanswered. (The PST results were insignificant.) In this case, therefore, our expectations are fulfilled. The students in the appropriate subject area (LMS) seemed to be able to read the test and answer the questions more quickly than the PST students who in their turn completed more of the paper than did the BSS students.

However if we look at the PST contingency table a different picture emerges. Here the major differences between the obtained and the expected figures appear in boxes 1, 3, 7 and 9. The BSS students finished *more* papers than expected, whereas the PST students finished *fewer*. For this module, therefore, it is not the case that students in the appropriate subject area finished the test more quickly. The reasons for this surprising result are probably complex, but it seems likely that they partly stem from the fact that the PST Module is the least subject specific of the three modules, and that the text on which all the last 22 items in the test are based is the (PST)Ship passage which is in a subject area which was unfamiliar to almost all the PST students (See Chapter 8). However, if the PST module was simply non-subject specific, the results of the chi squared test might be expected to be insignificant. The significant result may be due instead to the fact that 36% (142/395) of all the students failed to answer more than six of the last questions in the test, whereas for the LMS and BSS modules the numbers were 12% and 13% respectively. Items 19–27, the first section of the (PST)Ship subtest, consists of a very demanding, and time-consuming, set of items in which students have to identify the stages in a process. It is possible that the BSS students managed to answer this section more quickly than the PST students, and therefore had more time to attempt the final test items. A suspicion that the results might be due to higher proficiency on the part of the BSS students is not supported since not only is there little firm evidence that the BSS students are more proficient at reading comprehension, but there is evidence that the LMS students are (their mean scores on all three modules are higher than those of either the BSS or the PST students, see Tables 7.1 and 7.2). The repeated measures analysis of variance reported in Table 7.2 show that those LMS students who took both the LMS and the PST modules had significantly higher scores than the corresponding PST students. However, in spite of this difference in ability, the LMS results in the PST contingency table are insignificant.

Table 7.10

Uncompleted Test Papers

BSS Module

	Finished	1–4 blanks	>4 blanks	Total	Mean%
BSS	[1]E= 220.97* 223	[2]E= 51.04 50	[3]E= 42.99 42	315	50.8
LMS	[4]E= 63.84 50	[5]E= 14.75 9	[6]E= 12.41 4	63	51.4
PST	[7]E= 44.19 56	[8]E= 10.21 17	[9]E= 8.60 18	91	42.7
Total	329	76	64	469	

Chi squared = 7.25 Df = 4 Not significant
The BSS and LMS means are both significantly higher than the PST mean (p = <.05)
(ANOVA plus Scheffe's Test for differences between pairs)

LMS Module

	Finished	1–4 blanks	>4 blanks	Total	Mean%
BSS	[1]E= 109.63 98	[2]E= 38.57 42	[3]E= 20.8 29	169	46.7
LMS	[4]E= 81.74 96	[5]E= 28.75 23	[6]E= 15.51 7	126	55.5
PST	[7]E= 61.63 59	[8]E= 21.68 24	[9]E= 11.69 2	95	41.2
Total	253	89	48	390	

Chi squared = 13.45 Df = 4 Significant at .01

The LMS mean is significantly higher than both the BSS and the PST means (p = <.01)

PST Module

	Finished	1–6 blanks	>6 blanks	Total	Mean%
BSS	[1]E= 76.39 92	[2]E= 28.65 27	[3]E= 58.96 45	164	43.8
LMS	[4]E= 34.01 35	[5]E= 12.75 13	[6]E= 26.24 25	73	53.4
PST	[7]E= 73.6 57	[8]E= 27.6 29	[9]E= 56.8 72	158	45.8
Total	185	69	142	395	

Chi squared = 14.56 Df = 4 Significant at .01
The LMS mean is significantly higher than both the BSS and the PST means (p = <.05)
*E = Expected Frequency

To summarise the results from this analysis of uncompleted papers: the results for the LMS test, which was the module containing the two most highly subject-specific subtests, support the original repeated measures analysis of variance finding that students perform better at reading tests within their own discipline, than at ones outside it. However, the BSS results do not. They imply that LMS and PST students can read social science texts and answer the related questions as quickly as BSS students. The PST results are unclear.

Conclusion

The repeated measures analysis of variance studies have shown that the students tended to perform better on the reading module within their own subject area than on the one outside it. However, when performance on the individual subtests was compared, the students sometimes performed better on the subtest in their own subject area, and sometimes did not: the subtests seemed to vary in their specificity from ones which were highly specific to ones which were equally appropriate for students in all three subject areas. The two subtests which appeared to be most highly subject specific were (LMS)Genes, and (LMS)Nitro, and this was confirmed by the tests of bias which showed that some of their items were biased against BSS students. However, the other bias results seem to *disconfirm* the earlier findings. Except in the case of the two LMS subtests, the bias studies show no evidence of consistent bias against students outside the appropriate subject area. In the study into uncompleted papers, the results on the LMS Modules showed that students in the appropriate module completed more of the paper than did students outside it, but on the BSS module LMS and PST students left no more uncompleted answers than did the BSS students.

I am now in a position to add to my initial intuitive classification of the subtests into 'specific' and 'general', which was reported in Table 7.6. Table 7.11 includes the results of the repeated measures analysis of variance studies and also the evidence from the investigation into bias. Under each heading, subtests which appear to be specific are labelled 'S' and those which are not, 'G'. In the 'specificity' column, passages are labelled according to the majority of 'S's and 'G's in the other columns.

Table 7.11

Specificity Table 2

Subtest	Intuition	Repeated Measures An. of Var.	Bias.	Specificity
Qual	S	S	G	S
Educ	S	S	G	S
High	S	S	G	S
Tooth	G	G	G	G
Child	G	G	G	G
Genes	S	S	S	S
Nitro	S	S	S	S
Sun	S	S	G	S
Fuel	G	G	G	G
Ship	S	G	G	G

Since it is now clear that the subtests vary in their specificity, we need to find out why. In Chapter 8, therefore, there will be a description of the students' reported familiarity with the subject area and topics of the texts, and in Chapter 9 there will be an investigation into whether the source of the reading passages affected the specificity of the passages.

8 Reasons for the Variation in Subtest Specificity

In Chapter 7, I showed how the IELTS Reading subtests vary in their specificity. In this chapter I try to identify some of the causes of this variation. I shall look first at the comments by students on the familiarity of the subject matter and topic of each reading passage, and then at the comments by academic subject specialists on the texts' appropriacy for their students. In Chapter 9, I shall look at the source of each reading passage to see whether there is a simple relationship between this and text specificity, and shall then approach the problem another way by using Bachman's Test Method Characteristics (TMC) Instrument (see Bachman 1990) to analyse further aspects of the reading passages and their accompanying items.

Familiarity of Subject Area and Topic

Familiarity of Subject Area

In the questionnaire, students were asked whether they were familiar with the subject area of each reading passage they encountered. For example one question was as follows:

> *Were you familiar with this general area of physics before you read the passage? (Question relating to (PST)Fuel; see Appendix 6.4)*

Students had to say whether the text they had just read was 'Very Familiar', 'Familiar' or 'Not Familiar'. They were asked a similar question about each of the passages, whether the module was in their own discipline or not. Few students said they were *very* familiar with the subject area of any of the reading passages, so for this study the two 'familiar' categories have been conflated. (A complete list of the percentages of responses in each subject area is given in Appendix 8.1.)

On the whole the questionnaire answers tallied with my own guesses as to the specificity of each reading passage (see Table 7.6). For example, for the BSS reading passage, Educ, which I had identified as specific, the answers were as follows:

(BSS)Educ

BSS students:	*Familiar: 64%*	*Unfamiliar: 36%*
LMS students:	*Familiar: 39%*	*Unfamiliar: 61%*
PST students:	*Familiar: 47%*	*Unfamiliar: 53%*

64% of the BSS students said they were familiar with the subject area of the passage, as compared to 39% of the LMS students and 47% of the PST group. Differences of this kind between the answers of the three groups of students were particularly marked for (LMS)Nitro, which the Level of Significance Chart (Table 7.8) had identified as being highly subject specific:

(LMS)Nitro

BSS students:	*Familiar: 9%*	*Unfamiliar: 91%*
LMS students:	*Familiar: 64%*	*Unfamiliar: 36%*
PST students:	*Familiar: 41%*	*Unfamiliar: 59%*

However, not all the results were as expected. (PST)Sun, for example, which I had labelled specific, appeared to be slightly more familiar to LMS students than PST ones:

(PST)Sun

BSS students:	*Familiar: 33%*	*Unfamiliar: 67%*
LMS students:	*Familiar: 77%*	*Unfamiliar: 23%*
PST students:	*Familiar: 69%*	*Unfamiliar: 31%*

and (PST)Ship appeared to be unfamiliar to almost everyone:

(PST)Ship

BSS students:	*Familiar: 4%*	*Unfamiliar: 96%*
LMS students:	*Familiar: 11%*	*Unfamiliar: 89%*
PST students:	*Familiar: 10%*	*Unfamiliar: 90%*

In both cases these results can probably be explained. The Sun passage deals with the effect of ultraviolet waves on living forms, and might therefore be equally at home in an LMS textbook. On the other hand (PST)Ship may be so subject specific that it is unfamiliar to students in all three subject areas. Or possibly its linguistic style poses a problem. It contains a description from an engineering journal on the raising of a Tudor warship from the sea bed and the text is unlike any of the others in the three modules. It is in narrative style, giving a step-by-step account of the stages and dangers of the salvage process, and includes some highly technical terms. It also, as we saw in Chapter 6, includes a surprisingly large number of passive verbs. It seems likely that its approach is familiar to only a small group of engineers and that it was unsuitable for most of the PST students who took this module. This bears out a discovery we made during the content validation of the pilot IELTS reading modules. The pilot PST reading test contained two engineering texts which were in many ways similar in style to

(PST)Ship. The university lecturers in the physical sciences and technology (with the exception of the engineers) said that these texts were not appropriate for their students. A clause was therefore inserted into the test specifications to say that no more than one text of this kind should be included in future PST reading modules. It seems now as if even one such text is inappropriate; our attempt not to disadvantage engineering students has led to an unsuitable text for almost everyone. It is probable that it is because the Ship text is so unfamiliar that it shows few significant subject area effects, and none with the BSS subtests. Presumably this would be an ideal text for the Joint Matriculation Board University Test in English for Speakers of Other Languages (see McEldowney 1976), where the reading and listening passages are intentionally designed to be on topics which are equally unfamiliar to all candidates.

My intuition about (PST)Fuel, which I had classified as being general, was supported in that it turned out to be familiar to more than half of each group of students. However, it did become progressively more familiar from BSS to LMS to PST, which perhaps implies some degree of specificity:

(PST)Fuel
BSS students: Familiar: 60% Unfamiliar: 40%
LMS students: Familiar: 72% Unfamiliar: 28%
PST students: Familiar: 92% Unfamiliar: 8%

Table 8.1 gives the mean familiarity rating for each complete module. For this, each student's responses to the three reading passages were summed, divided by 3 and rounded to the nearest whole number, before being converted into percentages.

Table 8.1

Subject Area Familiarity – Whole Tests

		BSS Students	LMS Students	PST Students
M(BSS)	Familiar	57%	29%	34%
	Unfamiliar	43%	71%	66%
M(LMS)	Familiar	9%	68%	19%
	Unfamiliar	91%	32%	81%
M(PST)	Familiar	34%	61%	60%
	Unfamiliar	66%	39%	40%

It will be seen that for the BSS and LMS modules, students were consistently more familiar with the subject areas of the passages in their own academic disciplines than were students outside those disciplines. However, this was not the case with the PST module: more than half the PST *and* LMS students found the subject areas familiar, and the proportions were almost identical (60% and

61% respectively). This is partly caused by the fact that a higher proportion of LMS than PST students said they were familiar with the subject area of (PST)Sun, but it also reinforces my suggestion that the LMS students might be expected to do as well as the PST students on the PST module since all LMS students have to study the physical as well as the life sciences. PST students, on the other hand, are not familiar with LMS subject areas: only 19% were familiar with the subject areas of the LMS texts.

Topic Familiarity

The students were asked a similar question about the topic familiarity of each passage:

> *Were you familiar with this particular topic, i.e. the effects of the sun's rays? (Question relating to (PST)Sun, see Appendix 6.4)*

The questionnaire explained what was meant by the word 'topic', so that students could appreciate the difference between that and 'subject area'. Although the distinction between the two terms is perhaps fine, no students asked for the terms to be clarified, and their responses show a certain consistency which suggests that most of them did have an idea of the difference between the two. On the whole the responses were similar to those relating to subject area, though in most cases students were, not surprisingly, less familiar with the topic than they were with the general area (see Appendix 8.2). For example, although 50% of the BSS students said they were familiar with the general area of business studies, only 36% said they were familiar with the particular topic, i.e. Quality Circles. However the BSS students' responses to the PST reading passages were intriguing. They found the topics of all three texts more familiar than the general subject areas. Table 8.2 lists the percentage of BSS students who found the subject area and the topic of each reading passage familiar.

Table 8.2

BSS Student Familiarity with PST Subject Areas and Topics

	Sun	Fuel	Ship
Subject area	33%	55%	4%
Topic	49%	71%	6%

The differences in the familiarity of the subject area and topics of the Sun and the Fuel texts are striking. A markedly higher proportion of BSS students were familiar with the topics than with the general subject areas. This is presumably because the effects of the sun's rays and the provenance of alternative fuel resources have recently become areas of general interest. When the passages were originally selected for the PST module they were considered to be about interesting but little known topics. Unfortunately the topics quickly sprang to general prominence, and it is probable that the specificity of the PST module has suffered in consequence.

For most reading passages, the difference between the proportions of 'familiar' and 'unfamiliar' responses were in the same direction for both questions even if the actual proportions varied. So, for example, more than half the PST students were familiar with both the subject area and the topic of the (PST)Sun text (69% and 70% respectively). However, in the case of all three of the LMS passages, the proportions of LMS students who were familiar with the subject area and the topic were reversed. Table 8.3 lists the percentages of LMS students who were familiar and unfamiliar with the subject areas and topics of the three passages.

In all three cases, a higher proportion of students were familiar with the subject areas, and a smaller proportion were familiar with the actual topics. This is presumably as it should be for a test of this sort and so it is perhaps surprising that of the other 26 pairs of responses only seven show a reversal of this sort (see Appendix 8.2).

Table 8.3

LMS Students' Responses to LMS Passages

		Teeth*	Genes	Nitro
Subject area	Familiar	59%	66%	64%
	Unfamiliar	41%	34%	36%
Topic	Familiar	43%	37%	45%
	Unfamiliar	57%	63%	55%

* Teeth covers both the Tooth and the Child passages which are on the same topic – tooth decay.

Crosstabulations

In order to see how students' familiarity with the subject areas and topics of the texts accorded with their scores on each subtest, the scores were crosstabulated against the two categories in each question, 'familiar' and 'unfamiliar'. For ease of interpretation, scores were divided into two groups, those above the subtest median and those below. Scores on the median were omitted. Chi squared tests were carried out on the ensuing two by two contingency tables and the results are listed in Table 8.4.

The chi squared significance figures may not be a reliable guide to the relationship between subject-area and topic familiarity and test score, since the cut-off point of above and below the median was somewhat arbitrary. However, taking this into account, one interesting fact emerges. With the exception of the

LMS texts, Tooth and Child (here conflated into 'Teeth'), which I had classed as being general texts, and (PST)Ship, which seems in some ways to behave like a BSS text (there were, for example, no significant subject effects when it was paired with the BSS subtests [see Table 7.8]), the M(BSS) chi squared results are mostly not significant whereas those of the two Science Modules are. For M(BSS), there was no correspondence between familiarity with the subject area of the BSS texts and the test scores. This supports my suggestion in Chapter 7 that science students can cope with BSS texts (if they are not too highly specific), whereas BSS students need some knowledge of the subject before they can understand science texts. If this turns out to be the case it would provide a strong argument for IELTS dropping the three academic modules, and replacing them with one BSS-type one.

Table 8.4

Chi Squared Tests – Levels of Significance (P Values)

	Chi Squared	
Text	**Subj. Area**	**Topic**
Qual	NS	NS
Educ	NS	NS
High	NS	.001
Teeth	NS	NS
Genes	.0002	.03
Nitro	.01	.004
Sun	<.0001	.0001
Fuel	.0001	.00002
Ship	NS	NS

* Teeth is a conflation of the Tooth and the Child subtests, which are both on the same topic – tooth decay.

Subject Specialists' Views

When the prototype IELTS test was constructed, academic subject specialists were asked whether the sample reading passages and test tasks were appropriate for their undergraduate and postgraduate students. Their comments led to modifications of the draft tests and test specifications (see Chapter 4). This interviewing was very time consuming and depended heavily on university lecturers' goodwill, so it could not be carried out every time a new set of test modules was constructed. It was hoped that the test specifications would be sufficiently detailed to ensure that future passages and test items would be appropriate for their candidates. Since it seems that this may not be completely so in the case of the versions under discussion, a small number of university lecturers at Lancaster University were asked to read all ten passages and to say how appropriate the texts were for their students, and how familiar these students would be with the subject matter. The questions were designed to be as quick and easy to answer as possible, so lecturers were asked to rate the passages according to a three-point scale – three-point since that is the scale used in Bachman's Test Method Characteristics (TMC) Rating Instrument and since the lecturers were also asked four questions relating to this instrument. The responses to these four questions will be discussed in Chapter 9. (Appendix 8.3 contains the questions to the subject specialists.)

Nineteen lecturers filled in the questionnaire – seven from the BSS subject area, seven from LMS and five from PST. The two main subgroups for which each test was intended were all represented, so there were lecturers in business studies, the social sciences, medicine, biology, the physical sciences and engineering. Since each lecturer came from a specialisation within a discipline which was itself more restricted than the subject areas covered by BSS, LMS and PST, their comments naturally varied considerably. In addition, some lecturers had difficulty choosing which students to relate their answers to, since, following the orientation of the IELTS test battery, the questionnaire asked them to relate their answers to both undergraduate and postgraduate students. Some lecturers, it turned out, chose to answer with just undergraduates in mind and some chose postgraduates. The resulting answers, therefore, cannot be considered to give anything more than a general idea of the appropriacy and familiarity of the passages.

Appropriacy of Text

I shall not describe the subject specialists' answers in any detail, but shall highlight those responses that might contribute to our understanding of the specificity of the texts. One finding was that (PST)Sun, which showed no significant subject area effect when paired with the LMS subtests (see Table 7.8), and where the subject area appeared to be more familiar to LMS than PST students (see Appendix 8.1), was considered to be just as suitable for LMS as for PST students. It is based on a description of ultraviolet rays and their effect on living things, and clearly straddles the two subject areas. Another finding is that (LMS)Child, one of the 'general' passages, was considered highly appropriate for doctors but not for biologists. This lack of appropriacy is likely to account for the reduced subject specificity of that subtest. On the other hand, (LMS)Genes and (LMS)Nitro, the two 'highly specific' subtests, were considered highly appropriate by the LMS lecturers, and not at all appropriate by those in BSS. They were also considered highly inappropriate by the physics lecturer and the two engineers, but one of the two chemists thought they were fairly appropriate, and the other thought they were highly appropriate. This emphasises the disparate nature of the lecturers in the PST subject area, and also the indeterminacy of the boundary between LMS and PST. One intriguing discovery, which is unlikely to have been anticipated by the item writers, is that the two engineering lecturers both thought that (BSS)Qual, which is a business studies report about producing high quality goods in factories, was highly appropriate for their students, and one of the engineers said that it was more appropriate than the engineering text, (PST)Ship.

Familiarity with Topic

Since both subject specialists and the students commented on the familiarity of the reading passage topics it is possible to compare the two sets of answers. To make this easier, the subject specialists' answers have been conflated into two categories – Familiar and Not Familiar – so that they match the students' responses. Appendix 8.4 gives the percentages saying 'familiar' and 'not familiar'. These percentages cannot be directly compared to those of the students, since there were no more than seven subject specialists in any one of the three subject areas, and since the respondents by no means represent all the subject areas from which the students came. However, an initial comparison of the proportions serves to identify gross differences between the views of the two groups.

On the whole the proportions of respondents saying that a topic would be familiar or unfamiliar to their students were similar to the students' proportions. When there were marked differences, it was always the case that the lecturers overestimated the degree to which students in their field would be familiar with a particular topic. For example, all the seven BSS lecturers thought that the (BSS)Qual topic would be familiar to their students, whereas only 36% of the BSS students said it was. Similarly, all the LMS lecturers thought that (LMS)Nitro would be familiar to their students, whereas only 45% of the students claimed that it was. I have said that this was an overestimation, but it may be the case that the lecturers accurately estimated the knowledge of students within their own discipline, and that the students who said they were not familiar with the topic came from fields which were not represented by lecturers on the survey. However it is also possible that these overestimations were partly due to the fact that the lecturers had a wider knowledge of the world than the students, particularly than the undergraduates, and secondly that because they probably had British students in mind, they discounted the effect of home background. For example, the LMS and PST lecturers heavily overestimated the topic familiarity of the (LMS)Teeth topic for their students. Over 80% of both the LMS and the PST lecturers thought that the topic of water fluoridation would be familiar to their students. However, only 43% of the LMS students, and 26% of the PST students claimed to be familiar with this topic. It is probable that because the controversy as to whether the water supply should be fluoridated or not is frequently discussed in British newspapers, the lecturers, who were all British, would have expected the topic of fluoride and tooth decay to be generally familiar, but it is likely that many students from abroad would be unfamiliar with the subject. The discrepancy between the lecturers' and the students' responses may therefore be due to misapprehensions about students' general knowledge, rather than any disagreements about the three main subject areas.

Of course the fact that a topic is not familiar to students in the appropriate subject area does not mean that that text is unsuitable. Indeed probably the most suitable texts would be ones which were firmly based in the relevant subject area, so that the rhetorical style and conceptual framework were familiar, but which related to unfamiliar topics so that there was no possibility of the students answering the questions from background knowledge alone. The fact, therefore, that of the seven BSS subject specialists only the lecturers in Education and Politics thought that (BSS)High would be a familiar topic to their students does not imply that the reading passage is not specific, especially as five out of the seven thought that the subject area was appropriate. However, there is cause for concern when the topic of a reading passage which is supposed to be *outside* the students' subject area is familiar to them. All but one of the BSS subject specialists, and similar proportions of the LMS and PST subject specialists, thought that the topic of (PST)Fuel would be familiar to their students. This agrees with the fact that equal proportions of students in the three subject areas thought the topic was familiar, and further reinforces my belief that this text is not subject specific.

The fact that the topic of a text is unfamiliar to students outside the relevant subject area does not necessarily confirm the subject specificity of that reading passage, but it is interesting that all the BSS lecturers said that (LMS)Genes, and (LMS)Nitro, which are the most subject specific of the Reading subtests, would not be familiar to their students. These are the only two passages where there was universal agreement between the subject specialists who were outside the appropriate subject area.

The only clear-cut divisions between members of subgroups within the main subject areas were within PST: the physical scientists and the engineers sometimes disagreed with each other. For example, the two chemists and the physicist thought that the (BSS)Qual topic would be unfamiliar to their students, whilst the two engineers thought it would be familiar. Similarly, the physicists thought that (LMS)Nitro would be familiar to their students, whereas the engineers did not. These disagreements emphasise the fact that the inclusion of Physical Science and Technology in one module is not satisfactory – there is too great a divide between physical science and engineering. However, these subject specialists showed no such disagreements over the texts within their own subject area. There was full agreement about (PST)Sun, which is arguably the only one of the three subtests to be subject specific, and no clear distinctions between the lecturers' views on (PST)Fuel and (PST)Ship. This probably emphasises the lack of specificity of the reading passages, rather than any closeness in the views of the subject specialists.

Although only a small number of lecturers commented on the reading passages, and although their comments must have been influenced by their own interests and specialisations, these views were useful as they highlighted aspects of the reading passages which we might not have suspected. For example, no person with a background in BSS would have expected that (BSS)Qual would have been considered so suitable for engineers. Subject specialists have an overall feel for their general subject area which few EAP item writers are likely to have, and the choice of specific academic texts for both teaching and testing should not therefore be left solely to BSS based teachers.

Conclusion

Since we now have a better idea of the specificity of the different subtests, Table 8.5 presents an updated version of Table 7.11. My intuitive views are listed under 'Intuit.', the results of the Table 7.8 Significance Chart are under 'Sig,', those from the bias studies in Chapter 7 under 'Bias' and those of the Subject Area Familiarity Table (see Appendix 8.1) under 'Familiarity'. The information from all the columns is combined to produce one overall classification under 'Specificity', where General (G) texts are those with three or more 'G's in the other columns, Specific (S) texts have three or more 'S's, and Highly Specific (HS) texts have only 'S's. There is no major difference between this column and the same column in Table 7.11 except that (LMS)Genes and (LMS)Nitro are now labelled as Highly Specific rather than Specific, and (PST)Ship is labelled TS (Too Specific). The students' and lecturers' comments have therefore reinforced earlier findings.

Table 8.5

Specificity Table 3

Text	Intuit.	Sig.	Bias	Familiarity	Specificity
Qual	S	S	G	S*	S
Educ	S	S	G	S	S
High	S	S	G	S	S
Tooth	G	G	G	S	G
Child	G	G	G	S	G
Genes	S	S	S	S	HS
Nitro	S	S	S	S	HS
Sun	S	S	G	S (for all sciences)	S (for all sciences)
Fuel	G	G	G	G	G
Ship	S	G	G	TS**	TS

*Although unfamiliar to 50% BSS students

** TS = Too Specific

My assumption in the classification of these reading passages, and in particular in the analysis of the Level of Significance Chart (Table 7.8), was that there would be a significant subject area effect between two specific texts, such as (BSS)Educ and (LMS)Nitro (S/HS from the specificity column above), and no significant effect between two general texts, such as (LMS)Tooth and (PST)Fuel (G/G). Between a specific and a general text such as (LMS)Genes and (PST) Fuel (HS/G), there might be little or no significant subject area effect. This assumption was tested by entering the 27 pairs of subtests in a contingency table (see Table 8.6), each pair being entered under S/S if both passages were either S or HS in the Specificity column, and under S/G if one or both were G or TS. Each pair was then classified as to whether the significance level shown in Table 7.8 was +Sig or -Sig, the level of significance being set at .05.

Table 8.6

Specificity by Significance of Subject Area Effect

	S/S	S/G	Total
+Sig	10	4	14
-Sig	1	18	19
Total	11	22	33

The chi squared test reveals one tailed significance at <.0005, so my classification of the texts achieves some support. If I follow my intuition alone, and class the Ship text as S rather than G, the result is still highly significant (p = .005), but the value of chi squared drops from 13.03 to 8.56. This implies firstly that intuition alone cannot be relied on, and secondly that running chi squared tests on the specificity classifications may be a useful way of seeing how exact those classifications are.

Before we go on to look at the content of the three modules in more depth, I shall give a brief summary of what has been learnt so far. From the students' questionnaires I have shown that familiarity with subject area has an important effect on the specificity of reading passages. I have also shown that although subject specialists disagreed about the appropriacy of some of the passages there was no disagreement about the two highly specific texts. The lecturers agreed that they were suitable for LMS students and inappropriate for BSS and PST ones.

Although we now know more about the reasons for the specificity and non-specificity of the various subtests, we have not looked at the reading passages in any detail. Chapter 9 therefore looks first at the source of the reading passages to see how much effect this might have on the specificity of the subtests, and then uses an adaptation of Bachman's TMC Rating Instrument to find out more about the content of the reading passages and also the test items.

9 Text Source and the TMC Rating Instrument

In this chapter I look at the source and content of the reading passages and test items to see if there are further reasons for the variation in subtest specificity. First I shall look at the source and rhetorical style of the reading passages, and then I shall describe the responses of three raters to facets of Bachman's TMC rating scale relating to the subtests' linguistic complexity, subject specificity, and cultural content.

Source of Reading Passages

For ease of reference the list of passages originally given in Table 6.1 is repeated in Table 9.1 below.

Table 9.1

The Reading Passages

Business Studies and Social Sciences (BSS)

1. (Qual)	'Quality Circles' (from *Study Document 342*, Incomes Data Services Ltd., Great Portland Street, London. No date)
2. (Educ)	'The Purposes of Continuing Education' (adapted from *Adult Learning: Issues and Innovations*, edited by Robert M. Smith. ERIC Clearinghouse in Career Education. 1976)
3. (High)	'Access to Higher Education' (from a Department of Education and Science White Paper on Higher Education, London. 1988)

Of the three reading passages, one is in the field of Business Studies (Quality Circles) and the other two are in Education.

Life and Medical Sciences (LMS)

1. (Tooth)	'The Mystery of Declining Tooth Decay' (from *Nature*, Vol. 322, 1986)
2. (Child)	'Our Children's Teeth' (from *The British Medical Journal*, Volume 298, February 1989)
3. (Genes)	'Three Ways to Make a Transgenic Beast' (from the *New Scientist*, 7 July 1988)

4. (Nitro)	'Nitrogen Fixation' (from 'Agricultural Microbiology', *Scientific American*, 1981)

Passages 1 and 2 are in Medicine, and Passages 3 and 4 are in the Life Sciences.

Physical Science and Technology (PST)

1. (Sun)	'Life without a Sunscreen' (from the *New Scientist*, 10 December 1988)
2. (Fuel)	'Energy from Fuels' (from a textbook written for young non-scientists; source unknown)
3. (Ship)	'The Recovery of the Mary Rose' (from *The Structural Engineer*, Vol. 62a, No. 2, 1984)

Passages 1 and 2 are in the Physical Sciences, and Passage 3 is in Engineering.

As can be seen from this table, the passages come from varied sources. The BSS module contains the greatest variety of texts: one comes from a study document, one from a British government paper, and one from an adaptation of a paper on career education. The sources of the two science modules' texts are more uniform: with one exception, the passages come from either academic journals, *Nature*, *The British Medical Journal*, and *The Structural Engineer*, or popularisations of academic science reports, the *New Scientist* and the *Scientific American*. The one exception is the (PST)Fuel text which comes from a textbook written not for science students but for young *non-scientists*. This surprising fact may well account for the lack of specificity of this subtest as although the subject matter may be appropriate for PST students, they are not its intended audience. The content of the passage is not sufficiently specialised for PST students, and it is therefore insufficiently academic. It seems likely that the more academic a piece of writing is, the more highly specific will be its subject matter, as it will be aimed at a progressively more specialised audience. Articles in learned journals are considered to be more academic than ones in popularisations such as the *New Scientist* and the *Scientific American*, and these in turn are thought to be more academic than, for instance, articles in quality newspapers. However, it is not clear that there is any agreement about what the word 'academic' means. It is used frequently without explanation. Everyone would probably agree that a research article is academic. Writers such as Fahnestock (1986) and Myers (1991) have described the differences between research articles and popularisations, and imply that popularisations are less academic. Presumably the more academic an article is, the more closely it is related to one discipline, so it might follow that texts from academic journals would be more subject specific than popularisations. Since all but one of the texts from the two IELTS science modules come from either academic journals or popularisations, we can see whether this is the case here. Table 9.2 shows the source of each of the reading passages.

Table 9.2

Rhetorical Function of the Reading Passages

Specificity	Passage	Source	Rhetorical Function
S	Qual	Study document	Description of organisation
S	Educ	Adaptation of a university paper	Enumeration of list of aims
S	High	Government report	Record and prediction of facts
G	Tooth	Academic article	Introduction: reporting research
G	Child	Academic article	Introduction: listing findings
HS	Genes	Popularisation	Description of processes
HS	Nitro	Popularisation	Description of process
S	Sun	Popularisation	Introduction: explanation
G	Fuel	Textbook for non-scientists	Description of facts
TS	Ship	Academic Article	Description of plans and narration of outcomes

Key: HS = Highly Specific
S = Specific
G = General
TS = Too Specific
(Classification based on Table 8.5)

Contrary to expectations, with the exception of (PST)Ship, which has been shown to be 'too specific', neither of the passages from academic journals seems to be specific, whereas all the three popularisations *are*. This may seem surprising. However, although the descriptions of research articles given by genre analysts such as Swales (1990) may show that research articles are highly specific to their field of study (see also Bazerman, 1988), academic articles as a whole take so many forms that it is difficult to generalise about them. Even in one discipline they may vary from general survey articles which the lay reader would understand, to ones which are so technical that even experts in the esoteric sub-discipline have difficulty understanding them. In addition, different sections of an article may vary in their specificity. Within a publication there are likely to be many kinds of discourse. The introduction, for example, may contain an easily accessible review of the literature, but this may be followed by a highly specialist description of an experiment or process. Authors such as Dudley-Evans and

Henderson (1990) have described the different styles used in different parts of a single article. Although the two 'general' LMS passages come from learned journals they are both so general in approach that they present no problems to a BSS reader. The first passage, Tooth, introduces the concept of fluoridation in water supplies, and reports on some studies on the effect of fluoridation on tooth decay. Very few technical terms are used, and those that are, are explained in the text. The second passage, Child, discusses tooth decay in children, and shows how this is related to social class. This passage is perhaps slightly more technical than the first in that it includes some statistics, but there are no concepts or terms which would be unfamiliar to a social scientist. The section of an academic article from which a passage is selected will, therefore, itself have an effect on the suitability and easiness of the passage.

Unlike the two LMS texts which come from academic articles, the two that come from popularisations are not introductions or surveys. They are either wholly or partially descriptions of processes, and this may partly account for their difficulty for non-LMS students. (LMS)Genes describes methods of transferring genes to mice, and (LMS)Nitro describes the process of nitrogen fixation. Both passages have unexplained technical vocabulary, and both demand an understanding of biological concepts. To a biologist such concepts are elementary, but to a non-scientist they are obscure. Myers (1991) has shown how the lexical cohesion which makes scientific research articles so difficult to read for non-scientists is replaced by more helpful, explanatory cohesive devices in popularisations, but in (LMS)Genes and (LMS)Nitro any such devices seem inadequate for the layman. The texts are not contextualised for non-life-science readers (see Bachman 1990), and therefore BSS and PST students have difficulty with them.

In Table 9.2 above, the fourth column shows the rhetorical function of each passage and this does appear to have some relationship with the specificity or non-specificity of the passages. The passages from academic journals come from general introductions. The two 'highly specific' texts come from popular rather than academic journals, *but* they are descriptions of processes. I have already shown that the fact that a passage is extracted from an academic article will not itself guarantee that the text is subject specific, and although it is unwise to generalise from such a small number of texts, it seems as if the rhetorical function of an extract may be of more importance than the source.

Bachman's TMC Instrument

Bachman and his colleagues designed two rating scales, the TMC Scale and the Communicative Language Ability (CLA) Scale, for the Cambridge-TOEFL Comparability Study (Bachman *et al.* 1995). The scales were drawn up in order to find a quantifiable way of comparing the content of two test batteries;

Bachman (1990) discusses the theoretical framework behind these scales. The CLA facets are rated on a five-point scale and relate to the level of ability required of the test takers in the areas of grammatical, textual, illocutionary, sociolinguistic and strategic competence. The TMC facets relate to test items and test passages, and concern the testing environment, test rubric, item type, and the nature of test input. Among the facets of test input are: complexity of language, rhetorical organisation, degree of contextualisation, test topic, cultural bias, and pragmatic characteristics. For each facet, raters assess an item or a passage according to a scale which generally has three points. For example:

	Very Simple		*Very Complex*
Rhetorical Organisation	*0*	*1*	*2*

Some facets are rated according to the number of occurrences of a feature, for example:

	No Occurrences	*One Occurrence*	*Two or more Occurrences*
Cultural References	*0*	*1*	*2*

Since the first TMC trials, the scales have been steadily refined. They were used to investigate the relationship between item content and item difficulty (Bachman *et al.* 1989), and to compare the content of different versions of a test (Bachman, Davidson and Milanovic 1991). The version of the instrument that I have adapted was drawn up in March 1991.

The complete 1991 versions of Bachman's rating instruments consist of 63 TMC and 13 CLA facets. They are designed to apply to listening, speaking, reading and writing tests in a wide range of sociolinguistic settings. Not all these facets are applicable to the IELTS reading tests, which all have the same overall test design, test reading only, and contain only academic or quasi academic passages. For my version of the scales, therefore, I was able to reduce the number of facets. However, in spite of this reduction, my first draft of the scales consisted of 35 TMC and 12 CLA facets to be assessed for ten reading passages and 95 test items. This was too daunting for the volunteer raters and I had to reduce the scope of the exercise radically. Some facets were amalgamated – grammar, for example, was no longer assessed according to embeddings, sentence type and voice, but only under the one umbrella term, 'grammatical complexity' – and, because the CLA scale seemed more valuable for assessing tests of productive than receptive skills, it was dropped completely. The only addition I made to the existing facets was that 'Degree of Contextualisation – Topic Specificity' was expanded so that raters were asked to assign ratings for students in the three subject areas of BSS, LMS and PST separately. A copy of the modified instrument is provided in Appendix 9.1; it consists of 17 facets, some of which relate to items, some to passages and some to both. It is heavily based on Bachman's 1991 instrument, but has been adjusted to take account of perceived ambiguities in the original. It should be pointed out that although these alterations appeared to clarify the raters' task, they may in some cases have obscured the purpose of the Bachman originals. To try to guard against this, raters were referred where possible to the relevant sections of Bachman's 1990 explanations of the facets concerned.

The Raters

Three raters assessed the tests. All three were applied linguists and experienced teachers of English for Academic Purposes. Two were British and one had a Canadian/British background, and all three were educated in the humanities. Their most recent EAP experience covered Indonesia, Thailand and the UK. None of them was familiar with the tests before they embarked on the rating procedure.

The academic subject specialists, too, were asked to rate the passages according to the three of Bachman's facets which were directly related to subject specificity. These were 'Specialisation of Topic', 'Degree of Contextualisation – Topic Specificity' and 'Academic Specificity'. These subject specialists received no training. It was hoped that the reworded questions and the explanations of the facets (see Appendix 8.3) were sufficiently clear. The lecturers' responses will not be treated separately, but will be reported at appropriate points during the description of the EAP raters' responses.

Training the Raters

At an initial meeting to discuss the first draft of the modified TMC instrument, several problems arose because the group felt that the explanations of some of the facets were ambiguous. Possibly the most important problem, and certainly the most enduring, related to those facets where the assessment had to be made in relation to the expected test taker. For example, raters were asked to decide on the frequency of the vocabulary used in a test passage or item, using a three-point scale ranging from 0 if the vocabulary was frequent to 2 if it was infrequent. The raters wanted to know for whom this vocabulary should be considered frequent or infrequent. Should it be for all members of the English speaking world, or all learners of English, or the specific test takers, which in the case of IELTS meant prospective L2 university students? Eventually, with Bachman's approval, it was decided that the question should be related to IELTS candidates, but even then the raters wanted to know which IELTS candidate since there was no such thing as a typical IELTS test taker. Bachman supplied some very detailed comments on this and other problems which arose (Bachman, personal communication), and the TMC Instrument was further modified. Unfortunately, because of their other commitments, the raters were not able to meet again to discuss the rating procedure together, and so the training process went as follows. One of the raters and I went through all the facets relating to one passage and its accompanying test items seeing if we could agree on which level to assign for each facet. When we did not agree we adjusted the explanation of the facet until we did. After that we separately assigned ratings to all the other passages and items, and then compared our ratings. Wherever there was disagreement we had further discussion, and adjusted the rating rubric yet further. In some cases one of us changed a rating to fit in with the other; in some cases we agreed to differ. (My own ratings were discounted in the final analysis since I had edited some of the test items during the test construction phase, and might therefore not be impartial.) After that I went through one passage and its items with each of the other two raters, before they carried out the rest of the rating by themselves. The rating itself took each rater approximately three hours.

None of the three raters were confident about their assessments. They felt that although some facets were unambiguous and straightforward to answer (such as 'Figurative Language' [see Appendix 9.1], where the rating would be the same for any group of test takers), they were still worried by others. They all said that they would not expect to give the same ratings another time as they felt their internal rules for assessing the facets kept changing. As a check on this, Rater 2 carried out the whole exercise again eight weeks later (see below).

For the sake of clarity the raters' views on the reading passages and the test items will be reported separately.

TMC Ratings of the Reading Passages

Agreement among Raters

It was not possible to calculate a standard reliability index for the raters, as the number of rating categories was too small. It was only possible to check the agreement between raters. Table 9.3 gives the weighted Kappa statistic for each pair of judges, weighted to take account of the fact that if raters differed by two points when assessing a facet this was a more serious disagreement than if they had differed by one (see Everitt 1968). The fourth row shows the agreement between Rater 2's first and second marking. (Note that a Kappa of .00 would show that any agreement could have been due to chance, and that 1.00 would show perfect agreement.)

Table 9.3

Weighted Kappa Statistics

Raters	Weighted Kappa	95% Confidence Interval
1/2	.36	.21 to .51
1/3	.48	.34 to .61
2/3	.45	.31 to .58
1st/2nd rating	.45	.32 to .58

All four weighted Kappas are significantly different from chance at the .05 level, and show a moderate agreement between the raters. No one rater stands out as having performed very differently from the other two, but Raters 1 and 2 showed the least agreement. The fact that the agreement between Rater 2's first and second marking is also only moderate shows that the raters were probably right to doubt their consistency.

A weighted RAP (Rater Agreement Proportion) statistic was calculated on each facet for each passage. The RAP statistic is used by Bachman, Davidson and Milanovic (1991) to measure the proportion of rater agreement on each facet/item. In their research, for example, where there were five raters, the RAP was 1.0 (5/5) if all five raters agreed, and .8 if four did. If two agreed it was .2 and no agreement was .0. To find out the agreement over facets or items, the mean RAP was calculated for each variable. This statistic is easy to conceptualise but it has one disadvantage: it does not take account of extreme judgements. So, for example, if three raters gave a 2, and two raters gave a 0, the RAP index would be the same as if three had given a 2 and two a 1. For the present study, the scale has been adjusted to account for extreme ratings. If all three raters agree, the RAP figure is 1, and if two agree with a difference of 1 between the ratings, then the RAP is .67 (2/3). If, however, two agree, but the third is *two* points away from the others, then the RAP is .33. If no-one agrees the RAP is .0. This, therefore, could be called a Weighted RAP or WRAP. Table 9.4 gives the mean WRAPs for all the facets. (A more detailed table giving the mean WRAPs for all the passages is provided in Table 9.5)

It will be seen from these figures that, in spite of the raters' doubts, there was quite high agreement for some of the facets. The raters mostly agreed about 'Grammar' and 'Cohesion', although in both cases Bachman's detailed facets had been conflated into single variables. 'Frequency of Vocabulary', too, which had caused such anxiety, had a mean WRAP of .80 and so did 'Rhetorical Organisation', which two of the raters had felt was impossible to rate in the manner outlined in the TMC instrument (see Appendix 9.1, page 286). There was, however, little agreement on some of the other facets. Since it is possible to get a mean WRAP of .48 by chance alone, it seems that where facets have a mean WRAP of less than about .7 there is too little agreement for us to believe that the raters were using the same criteria for their judgements. What is interesting is that the facets which led to such disagreement are, with one exception, those most obviously related to the problem of text specificity, namely 'Contextualisation – Topic Specificity', 'Specialised Topic' and 'Academic Specificity'.

Table 9.4

Mean WRAPs* for Each Facet Relating to the Reading Passages

Facet	Mean WRAP	Facet	Mean WRAP
Vocabulary		**Organisational Characteristics**	
Infrequent	.80	Grammar	.87
Specialised	.77	Cohesion	.90
Ambiguous	.80	Rhetorical Organisation	.80
		N. of Types of Rhet. Org.	.70
Degree of Contextualisation			
Cultural Content	.73	**Sociolinguistic Characteristics**	
Topic Specificity		Cultural References	.68
BSS	.70	Figurative Language	.80
LMS	.66		
PST	.66		
Topic			
Culture Specific	.77		
Academic Specific	.57		
Specialised Topic	.67	**Overall Mean WRAP**	.74

* All the mean WRAPs are slightly inflated since one passage, (PST)Sun, was used for training purposes.

The only exception is 'Cultural References' (mean WRAP = .68) where raters had to rate passages according to the number of cultural references in the text (see Appendix 9.1, page 287). This appeared to be a fairly mechanical task which should have led to high agreement. It may be that the raters interpreted this facet in different ways, or it may be that they have varying perceptions of what cultural references are. The facet which led to the greatest disagreement among raters was 'Topic: Academic Specific' (see Appendix 9.1, page 285). Raters were asked to state how specifically academic each reading passage was, regardless of the test taker. It is possible that since all the IELTS passages are supposedly suitable for academic study, there was not enough range in the passages for the raters to gain an impression of what a 0, a 1 or a 2 would mean. On the other hand it may be that some passages are very difficult to rate in this way. As was discussed earlier, there may be no clear concept of what 'academic' means. There was total disagreement over the academic specificity of two of the passages – (LMS)Tooth and (LMS)Child: in each case one rater gave a 0, one a 1, and one a 2. These were two of the passages that I had labelled 'general', and, although they both came from academic journals (see Table 9.2), the subject matter was not only familiar to readers in many disciplines, but it was presented in a non-specialised way. Interestingly, the academic subject specialists also disagreed with each other about this, though not so strongly. Most of them gave a non-committal rating of 1 each time, but six said that (LMS)Tooth was not at all academically specific, and one said it was highly so. The responses for (LMS)Child were similar. The other passage where 'Academic Specificity' produced little agreement among the EAP raters was (PST)Ship, which I have labelled 'too specific'. Its WRAP was .33. What the TMC instrument may be confirming is what I suspected earlier, namely that there is as yet no consensus on what makes an 'academic' text.

The other facets with low mean WRAPS were the three 'Contextualisation - Topic Specificity' ones (see Appendix 9.1, page 284). Here the raters had to rate the passages according to the viewpoint of BSS, LMS and PST students. This may seem to have been a strange thing to ask them to do, but that is what item writers are doing when they select a text for an ESP test. Since all three raters had social science backgrounds it might be expected that they would agree most on the BSS facet, and indeed they did, but the mean WRAP was still low, only .70. It may not only be the difficulty of judging the contextualisation from different viewpoints which is the problem here. All three raters found it difficult to grasp the concept behind this facet. They said that they never managed to internalise it, and kept having to re-read the instructions. The academic subject specialists were also given this question in the pilot questionnaire, but it caused such problems that it was turned into two questions, 'New Knowledge' and 'Explanation of New Knowledge' (see Appendix 8.3). What can be deduced from the above, I think, is that for some facets the low agreement is due to inadequate explanation, and in some it is because the concepts are not yet sufficiently defined. In the case of the 'Contextualisation – Topic Specificity' facets, the raters may not have the appropriate background knowledge to be able to agree on an answer, but they are also uncertain about the concept.

Use of the TMC Instrument

Linguistic Complexity

We now need to see whether any of the facets can throw light on the specificity of the reading passages, and it is certainly possible to show the ones which do not. Table 9.5 gives a complete picture of the ratings on all facets for all passages. The three raters' assessments have been totalled, so that, for example, if each gave a 1, the total is 3. So for the (LMS)Child passage, the raters' combined rating for 'Infrequent Vocabulary' (separately 2, 2, and 1) is 5. In this table the facets are grouped under three headings: Linguistic Complexity, Subject Specificity, and Culture. Looking at the facets across the three subject areas of BSS, LMS and PST, it can be seen that those grouped under Linguistic Complexity do not seem to differentiate across the modules. For 'Ambiguity', for example, all the passages are considered to be relatively unambiguous, and have total scores of 3 or less. Some of the facets seem to work in pairs, witness 'Grammar' and 'Cohesion', which appear to be almost identical, and are very steady across all ten passages. With the exception of (BSS)Qual, which has a lower rating for 'Cohesion', the passages have all been rated as having very complex Grammar and Cohesive Devices. This consistency may show that it is difficult to distinguish between such academically sophisticated texts, or it may be that my reduction of Bachman's more detailed facets has obscured interesting differences. On the other hand, since the 'Cohesion' facet comes directly after 'Grammar', it may be that 'Cohesion' is suffering from a halo effect. It is possible that the raters were rating both facets according to some general recognition of linguistic complexity. 'Rhetorical Organisation' and 'Number of Types of Rhetorical Organisation' also have almost identical ratings, and so do 'Specialised Vocabulary' and 'Specialised Topic'. This latter pair is particularly interesting as the facets appear in different parts of the TMC instrument, and might not therefore be expected to suffer from a halo effect. It seems that the raters are using the same criteria to assess these facets, although for 'Specialised Vocabulary' they are supposed to hold IELTS test takers in mind, and for 'Specialised Topic' they are not.

It seems from the above that the raters did not find any striking differences in the linguistic complexity of the passages in the three subject areas, so this aspect of the reading passages does not add to our knowledge of why some texts are more specific than others.

Table 9.5

TMC Ratings on All Facets for All Passages

	BSS			LMS				PST			
	S	S	S	G	G	HS	HS	S	G	TS	
	Qual	Educ	High	Tooth	Child	Gene	Nitro	Sun	Fuel	Ship	Mean
Linguistic Complexity											
Infr.Vocab.	5	5	6	4	5	6	5	6	4	6	.80
Fig.Lang.	6	6	5	3	3	6	2	0	2	6	.80
Ambiguity	0	1	3	0	0	1	0	3	0	2	.80
Grammar	5	5	6	6	6	6	6	6	5	5	.87
Cohesion	3	5	6	6	6	6	6	6	5	5	.90
Rhet.Org.	3	3	5	3	3	3	2	3	3	2	.80
N Rhet.Org.	6	3	5	3	3	3	2	4	4	3	.70
Subject Specificity											
Context:											
ASS	3	2	5	3	2	5	6	5	2	4	.70
LMS	4	2	6	1	2	2	2	3	2	4	.66
PST	4	2	6	3	2	4	4	2	1	2	.66
Ac.Specific	2	2	6	3	3	4	6	4	4	2	.57
Special.Top.	5	2	6	3	3	6	6	4	3	6	.67
Special.Voc.	5	4	6	4	4	6	6	6	3	6	.77
Culture											
Cult.Ref.	5	0	4	3	3	1	1	0	0	2	.68
Cult.Cont.	3	1	5	1	5	1	1	3	2	1	.73
Cult.Specif.	4	2	6	2	5	1	1	0	0	2	.77
Mean WRAP	.81	.63	.79	.79	.60	.81	.75	.92	.60	.74	

Culture

Here differences can be seen between the three subject areas. The most outstanding feature is that the PST passages appear to be the least culturally specific. Two out of the three texts are considered to contain no Culture Specific content (Sun and Fuel) and the amount of unexplained information for which prior cultural knowledge is required (Contextualisation – Cultural Content) is low for all three passages. In the BSS and the LMS modules, however, the figures are much higher. (BSS)Qual and (BSS)Higher have high ratings on at least two of these variables, and so does (LMS)Child. In the case of Higher and Child, at least, these are easily explained.

(BSS)Higher is part of a British government report on projections for higher education in Britain. Since the text is aimed at a readership which is familiar with the British educational system it uses terms such as 'polytechnic' and 'Highers' without any explanation. Similarly, the (LMS)Child passage compares tooth decay in different areas of Britain. This time it is not the vocabulary that might give problems to candidates without prior knowledge of Britain, but knowledge of the relative wealth of different areas in Britain. Without this knowledge it is difficult to make sense of various of the references.

From the raters' answers it seems that two of the texts, one in BSS and one in LMS, contain uncontextualised cultural content which might be inappropriate for a test such as IELTS which is targeted at students from many countries. This certainly casts doubt on the content validity of the subtests in which they appear. However, it is not clear whether these differences relate to the texts' subject specificity. Without more text samples we cannot know if there is a general trend for BSS and LMS material to contain more uncontextualised cultural references than PST texts; and this would be worth investigating further. From this study there is no evidence that the cultural specificity of two of the texts has contributed to the subject specificity of the subtests in which they appear.

Subject Specificity

Since all the subject specificity facets had low agreement among raters it is not possible to make any deductions from the total ratings. However, it is possible to look at them in more detail, and to see if we can account for the lack of agreement. Table 9.6 shows the totals for the facets relating to subject specificity for the three passages which had been labelled 'general'. The WRAP statistic for each facet is shown in brackets. Table 9.7 shows the same facets for the 'highly specific' subtests.

Although most of the facets in Table 9.6 show moderate agreement (.67), it is interesting that there is total disagreement among the raters concerning the 'Specialised Topic' facet for two of the passages, (LMS)Child and (LMS)Fuel. Although this may be accounted for by a lack of clarity in the TMC Instrument, it is more likely to be the result of the generality of the texts. In Table 9.7 it will be seen that there is total agreement as to the specificity of both the 'highly specific' texts.

Table 9.6

Totals and WRAP Statistics for the Three 'General' Passages

	(LMS)Tooth	(LMS)Child	(PST)Fuel
Contextualisation			
BSS	3 (1.0)	2 (.67)	2 (.67)
LMS	1 (.67)	2 (.67)	2 (.67)
PST	3 (1.0)	2 (.67)	1 (.67)
Academic Specific	3 (0)	3 (0)	4 (.67)
Specialised Topic	3 (.67)	3 (0)	3 (0)
Specialised Vocabulary	4 (.67)	4 (.67)	3 (0)

Table 9.7

Totals and WRAP Statistics for the Two 'Highly Specific' Passages

	(LMS)Genes	(LMS)Nitro
Contextualisation:		
BSS	5 (.67)	6 (.67)
LMS	2 (.67)	2 (.67)
PST	4 (.67)	4 (.67)
Academic Specific	4 (.33)	6 (1.0)
Specialised Topic	6 (1.0)	6 (1.0)
Specialised Vocabulary	6 (1.0)	6 (1.0)

It seems clear that however we approach those 'general' passages their properties are not as easily distinguishable as are those of the 'highly specific' ones. If we look, for example, at the contextualisation facets, we can see that for the 'highly specific' passages, the raters have judged them most contextualised for LMS students, and least for BSS ones, which is just as might be expected. For the 'general' texts too the ratings are as expected: there is no such clear-cut delineation, and indeed (LMS)Child is considered to be equally contextualised for BSS, LMS and PST students.

The academic subject specialists' answers to their simplified version of the 'Contextualisation - Topic Specificity' question (see Appendix 8.3, Questions 4a and 4b) followed the same general pattern, with the 'general' LMS subtest showing mixed responses, and the 'highly specific' ones showing clear divisions into subject areas. For example, only one or two specialists in any of the three subject areas said there was a lot of new knowledge in (LMS)Child, and they all said that any new knowledge was either fully or partially explained. However, for (LMS)Nitro, again as expected, all the BSS specialists said there was a lot of new knowledge which was not explained, whilst most of the LMS specialists thought there was little or no new knowledge and that it was all fully explained.

This trend is to some extent reinforced if we look at how far the raters agreed over individual passages. Of course not too much should be made of these findings as there is so much room for error, but it does seem to be the case that the 'general' passages tend to have lower agreement among judges. The two which have the lowest agreement (Mean WRAP .60) are (PST)Fuel and (LMS)Child (see Table 9.5).

Although the lack of agreement among the raters means I cannot draw any new conclusions about the subject specificity of the reading passages, an interesting point emerged from the survey. Raters had no difficulty agreeing about the specificity of the topics and vocabulary of the two passages which were classed in Table 8.5 as 'highly specific', but they were not able to agree about the 'general' texts. Test writers might find it useful, therefore, to ask judges to rate possible reading passages for topic and lexical specialisation before making final selections for ESAP tests. If the judges cannot agree on these two facets, the passages are probably not suitable.

TMC Ratings on Passages – Conclusion

To conclude this section on the reading passages, I have not succeeded in finding a formula for identifying factors contributing to the specificity of texts. It may be that the passages are too similar to each other to be suitable for this kind of analysis, or that the training of raters was not sufficiently detailed, or that the modified TMC Instrument is ambiguous in parts, or that the three point scale is too limited. On the other hand it may be that it would be very difficult to get more agreement on some of these facets under any circumstances. Although it should surely be possible to achieve total agreement on the facets relating to culture, and those requiring simple counts such as 'cultural references' and 'figurative references', and high agreement on those concerning linguistic complexity, as indeed has been shown in the various Bachman studies, agreement on those relating to specialised topics may be more elusive. We may still not know enough about academic and subject specificity for the raters to be able to agree.

TMC Ratings of the Test Items

The raters assessed each test item for the clarity of the rubric, the familiarity of the item type, the frequency, specificity and ambiguity of the vocabulary, and the complexity of the grammar (see Appendix 9.1, page 286). The introductory rubric for a set of items was included in the assessment of the first item in the set, and so was any additional material which related to the complete set. So, for example, the list of phrases to be matched with M(BSS) Items 13–19 (see Appendix 6.1) are considered to be part of Item 13. Items such as M(BSS) 14–19, which have no new text, were not assessed.

There was one TMC question, 'Relationship of item to reading passage' (see

Appendix 9.1, page 999), which was based on a 7-point rather than a 3-point scale, so the responses to it could not be analysed with the responses to the other TMC questions. Responses to this question will therefore be reported first.

The three judges generally agreed about this facet: the mean WRAPS for the LMS and PST modules were .88 and .89 respectively, and for M(BSS) there was almost total agreement at .99. The judges almost unanimously said that all but five of the 115 items related only to one part of a passage, and required only localised understanding of that part. This supports my comment in Chapter 6 that few items seemed to be testing the wider reading skills such as 'finding main ideas', 'identifying the relationships between ideas in a text' and 'reaching a conclusion by relating supporting evidence to the main idea' which are listed in the specifications, and which require a reading of more than one sentence (see Appendix 4.5).

For the other six facets there was much less agreement, as can be seen from Table 9.8. All four Kappas are again significant at the .05 level, but the agreements between the pairs of raters are low. That between Raters 1 and 3 is the strongest, but even so .30 is very low, and it should be held in mind that the agreement between Rater 2's first and second ratings is only .22.

Table 9.8

Weighted Kappa Statistics – Test Items

Raters	Weighted Kappa	95% Confidence Interval
1/2	.28	.13 to .42
1/3	.30	.16 to .44
2/3	.21	.06 to .36
1st/2nd rating	.22	.08 to .35

The mean WRAPS in Table 9.9, show that there was more agreement for some facets than others. All three mean WRAPS for Vocabulary Ambiguity, for example, are above .80, and the total ratings for each item are mainly 0s and 1s. The item vocabulary in all three modules was therefore considered to be clear and unambiguous.

Table 9.9

TMC Total Ratings on Items

Item	Rubric Clarity	Item Famil.	Vocab. Freq.	Vocab. Spec.	Vocab. Ambig.	Grammar Level	Mean WRAP
BSS MODULE							
1	4	3	2	2	2	3	.39
5	-	3	2	2	2	3	.53
6	-	4	2	2	0	2	.60
7	-	2	2	2	0	2	.53
8	-	2	2	2	0	2	.53
9	-	2	2	2	1	1	.53
10	-	3	0	2	0	1	.60
11	-	2	2	2	0	2	.60
12	-	3	2	2	0	5	.53
13	1	3	4	3	0	2	.61
20	2	2	3	3	0	4	.78
28	0	0	3	2	1	1	.83
29	-	-	3	0	0	1	.91
30	-	-	0	0	0	1	.91
31	-	-	3	2	0	2	.83
32	5	5	1	3	0	2	.78
M.WRAP	.74	.41	.69	.60	.81	.65	

Item	Rubric Clarity	Item Famil.	Vocab. Freq.	Vocab. Spec.	Vocab. Ambig.	Grammar Level	Mean WRAP
LMS MODULE							
1	1	3	4	4	0	5	.78
19	0	0	2	3	0	4	.72
20	-	-	3	4	1	4	.75
21	-	-	4	4	0	4	.75
22	-	-	5	4	0	5	.75
23	-	-	3	3	0	4	.67
24	3	4	5	4	0	5	.78
28	1	3	3	5	0	3	.56
29	2	3	5	5	1	1	.56
30	-	-	5	5	1	1	.67
31	-	-	4	4	0	1	.75
32	-	-	4	4	0	1	.75
33	0	0	3	3	0	3	.50
34	-	-	2	2	0	2	.58
35	-	-	4	5	0	0	.84
36	-	1	3	4	1	3	.60
37	1	2	4	6	1	0	.72
M.WRAP	.76	.75	.57	.51	.90	.59	

Table 9.9 Continued

Item	Rubric Clarity	Item Famil.	Vocab. Freq.	Vocab. Spec.	Vocab. Ambig.	Grammar Level	Mean WRAP
PST MODULE							
1	1	0	4	4	0	0	.84
2	-	-	3	3	0	0	1.00
3	-	-	3	3	1	0	.92
4	3	3	-	-	-	-	.00
12	0	0	2	0	1	1	.84
13	-	-	0	1	1	1	.75
14	-	-	2	2	0	2	.75
15	-	0	1	1	0	3	.67
19	2	4	4	4	0	1	.73
28	1	2	5	4	1	6	.67
M.WRAP	.60	.72	.78	.78	.85	.74	

For the other facets the agreement among the judges varied according to the modules, with agreement on the PST items being the strongest. Overall the mean WRAPs were not high, with several below .70, but in only 17 out of the 212 ratings was there a difference of more than one between the three raters. If we accept facet ratings with mean WRAPs of more than .70, then we can presume that the rubric for each set of M(BSS) and M(LMS) items was generally clear (the total ratings were 3 or less), with the exception of M(BSS) Items 1 and 32, which totalled 4 and 5 respectively.

However, the grammatical complexity of the instructions to the summary completion tasks in all three modules, M(BSS) Item 20, M(LMS) Item 1, and M(PST) Item 28, is considered to be very complex (total ratings of 4 or more). This is not surprising since for these items the item stimulus is taken to include the complete text of the gap-filling tasks. The lowest of all the facet mean WRAPs is for M(BSS) Item Familiarity, where the raters show almost no agreement. They have very different ideas of what sorts of items are likely to be familiar to IELTS candidates.

As far as specificity goes, two out of the three raters, all of whom, it must be remembered, have social science backgrounds, did not think that the language of the M(BSS) items had specialised vocabulary – there were no total ratings above 3. However, one of the three gave Items 6 to 12 a 2 each time, while the others gave 0s. Three items in M(PST) and eight in M(LMS) had total ratings of 4, and in addition M(LMS) had 4 items with total ratings of 5, and one of 6. These items all cited specialised words from the text. For example, M(PST) Item 1 contained the acronym 'CFC', and in Items 19 and 28 the lists of words and phrases from which the candidates had to select answers included 'hull', 'cradle' and 'lifting frame' (Item 19) and 'sonar docking', 'stabbing guides' and 'jacks' (Item 28). In the LMS module, Item 37, which was rated 6, contained the title of the reading passage 'Nitrogen Fixation' and Item 35, which was rated 5, asked for the definition of a 'mosaic animal'. Although it is clear from this that some items do have specialised vocabulary, they are only specialised because they include words from the reading passages. Therefore the items are not increasing the specialisation of the subtests; they are simply reinforcing any specialisation that there may already be in the passages.

Overall, there do not seem to be any great differences in the rubric, item types, frequency of vocabulary, and grammatical complexity of the items across the three reading modules. On the whole the raters considered that, although some of the item types were unusual, the items were clearly explained, and were generally presented in simple language. There appeared to be no marked differences among the subtests. The tentative conclusion from the analysis of the items is that item content does not appear to have affected the specificity of the subtests.

This analysis of the facets of the test items and passages reveals weaknesses in the tests which might not have been revealed otherwise. For example, two of the reading passages appear to contain uncontextualised cultural knowledge which many students might not have. In addition, most of the test items only require students to read a phrase or at most a sentence, and do not make students use such skills as 'finding main ideas' and 'identifying the relationship between ideas in a text'. The tests do not, therefore, fully match the test specifications (see Appendix 4.5) and therefore lack content validity.

This analysis also reveals weaknesses in this modified TMC instrument. It seems likely that for some facets raters are not sure what they are supposed to do, and the present system of rating the items does not suit matching and gap-filling tasks. However, this does not invalidate this TMC instrument. The very process of trying to assign a rating forces raters to think about aspects of the test which might not otherwise have occurred to them. Were it not for the length of even the modified TMC instrument I would recommend that it be used as a standard tool for the content validation of new tests.

Conclusion

Chapters 8 and 9 have identified some of the factors which appear to affect the specificity of ESP reading tests. By comparing various features of the 'highly specific' subtests with those of the 'general' ones it has been possible to show the following:

1 The reading subtests' subject specificity partially depends on students' familiarity or lack of familiarity with the subject area of the reading passages.

2 Some of the passages are either too general or too specialised for students in the intended subject areas. This has led to the poor subject specificity of some of the subtests.

3 It is not clear how much effect source of text has on the specificity of a subtest. It is not the case that an extract from an academic journal is automatically specific to students in that subject area. Some extracts may be too general, and some too specific. Specificity may well depend on the rhetorical function of the passage. It may also depend not so much on the presence of subject specific vocabulary as the presence of unexplained subject specific concepts. There was not enough agreement among raters for the Bachman 'Degree of Contextualisation: Topic Specificity' facet to throw any light on this. It needs more research.

4 According to the TMC ratings, subject specificity seemed to have no effect on linguistic complexity. However, it may be that the TMC three-point scale is insufficiently detailed to distinguish between texts of the complexity of the IELTS ones.

5 There was some difference in the Cultural Content of the three modules but it is not clear whether this affected subject specificity. This needs to be considered further.

6 The study of the TMC ratings of the test items provided no evidence that the items' content was affecting the subject specificity of the individual subtests.

Chapter 10 reports the final research findings. It looks beyond the effect of academic discipline on reading comprehension to other aspects of background knowledge, and sees whether the effect of background knowledge varies according to the level of students' L2 language proficiency.

10 Language Proficiency and Background Knowledge

In order to study the effect of academic subject area on reading test performance, I have based most of my studies on students' scores on the three IELTS academic reading modules. However, since I have now shown that some of the components of these modules are not specific for their intended audience, scores on the complete modules may not provide the most useful information for my purposes. In this chapter, therefore, I describe the results of a new set of repeated measures analyses of variance which follow the design of those described in Chapter 7, but which are based on students' scores once the non-specific subtests have been omitted from the modules.

In these analyses of variance the students are classified according to their field of study. However, such a classification relates to only one aspect of background knowledge, and takes no account of other ways in which students might have become familiar with the subject areas of the reading passages. I therefore also investigate the questionnaire answers to see how much effect other aspects of background knowledge seem to have had on students' reading comprehension.

Later in the chapter I use multiple regression analysis to look at the effects of language proficiency and various aspects of background knowledge on reading test performance, and in the final section I see whether the effect of background knowledge on test performance remains constant at different levels of language proficiency.

The Revised Reading Modules

Repeated Measures Analysis of Variance on Selected Subtests

If we omit the scores on the non-specific subtests, (LMS)Tooth, (LMS)Child, (PST)Fuel and (PST)Ship from the data set, we are left with the results of three modules which are more subject specific. Table 10.1 gives the distribution statistics of these revised modules, and Appendix 10.1 contains the score distribution tables.

Table 10.1

Revised Reading Modules – Distribution Statistics and Reliability

	M(BSS) (unchanged)	M(LMS) (Genes + Nitro)	M(PST) (Sun)
Number of Students	634	513	527
Number of Items	35	13	11
Mean Raw Score	16.87	4.54	6.75
Raw Score S.D.	6.60	3.26	2.75
Mean Score as a %	48%	35%	61%
Percentage Score S.D.	18.86	25.07	25.00
Mode as a %	43%	0%	82%
Median as a %	46%	31%	64%
Max/Min as %	97%/6%	100%/0%	100%/0%
Reliability (KR 20)	.85	.81	.77

Percentages are rounded up to the nearest whole number.

It will be seen from these figures that the revised modules vary widely in the number of test items, and also in level of difficulty. The mean score for the revised M(PST), for example, is 61%, whereas for the revised M(LMS) it is 35%. The revised LMS appears so difficult partly because the two remaining subtests are the last two in the original module, and many students left some items blank. This difference in the level of difficulty in the three revised modules is unfortunate as it is likely to obscure genuine differences in the effect of subject area and background knowledge. Table 10.2 gives the repeated measures analysis of variance F-ratios and their significance levels.

These results highlight the difference in the difficulty levels of the three revised modules. In all three pairings there is a highly significant difference between the two tests ($p = <.001$).

By omitting all the non-specific tests we are naturally maximising the subject area effect between the three pairings, and it is not surprising therefore that the BSS/LMS and the BSS/PST subject area effects have p values of $<.001$. However, what is interesting is that the subject area effect between LMS and PST, which is significant at .05 when the whole modules were included (see Table 6.2), is now not significant. This may add further reinforcement to the suggestion that there was not a sufficient difference between the two academic subject areas for it to be necessary to give LMS and PST students separate reading modules. On the other hand any subject area effect may have been obscured by the fact that both the group and test variables show highly significant effects. (In the original analysis in Chapter 7, too, the LMS students scored significantly higher than the PST students ($p = <.001$), but there was no test effect. The wide disparity in the difficulty levels of the revised LMS and PST when combined with the disparity in the ability levels of the two groups may have reduced the overall subject area effect.

Table 10.2

Repeated Measures Analysis of Variance – Revised Modules

BSS/LMS (n = 175/82)		**Means (%)**			**F**	**P**
Groups	(BSS)	39.7	(LMS) 40.7	Between Groups:	.12	NS
Tests	M(BSS)	50.1	M(LMS)29.9	Between Tests:	154.63	<.001
Diag. means: Disadvant.		33.1	Advant. 46.9	Subject Area Effect:	28.74	<.001
BSS/PST (n = 165/105)						
Groups	(BSS)	51.1	(PST) 54.6	Between Groups:	2.37	NS
Tests	M(BSS)	46.9	M(PST) 58.0	Between Tests:	79.19	<.001
Diag. means: Disadvant.		49.8	Advant. 54.4	Subject Area Effect:	27.16	<.001
LMS/PST (n = 70/89)						
Groups	(LMS)	62.1	(PST) 48.8	Between Groups:	17.93	<.001
Tests	M(LMS)	40.3	M(PST) 69.0	Between Tests:	169.06	<.001
Diag. means: Disadvant.		50.9	Advant. 58.4	Subject Area Effect:	3.62	NS

The results from these revised modules will be used again when I look into the effect of level of language proficiency on the use of background knowledge on reading comprehension.

Effect of Background Knowledge on Reading Comprehension

It is, of course, difficult if not impossible to gain a complete picture of a person's background knowledge, and, as we saw in Chapter 3, different researchers have looked at it in different ways. Some researchers have equated it with aspects of a student's past experience. For example, as we saw in Chapter 3, Johnson (1981) used 'home culture' as the independent variable when assessing the effect of background knowledge on reading comprehension, and Carrell (1983) looked at the effect on recall of students' familiarity with text topics. Other researchers have assessed students' subject knowledge by giving them tests. Bernhardt (1991), for example, asked students to do a free association task, and Tan (1990) gave students tests of knowledge about the subject areas in which she was interested (see Chapter 1). None of these researchers gained a complete picture of their students' background knowledge, and indeed it is difficult to know how they could have. Background knowledge comes from a wide range of different sources; it can be built up systematically or acquired by chance and it is unlikely ever to be fully charted. All we can do is to gather as much information as possible about the sources and scope of this knowledge, and accept that this information will be incomplete. In my study, time constraints limited the number of questions about background knowledge that could be asked, but the responses to those questions that I did ask are reported on here.

Of the questionnaire items which were directly related to background knowledge, Questions 8, 9, and 10 (see Appendix 6.4) concerned students' past and present fields of study and were used to classify students into academic subject areas (Chapter 7). Question 17 asked students about their reading habits, and will be discussed below. Questions 20 and 22, which referred to the (PST)Sun reading passage, and the parallel questions which related to other texts, asked students whether they were familiar with the subject area and topic of each of the reading passages. The responses to these questions helped to identify those reading passages which were not subject specific (see Chapter 8). Questions 21 and 23, and their parallel versions, asked those students who were familiar with a subject area or topic whether this familiarity had helped them answer the test questions. The answers to these two questions will be briefly discussed later.

Background Reading

The fact that people study in one particular subject area does not, of course, mean that they are ignorant about other subjects and unfamiliar with other rhetorical styles, since they may well read books and articles in subjects outside their own academic field. The student questionnaire, therefore, included the following question:

> *17. Think about the reading you do for your work or during your spare time. Do you read books, magazines, academic papers or newspaper articles on any of the following subjects? (Please circle 1, 2, or 3 for each subject.)*

	Often	*Sometimes*	*Never*
Physical Science	*1*	*2*	*3*
Natural (Life) Science	*1*	*2*	*3*
etc.			

(See Appendix 6.4 for the complete question.)

The students' answers were coded in the same way as 'Past Subject', so that, for example, material in the Life and Medical Sciences was coded LMS. If a student read, say, articles on history and physical science, that student was coded as BSS + PST. The students' answers to 'often' and 'sometimes' were afterwards conflated, since it was impossible to distinguish between them.

The question about background reading is very general and may well have been interpreted by different students in different ways. However, it does enable us to have a rough idea of what students read in areas outside their own subject area.

Table 10.3 shows what percentage of BSS, LMS and PST students read material in each of the three subject areas. For example, 36% of the BSS students read nothing but BSS texts, and 24% read BSS and LMS texts. What is noticeable is that although 36% of the BSS students did not read outside their own broad subject area, most of the science students did.

Table 10.3

Choice of Reading Matter

	Students' Academic Subject Area		
Reading Matter	**BSS**	**LMS**	**PST**
BSS	36%	1%	5%
LMS	-	8%	1%
PST	-	1%	9%
BSS and LMS	24%	21%	5%
BSS and PST	6%	3%	21%
LMS and PST	-	4%	3%
BSS and LMS and PST	34%	62%	56%

Only 8% of the LMS students and 9% of the PST students read nothing but texts in their own subject area. 87% of both LMS and PST students said that they read at least some BSS materials, whereas only 64% of BSS students said they read science texts. This finding agrees with popular conceptions about the differences between scientists and non-scientists. Most people, whether they are scientists or not, are expected to read novels, and comments on social and political events, but it is widely accepted that many non-scientists show little interest in scientific matters. It is also true that although it is possible to appear to function adequately in society with little or no scientific knowledge, it may not be possible to do so without some knowledge of the main social and economic issues of the times. It is therefore not surprising that so many science students read about BSS subjects.

In order to assess the effect of academic subject area plus reading background on students' test performance, I reclassified the students into three groups according to their present/past academic subject area and their reading interests. Those LMS students, for example, who read BSS as well as LMS texts, were reclassified as BSS + LMS, and those who read in BSS and PST areas as well as LMS were reclassified as BSS + LMS + PST. Repeated measures analyses of variance were run on this reclassified data, but so many of the science students had to be omitted from the calculations because they were now classified as BSS

+ LMS + PST that there were too few scientists for the results to be worth considering. For example, only nine of the reclassified LMS students took the BSS and LMS modules, and only 16 took M(LMS) and M(PST). This confirms my earlier suggestion that once students are classified according to more than one aspect of background knowledge the boundaries of the different categories become fuzzy, and it is no longer so easy to distinguish one student from another. It also confirms the somewhat arbitrary nature of classifying students according to their past subject area. However, so many BSS students were still classed as BSS under the new classification that 89 BSS students took M(BSS) and M(LMS), and 99 took M(BSS) and M(PST). This reinforces the finding that although very few science students restrict their reading to scientific subjects, many BSS students do not read about scientific subjects.

These findings have two implications: firstly the fact that so many students read outside their own broad subject area means that it is inadequate to classify students only according to their academic subject area when researching into the effect of background knowledge on comprehension. Secondly, it may mean that it is impossible accurately to assign students to subject areas because of the breadth of their interests outside academic study, and that dividing students up into academic subject areas, certainly at the superficial level of the IELTS test, may therefore be both unnecessary and impossible. In addition, the fact that 35% of the BSS students read nothing but BSS texts, whereas only 8% of the LMS and 9% of the PST students read only in their own subject area, supports my suggestion in Chapter 8 that IELTS examiners might be justified in giving BSS texts to all students.

Questions 21 and 23: The Role of Familiarity of Subject Area or Topic

Table 10.4 gives the percentages of students who thought that their familiarity with the subject area or topic of a reading passage helped them to answer the comprehension questions.

Table 10.4

Percentage of Students Finding Familiarity Helpful

		BSS			LMS			PST	
	Qual	Educ	High	Teeth	Genes	Nitro	Sun	Fuel	Ship
Area Familiar	72%	70%	65%	60%	63%	61%	76%	77%	47%
N. of Students	265	356	262	193	157	169	300	358	44
Topic Familiar	78%	68%	68%	61%	61%	67%	73%	78%	55%
N. of Students	190	328	136	163	89	124	338	393	34

It is difficult for examinees to know how much background knowledge has helped them to answer test questions since such help may be received subconsciously. For example the reader may not be aware while reading a text that familiarity with the rhetorical style or the underlying scientific concept is making comprehension easier. The above percentages, therefore, are likely to be underestimates, and can give only a rough idea of how much the students were helped by their familiarity with the subject area or topic. It would be interesting in a future study to compare such quantitative data with the results of test-taking introspections during which students specifically focused on the effect of background knowledge on their understanding. (See below for a suggestion of how this might be done.) For the present all that can be learnt from these figures is a) that approximately two thirds of the students who were familiar with a subject area or topic thought this helped them answer the questions; b) that on the whole, knowledge of the topic was considered to be more helpful than knowledge of the subject area (in six out of the nine cases a higher percentage of students found the topic more helpful than the subject area); and c) that (PST)Ship confirmed its status as a maverick subtest by having the lowest percentage of students to find prior knowledge a help. Whether this was caused by the choice of reading passage or the types of questions cannot be known, but the number of students concerned was very small: only 21 students (57% of 44) said they were helped by their familiarity with the subject area, and only 19 (55% of 34) with the topic.

These findings would be much enriched by students' introspective reports on their reading processes. Students would be asked to introspect aloud while taking reading modules within and outside their field of study. They would be asked, and perhaps trained, to concentrate on the effect that background knowledge had on their responses, and the data could be analysed according to a schema theoretic model. If Schank (1980) is right in believing that schema theory accounts for those high level mental processes of which we are aware, it should be possible for students to describe their activated schemata as they answer a test question. For example, if they were asked to introspect as they answered (PST)Fuel Question 12 ('What kind of fuel was predominantly used before coal?' [see Appendix 6.1]), high proficiency students might show that they were able to answer the question without using any of the schemata that was activated when they read the test question because the answer was transparent from the text – in Bachman's terms it was fully contextualised (Bachman 1990:132). For lower proficiency students for whom much of the vocabulary was new, schemata relating in particular to world economic history might lead them to the correct answer. Background knowledge might therefore help the weaker students to answer this question, but the background knowledge would be BSS- rather than PST- based in spite of the fact that this text comes from the PST module. If it proved to be the case that low proficiency students with a knowledge of world history were able to answer this question more easily than the scientists, this would help explain why this subtest was not subject specific.

Multiple Regression Analysis

Each of the above questions has given us extra information about the students' background knowledge. To see whether these appear to affect students' reading test performance, and at the same time to compare the relative contribution of background knowledge and language proficiency to students' test scores (Research Question 7), we can enter the variables into a multiple regression analysis, in which the dependent variable is students' scores on the module in their own subject area ('Own-Subject-Module'), and the independent variables are:

1. Scores on the Grammar test (Grammar)
2. School subject (School)
3. Reading background (Reading)
4. Familiarity of subject area (Area Familiarity)
5. Familiarity of topic (Topic Familiarity)
6. Students' academic subject area (BSS, LMS or PST)

The Grammar Test

The Grammar test, which was described in Chapter 6, tests reading among other skills, and its results correlate fairly highly with those of the reading modules. Table 10.5 shows that although the correlations between Grammar and the reading modules are lower than those between the reading modules themselves, the differences are not great. The correlation between M(BSS) and Grammar, for example, is .59, and between M(BSS) and M(PST) is .63.

Table 10.5

Correlations between Grammar and the Three Reading Modules

	M(BSS)	M(LMS)	M(PST)	Grammar
M(BSS)	-	.73 (310)	.63 (328)	.59 (634)
M(LMS)	.73 (310)	-	.71 (204)	.57 (506)
M(PST)	.63 (328)	.71 (204)	-	.51 (526)
Grammar	.59 (634)	.57 (506)	.51 (526)	-

Bracketed figures give the n. of students.

One of the reasons that the Grammar component was dropped from the IELTS battery was that in the main IELTS trials, where a different set of academic modules were used (see Chapters 4 and 5), the correlations between the Grammar test and the reading modules were so similar to the modules' correlations with each other that the test was considered to be assessing similar skills to those tested in the reading modules (see Alderson 1993a). It is reasonable to suppose, therefore, that the Grammar test can provide a useful measure of the students' L2 language ability for the purposes of this study.

School Subject and Background Reading

Since regression analysis requires all categorial data to be in binary form, the 'School Subject' and 'Background Reading' variables were each recoded as 1 and 0. For School Subject, students were assigned a 1 if they had studied their present academic subject in their last two years at school, and 0 if they had not. For Background Reading they were assigned a 1 or a 0 according to whether or not they read books and articles in their own subject area, that is, the area related to the score on their own-subject reading module.

Area and Topic Familiarity

Each of the questions relating to Area and Topic Familiarity referred only to one of the three areas or topics in a module, so the three answers relating to each module had to be pooled to make one overall measure of the familiarity of a complete module. After being recoded into binary form – 1 for familiar, and 0 for not familiar – the students' three ratings for each module were summed, and once again recoded (3 or 2 = 1; 1 or 0 = 0). In the process of being conflated these variables lose much of their value and give only a general idea of the familiarity of the subject matter or topic of each module.

Academic Subject Area

Since account must be taken of the level of language proficiency of each of the three groups and the difficulty of the three modules, two dummy variables were introduced into the equation, for each of which students were given a dummy score of 1 if they were in a given academic subject area, and 0 if they were not. Only two variables represent the three subject areas, since if all three were included it would always be possible to deduce from a student's first two dummy scores what the third one would be. For example, a student with 0s for BSS and LMS would inevitably have a 1 for PST, as each student has to be in one of the three subject areas. The third variable, therefore, would be an 'exact linear function' (Lewis-Beck 1980:68) of the other two. The two dummy variables represent LMS and PST; BSS is excluded and 'serves as a reference group against which comparisons can be made' (Schroeder, Sjoquist and Stephan 1986:57).

The Regression Equation

The model for the regression equation is as follows:

Own-Subject-Module = constant + b1 x Grammar + b2 x LMS + b3 x PST + b4 x other (BSS) + + b5 x School + b6 x Reading + b7 x Area Familiarity + b8 x Topic Familiarity

Stepwise multiple regression was calculated using SPSS. The data fitted the model satisfactorily as there were no more than 10 out of 842 outlying cases, the residuals were normally distributed and cumulative distributions of the observed and the expected residuals were almost identical.

Table 10.6 gives the intercorrelations of the variables included in the regression analysis, and from these it will be seen that the correlation between Grammar and 'Own-Subject-Module' is .66; this is the only correlation between the dependent and the independent variables that is higher than .17. Most of the independent variables show little or no intercorrelation, but Area Familiarity and Topic Familiarity have a correlation of .46, which is quite high, and might have an effect on the regression results – I will discuss this overleaf. Table 10.7 gives the estimated value of the slopes (b), standardised regression coefficients (ß), t ratios, p values and the proportion of variance accounted for (r^2).

Table 10.6

Intercorrelations between the Variables (n = 515)

	Own-Sub.	Gram.	LMS	PST	School	Read.	Area	Topic
Own-Subj.-Mod.	-	.66	.16	-.17	-.04	.08	.13	-.04
Grammar	.66	-	.08	-.14	-.08	.03	.05	-.10
LMS	.16	.08	-	-.31	-.23	.04	.10	-.01
PST	-.17	-.14	-.31	-	.08	-.27	.02	.20
School	-.04	-.08	-.23	.08	-	.05	-.02	.01
Reading	.08	.03	.04	-.27	.05	-	-.01	.01
Area Fam.	.13	.05	.10	.02	-.02	-.01	-	.46
Topic Fam.	-.01	-.10	-.01	.20	.01	.01	.46	-

Table 10.7

Multiple Regression Analysis – Complete Modules

	b	ß	t	p	r^2
Grammar	.79	.65	19.73	<.0001	.44
LMS	4.75	.09	2.87	.01	.45
Area Familiarity	3.63	.09	2.63	.01	.45
Constant	-5.76	.09	-1.98	.05	
School Subject			1.13	NS	
PST			-1.87	NS	
Reading			1.89	NS	
Topic Familiarity			.76	NS	

Those variables with p values greater than .05 are not included in the equation as the value of b does not add significantly to the slope of the regression line. From the above table it will be seen that only three variables were significantly related to the students' scores. Grammar, not surprisingly, had far the strongest relationship, since Grammar scores correlate highly with students' scores on the reading modules. LMS is distinguished from BSS and PST because LMS students have higher mean scores on their own reading module than BSS and PST students have on theirs (see Chapter 7). How this fits into the regression equation is most easily understood if we imagine predicting an 'Own-Subject-Module' score from the regression equation. If we incorporate the values from Table 10.7 into the multiple regression equation, it looks like this:

$$\text{Own-Subject-Module Score} = -5.76 + 0.79 \times \text{Grammar} + \begin{Bmatrix} 4.75 \text{ (if LMS)} \\ 0 \text{ (if not LMS)} \end{Bmatrix} + \begin{Bmatrix} 3.63 \text{ (if Area Fam. is yes)} \\ 0 \quad \text{(if Area Fam. is no)} \end{Bmatrix}$$

This means that if the student whose score we are predicting comes from an LMS background, then 4.75 should be added to the equation, but if he or she comes from BSS or PST, nothing should be added. Similarly if the student is familiar with the subject area of the reading module, then 3.63 should be added (i.e. 3.63 x Area Fam. = 3.63 x 1 = 3.63), but otherwise not.

In this equation School Subject, PST, Background Reading and Topic Familiarity are omitted because they do not have significant b values. They would not therefore significantly increase the amount of variance accounted for. It is not surprising that School Subject is excluded. Most students studied a wide range of school subjects in their last two years, and many of them studied all three subject areas (see Table 10.8). It is possible, too, that Background Reading does not have a significant b value because so many science students read in more than one subject area; however, see Table 10.9 below.

Table 10.8

Subject Areas Studied during Last Two Years of School

Subject area	N
BSS	110
LMS	3
PST	15
BSS + LMS	49
BSS + PST	131
LMS + PST	47
BSS + LMS + PST	454

The higher the correlation between two independent variables, the less clear it is how much they individually contribute to a regression equation, and it may be for that reason that Area Familiarity has a significant t ratio, but that Topic Familiarity does not. On the other hand, knowledge of topic area may not necessarily help reading comprehension. It is possible that when students are already familiar with a topic, their knowledge, especially if it is different from what is in the reading passage, interferes with their understanding of the text (see Lipson 1984). Or it may be, as Carrell (1983) found, that a familiar topic is sometimes less 'salient' than an unfamiliar one (see Chapter 3). In her case she found that less salient texts were *remembered* less well than salient ones, but it is possible that less salient texts are also sometimes less well *understood.*

The total variance (R^2) accounted for by the above equation is only .45, which is not unexpected as not only are there many other aspects of background knowledge which have not been included in the equation, but other aspects of reading skill have also been neglected. For example no account has been taken of students' reading proficiency in their first language, and, as Bossers (1992) shows, this is related to reading proficiency in the L2. It seems from the equation that the students' performance in their own reading module is chiefly affected by level of language ability, and that background knowledge plays a very small part. However, this regression analysis was based on scores on the complete academic modules, which, as we have already seen, include some non-specific subtests. The analysis was therefore carried out again, with a revised dependent variable based on students' scores on the revised modules from which non-specific subtests had been omitted.

Table 10.9 lists the variables which help account for the variance of the scores on the revised modules.

Table 10.9

Multiple Regression Analysis – Revised Modules

	b	ß	t	p	r^2
Grammar	.73	.55	15.59	<.0001	.26
PST Subject Area	17.00	.35	9.64	<.0001	.37
Background Reading	7.17	.07	1.97	.05	.37
Area Familiarity	3.25	.97	1.97	.05	.38

This time it can be seen that although language proficiency makes the greatest contribution to the variance, its effect is less marked: Grammar alone only accounts for 26% of the variance. The effect of subject area is stronger than when scores from the complete modules were used. It is possible that as tests become more specific, background knowledge becomes more important and language proficiency less. If this is the case, any research into the comparative effect of background knowledge and language proficiency on reading comprehension should take account of the subject specificity of the reading passages. I shall return to this in Chapter 11.

It is also possible that the PST subject area stands out from BSS and LMS because the revised PST module, which consists of only one subtest, (PST)Sun, is much easier than the BSS and LMS modules (see Appendix 10.1). PST students therefore score very highly in their own subject module.

Interestingly, now that the modules are more subject specific, Background Reading is shown to be related to scores on the Own-Subject-Module. Its omission from the previous analyses, therefore, may not have been only due to the breadth of the science students' reading, but may also have been due to the fact that the complete reading modules were not sufficiently subject specific. It now seems, as might be expected, that background reading is an important source of subject specific background knowledge.

Discussion

The main findings from these multiple regression analyses are that:

1 Language proficiency accounted for more of the test performance variance than did any of the measures of background knowledge. This appears to go against the findings of Johnson (1981) and Floyd and Carrell (1987), who found that background knowledge had more effect on test scores than did the syntactic complexity of the passages. However, there is no direct relationship between syntactic complexity and level of language proficiency, and in addition, the different effects of background knowledge and syntactic complexity must vary according to the level of the students, the amount of background knowledge required in the passage, and the level of difficulty of the texts. I shall return to this point in Chapter 11.

2 Although Background Reading did not account for any of the variance when reading performance was equated with scores on the complete reading modules, once the non-specific subtests were omitted it was more closely related to reading test performance than was Area Familiarity. Since academic reading must almost always increase one's background knowledge, this finding perhaps tells us more about the lack of specificity of the complete reading modules than anything very startling about the effect of Background Reading.

3 What may be an interesting discovery is the fact that Topic Familiarity seemed to have less effect on test performance than did knowledge of the subject area of a reading passage. This is only a tentative finding, as the conflation of three different ratings for each module has lessened the validity of the ratings, but it would be interesting to investigate this in more detail. Again I shall discuss this further in the final chapter.

4 Much of the Own-Subject-Module variance is still unaccounted for. It is probable that if students were asked to demonstrate their background knowledge by taking tests of their subject knowledge, the relationship between test performance and subject knowledge might become clearer. It would also be interesting to include a measure of L1 reading proficiency since this might correlate with the reading test score, and might account for a substantial part of the variance.

Language Proficiency and the Effect of Background Knowledge

The question of whether L2 readers vary in their use of top-down skills according to their level of proficiency in the target language has evoked much interest in recent years. In the research described in Chapter 3 it was generally found that advanced language learners appeared to make more use of top-down skills than did lower level students. Coady (1979) claimed that as students became more proficient at a second language, they came to place more emphasis on cognitive strategies than decoding skills. Cziko (1980) and McLeod and McLaughlin (1986) thought that advanced students used interactive skills, while lower ability students depended on bottom-up ones. Clarke (1980) concluded that low level learners could not decode enough graphic and lexical symbols to be able to bring top-down processing systems to bear, and Hudson (1982), while agreeing with Clarke, described how low level students could be taught to make more use of top-down processing. The only counter evidence is offered by Wolff (1987), who, looking at the question of text difficulty rather than language proficiency, found that the more difficult the text, the more subjects depended on top-down processes. Wolff's findings are intuitively satisfying, since anyone struggling to understand a text might be expected to use all available top-down and bottom-up clues. However, this may depend on the sophistication of the reader, and it may be, as Hudson (1982) shows, that some low level readers have to be taught to make use of contextual clues or background knowledge.

One of the problems with attempts to answer questions related to the level of students' language proficiency is, as I pointed out in Chapter 3, that different researchers' ideas of the meaning of 'low' and 'high' level language proficiency vary, and it is quite possible that one researcher's 'high' may coincide with another researcher's 'low'. My own study is linked to the proficiency levels required for university entrance, and it can be assumed that students with high scores on the reading modules had a very advanced knowledge of English. It can also be assumed that although the test results reveal a wide range of language ability, none of the students were beginners, because even the weakest ones were at a level suitable for university pre-sessional courses. Few students, if any, are likely to have problems with basic graphemic and lexical decoding. There is not, therefore, a full range of language proficiency levels in the study, and differences between the top-down skills of high and low level students may not be fully revealed.

In Chapters 2 and 3 it became clear that background knowledge facilitated reading comprehension, and it therefore seems probable that students with weak L2 skills will lean more heavily on background knowledge to help them to interpret reading passages. Wolff's findings confirm this. In spite of the fact, therefore, that most of the research described above showed that students make increasing use of background knowledge as they become more familiar with a second language, I shall hypothesise that low level students in my study will make more use of top-down reading strategies than will the higher level ones because their limited vocabulary and syntactic knowledge will lead them to rely on guesswork and background knowledge. I therefore expect that a subject area effect will be more apparent between low level than between high level students. I shall use two methods of assessing this: firstly repeated measures analysis of variance for students at three levels of English proficiency, and secondly a classical analysis of variance based on a comparison of the mean differences between students' scores on their two reading modules.

Repeated Measures Analysis of Variance

In order to get an initial idea of how much effect students' level of language proficiency had on their ability to use background knowledge when reading, students were divided into three groups according to their scores on the Grammar test: below 60%, 60%–80%, and above 80%. (No students had scores of exactly 60% or 80%.) Repeated measures analyses of variance similar to those described in Chapter 7 were calculated separately for the three groups. Table 10.10 gives the numbers of students in each test pairing, and the p values for the subject area effects. Since the BSS and the LMS modules contain the highest proportion of subtests which appear to be subject specific (see Table 8.5), the results of the BSS/LMS pairing will be considered first. For the 33 BSS and the 19 LMS students who had Grammar scores of less than 60%, the effect of field of study is not significant. However, for students between 60% and 80% the effect is significant at $p = <.001$. This result is strikingly different from the previous one and suggests that lower level students were not using their subject knowledge, whereas higher level students were.

Table 10.10

Repeated Measures Analysis of Variance – Subject Area Effect at Three Levels of Proficiency – Complete Modules

Grammar Score	BSS/LMS		BSS/PST		LMS/PST	
	n	p	n	p	n	p
<60%	33/19	NS	45/36	.05	21/30	NS
>60% – <80%	94/36	<.001	75/46	NS	25/32	NS
>80%	48/21	.05	46/23	NS	20/26	NS

The low level students were perhaps trying so hard to make sense of the grammar and vocabulary in the reading passages that they were not able to bring background knowledge to bear, and it is also possible that their English was so poor that they could not yet distinguish between specialist and non-specialist expressions. Further up the proficiency scale, the students had enough English to read passages in their own subject area competently, although they were not yet able to cope so well with texts in other areas. From this result it looks as if my hypothesis should be rejected. However, the effect of academic discipline appears less marked for the students with scores above 80% ($p = .05$). This drop in the level of significance may be caused by the smaller numbers of students in the top groups, or it may be that students with this high level of language proficiency are able to compensate for any lack of background knowledge. It may be, therefore, that acceptance or rejection of my hypothesis depends on the level of ability of the students being tested.

The BSS/PST results are diametrically opposed to those of BSS/LMS: the lowest level students appear to make more use of subject area knowledge than either of the upper groups. This surprising result may be, at least in part, due to the lack of specificity of two of the three PST subtests. (Note that the LMS/PST pairing shows no significant effects at any of the three levels.) These results should therefore be compared with those of the revised modules from which the non-specific subtests have been removed.

Revised Modules

Because of the differences in the difficulty levels of the revised modules, there are highly significant test effects (p = <.001) for all but one pair of tests, and this test disparity probably confounds the results. Fortunately, however, M(BSS) is unchanged since all its three subtests are subject specific, and the revised M(LMS) still has two out of its four original subtests, so I shall report the results of this pairing in the most detail.

Table 10.11 gives the results of the repeated measures analysis of variance for the revised modules, and it will be seen that the BSS/LMS results are similar to those of the complete modules (see Table 10.10), but that the p value for students with Grammar scores above 80% is now <.001. This serves to confirm the finding that the students appear to make more use of their background knowledge once they achieve Grammar scores of at least 60%. It does not support my contention that further up the proficiency scale background knowledge becomes less important.

Table 10.11

Repeated Measures Analysis of Variance – Subject Area Effect at Three Levels of Proficiency – Revised Modules

Grammar Score	BSS/LMS		BSS/PST		LMS/PST	
	n	p	n	p	n	p
<60%	33/19	NS	45/36	.001	21/30	NS
>60% – <80%	94/39	<.001	75/46	.01	25/32	NS
>80%	48/22	<.001	46/23	.05	20/26	.05

This time the LMS/PST pairing, too, shows an increase in the effect of subject area, but the increase is not great, and is only significant for students with Grammar scores of above 80%. The revised M(PST) module consists solely of the Sun subtest, and since subject specialists considered that the Sun reading passage was suitable for both LMS and PST students (see Chapter 8), it is perhaps surprising that *any* of the results are significant. The BSS/PST pairing again shows a decreasing subject area effect as the students become more proficient at English, but there is now a significant subject area effect at all three levels. Because of the gross disparity between the difficulty levels of these two modules, it is not clear how this should be interpreted.

Difference Scores

Another way of seeing whether students' use of field specific knowledge varies according to their level of language ability is to find the difference between each person's scores on the two academic modules, and to compare this to the student's level of language proficiency. 'Level of language proficiency' is again represented by scores on the Grammar test. Since we expect students to have a higher score on the module within their own subject area than on the one outside it, each student's score on the 'Not-Own-Subject' module is subtracted from the score on the 'Own-Subject-Module'. For example, a BSS student with a score of 60% on M(BSS) and 55% on M(LMS) would have a difference score of 60 - 55 = +5. Classical experimental analysis of variance (SPSS 1990:65) can then be used, based on the equation:

difference = own subj.area effect + not-own subj.area effect + Grammar score + interactions

Since an initial attempt to plot a regression line of Difference on Grammar score showed that there was no consistent effect and therefore no linear relationship between the two, Grammar scores were again divided into three levels: below 60%, between 60% and 80%, and above 80%.

Table 10.12 shows that, as might be expected from the earlier repeated measures analysis of variance results in Chapter 7, there are no significant main effects. Nor are there for two of the two-way interactions. There is, however, an interaction between 'Own Subject-Area' and Grammar ($p = .01$). The students' difference scores are therefore significantly affected by the interaction between their own subject area and their level of language proficiency. Figure 10.1. represents the results graphically, with the difference scores plotted against the grammar scores (the Xs represent BSS students, the triangles LMS, and the stars PST).

The fitted values lines, which were plotted using the GLIM program for fitting generalised linear models (see Francis, Green and Payne 1993), show that for LMS students with Grammar scores of less than 60%, the fitted values line is close to zero, and that there is therefore no difference between students' LMS scores and their scores on either of the other two academic modules. However, LMS students with scores of above 60% did significantly better ($p = <.001$) at the test in their own subject area.

Table 10.12

Analysis of Variance of Mean Differences in Scores According to Level of Language Proficiency – Complete Modules

Source	SS	df	MS	F-ratio	p
Main Effects					
Own-Subject-Area	968	2	484	2.397	NS
Not-Own-Subj.Area	73	2	36.5	0.18	NS
Grammar	583	2	291.5	1.444	NS
2-Way Interactions					
Own Subj.Area by Not-Own Subj.Area	214	1	214	1.06	NS
Own Subj.Area by Grammar	3119	4	779.75	3.862	.01
Not-Own-Subj.Area by Grammar	1014	4	253.5	1.255	NS
Residual	118542	587	201.9		
Total SS	124513.4				

The BSS results are similar. These results, therefore, confirm the earlier finding that low proficiency BSS and LMS students do not seem to make use of their subject area knowledge to help them understand written text, whereas higher proficiency students do. The PST students' results, once again, do not confirm this trend, but this result may be at least partly due to the unspecificity of the PST module.

Table 10.13 gives the analysis of variance results once the non-specific subtests have been omitted. The effect of the interaction between 'Own-Subject-Area' and Grammar is similar to what it was in Table 10.12, but inevitably, the 'Own Subject Area' and 'Not-Own Subject-Area' main effects are highly significant since there is such a disparity in the difficulty levels of the revised modules. Because of these significant differences, a separate fitted values line is required to demonstrate the differences between students' scores on their Own-Subject-Area module, and that of each of the 'Not-Own-Subject' modules. The graph, therefore (see Figure 10.2), now consists of six, rather than three, fitted values lines – one for each pair of tests. Each of these lines names the students' own subject area first, so that the LMS–BSS line, for example, shows the mean difference between LMS students' scores on the LMS and BSS modules.

It will be seen in this graph that the three pairs of lines each follow a distinctive pattern, but, because of the magnitude of the difference in scores between each pair of tests, start from different points on the y axis. For example, for students with scores of less than 60%, the BSS–LMS line and the BSS–PST lines show mean differences of +20 and -11 respectively. For these students the BSS module is easier than the revised M(LMS) (difference = +20), and more difficult than the revised M(PST) (difference = -11). What is interesting is that both fitted values lines rise in steady steps through the three language proficiency levels, so that as the BSS students' language proficiency improves, the difference between the scores on the two tests becomes greater for BSS/LMS, and smaller for BSS/PST.

Figure 10.1

Difference Scores against Grammar Scores – Complete Modules

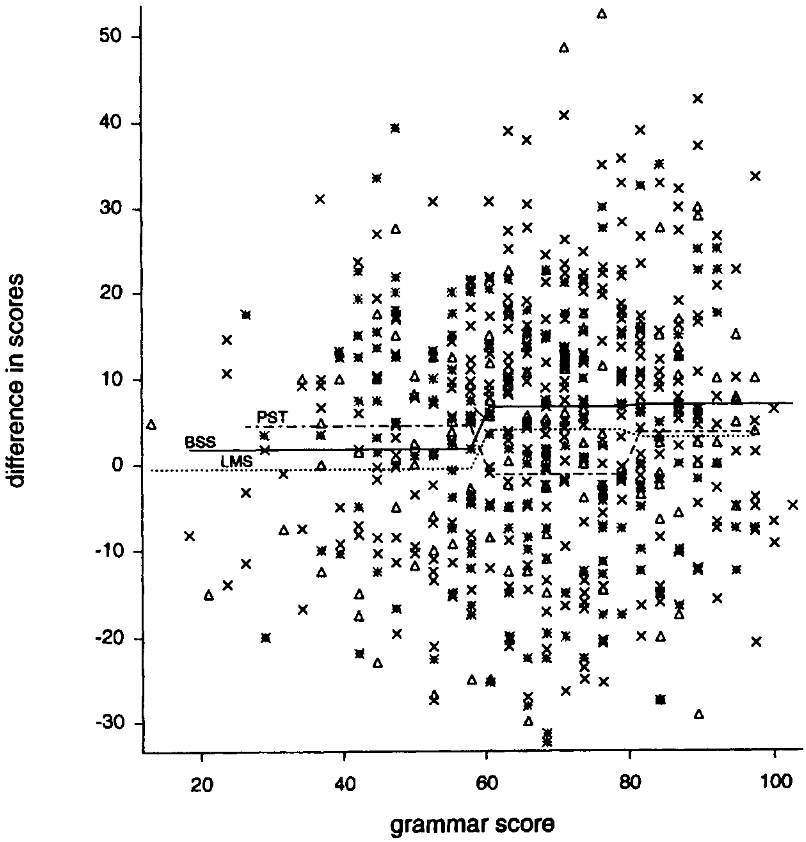


Table 10.13

Analysis of Variance of Mean Differences in Scores According to Level of Language Proficiency – Revised Modules

Source	SS	df	MS	F-ratio	p
Main Effects					
Own-Subject-Area	47314	2	23657	44.2	<.001
Not-Own-Subj.Area	92866	2	46433	86.76	<.001
Grammar	3788	2	1894	3.539	.05
2-Way Interactions					
Own-Subj.Area by Not-Own Subj.Area	182	1	182.0	.3401	NS
Own Subj.Area by Grammar	8481	4	2120	3.962	.01
Not Own Subj.Area by Grammar	5268	4	1317	2.461	.05
Residual	314155	587	535.2		
Total	472054				

The effect of subject area therefore increases with the higher levels of language proficiency. This agrees with the repeated measures results reported above. The two LMS lines show a similar tendency, although this time there is little change between students with scores of less than 60% and between 60% and 80%. There is, however, a sharp rise for students with Grammar scores above 80%.

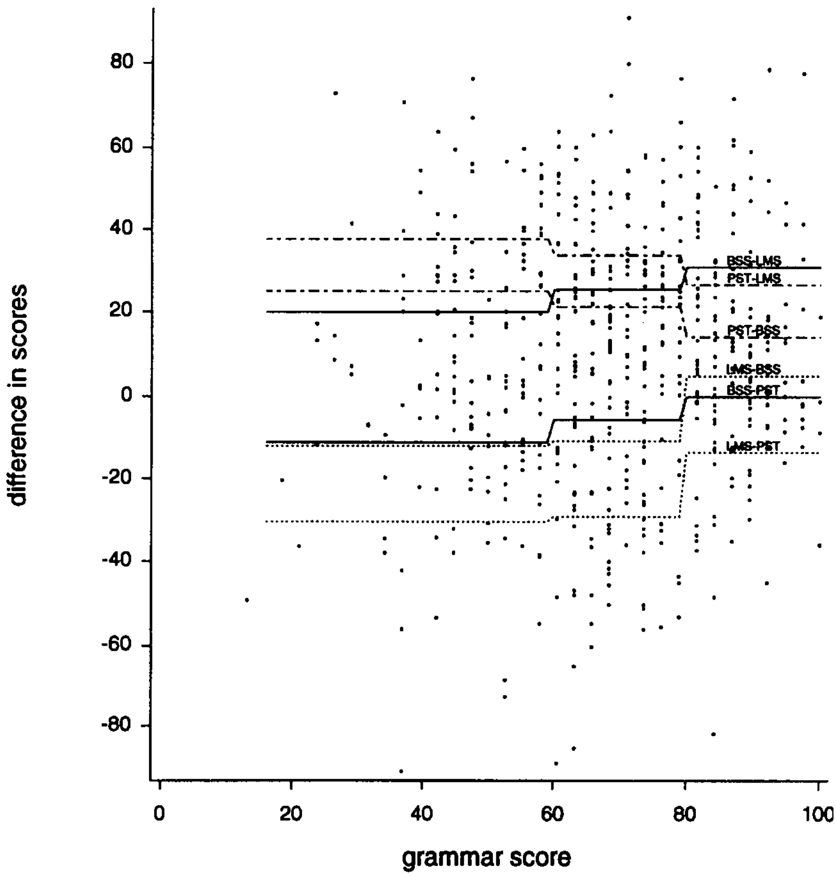
Not surprisingly, the two PST fitted values lines once again go against the general trend, probably because of the lack of specificity of (PST)Sun, the only remaining subtest in the revised M(PST).

Discussion

The value of those of the above analyses which were based on the complete academic modules is limited because of the lack of subject specificity of some of the reading sub-tests, and the results based on selected subtests were confounded by the variation in the difficulty of the revised modules. Both sets of results may therefore be on the conservative side. In spite of this, however, the results from the most subject specific of the three modules – BSS and LMS – seem to confirm that L2 readers make more use of background knowledge as their language proficiency improves. However, the fact that it was impossible to fit a straight regression line suggests that there is no steady increase in the use of background knowledge as language ability improves, but that students need to reach a certain level of proficiency before they can start to make use of this knowledge. At low levels of language proficiency they may be too concerned with bottom-up skills to make full use of their top-down skills, but at some stage they are able to decode the L2 sufficiently easily to be able to do so. Whether

Figure. 10.2

Difference Scores against Grammar Scores – Revised Modules



there is a sudden moment when this becomes possible, or whether there is a gradual change once a certain level of language learning has been reached is not yet clear, but in either case this would suggest that, as researchers such as Clarke (1980), Alderson (1984) and Laufer and Sim (1985) propose, there is a threshold level below which students are unable to make use of their background knowledge and L1 reading skills. Presumably, at some higher stage of ability, beyond the levels of proficiency included in this data set, background knowledge again becomes less important, as students learn to make maximum use of all the linguistic clues in a text. Many highly educated native speakers, for example, might be expected to achieve almost full marks on all three tests, regardless of their academic subject area.

Conclusion

Since this chapter has contained a variety of disparate studies I shall now briefly review the findings.

The chapter started with three repeated measures analyses of variance which were similar to those described in Chapter 7, and which looked at the effect of academic subject area on the three pairings, BSS/LMS, BSS/PST and LMS/PST. This time, however, the analyses were based on the students' scores once the non-specific subtests had been omitted from the reading modules. The revised modules varied widely in length and difficulty and were not therefore ideal instruments for this study. The subject effects for the BSS/LMS and the BSS/PST pairings were highly significant ($p = <.001$), so students in these subject areas did perform better at the reading test in their own subject area. However, this time there was no significant effect for the LMS/PST pairing, although there had been one ($p = .05$) previously. This result was caused either by the gross disparity in the difficulty of the two revised modules (see Table 10.1), or because the two science subject areas overlap. Although the only subtest to remain in the revised PST modules, (PST)Sun, is subject specific according to Table 8.5, it was considered by many subject specialists to be suitable for both LMS and PST students (see Chapter 8).

A simple crosstabulation between students' subject area and their background reading revealed that although 36% of the BSS students read only BSS material, only 8% and 9% of LMS and PST respectively restricted their reading to their own subject area. 87% of the science students read BSS texts. This suggests that most science students are sufficiently familiar with BSS materials to be able to cope with a BSS based test, and argues against the provision of science based ESAP tests.

An investigation into whether students found knowledge of the topic or the subject area of a reading passage more helpful in answering the questions showed that more found knowledge of the topic helpful. This is contrary to the regression analysis findings which are reported next.

A multiple regression analysis of students' scores on the complete test module showed that the major contribution to their test scores appeared to be level of English language proficiency. The students' field of study and their familiarity with the subject area were also significantly related to test scores, but much less strongly. A similar analysis based on the revised modules again emphasised the importance of language proficiency, and added background reading to the regression equation. Students' field of study, background reading, and familiarity with text subject area were all shown to have an effect on the variance of the test scores, but familiarity with the topic of a text did not have a significant relationship, although the students had said they found it more helpful for answering the test questions than knowledge of the subject area (Table 10.4). The absence of a significant topic effect may be because the two variables, subject area familiarity and topic familiarity, correlated with each other.

The final section of this chapter looked at whether the effect of background knowledge on reading comprehension varied according to students' level of English proficiency. Two methods of analysis were used. The first was repeated measures analyses of variance, with students divided into three levels according to their scores on the Grammar test. The second divided the students into the same three levels and looked at the differences between students' scores on the two modules, one within and one outside the students' subject area, to see whether this difference varied according to level of proficiency. Both studies were carried out on the results of the complete and the revised modules. Only the results on the BSS/LMS pairing are reported here, as there are problems with the other two pairings, since the complete PST module has only limited subject specificity, and the revised version is very short and very easy. The BSS/LMS results on both the complete and the revised modules showed that students did not appear to be affected by background knowledge until they achieved scores of over 60% on the Grammar test. The results of the complete modules then showed that students with Grammar scores of over 80% appeared to make less use of background knowledge than the intermediate students, but there was no such difference for the revised modules. It seems clear that the low proficiency students could not take advantage of their background knowledge, but from the results it is not clear whether as they became more proficient they were able to use linguistic skills to compensate for any lack in background knowledge.

I shall discuss these points more fully in the next chapter, when I discuss the different findings that have been reported in this book.

11 Conclusions

The two main purposes of this study were to investigate the value of giving academic students reading modules in different subject areas, and to find out more about the effect of background knowledge on reading. I shall start this concluding chapter by listing each of the research questions which were introduced in Chapter 6, and commenting briefly on the results. I shall then discuss the implications of these findings in more detail, and relate them to the broader issues of EAP proficiency testing. I shall conclude by briefly describing the areas in which I think more research is needed.

The Research Questions

Research Question 1

Do students in the three broad subject areas of BSS, LMS and PST achieve significantly higher scores in a reading test within their own subject area than they do in a test in one of the other two subject areas?

In the main study, repeated measures analysis of variance showed that students achieved significantly higher scores on the module in their own subject area than on the module outside it. ('Own subject area' refers to the students' present or past subject area, and not to their future one.) At first glance, therefore, it seems that the simple answer to this research question is 'Yes'. However, in the pilot study, where students took a different set of reading modules from those in the main study, there were no significant differences between the Whole Sample's scores on modules within and outside their own subject area. This difference between the two sets of results suggested that the subtests in the two test versions varied in their subject specificity. This was confirmed when the subtests in the main study were analysed individually (see Research Question 4 below). Four of the ten subtests were either too general or too specific for their designated subject area. When these inappropriate subtests were removed from the modules, the subject area effects for the BSS/LMS and BSS/PST pairings were, not surprisingly, stronger, and reinforced the assumption that students do better in a test in their own subject area. However, there was no longer a significant subject area effect for LMS/PST, probably because the only remaining text in the revised PST module was equally suitable for PST and LMS students.

From this, it is clear that there is no simple answer to Research Question 1: the results depend on the specificity of the individual subtests. I shall return to this question of subject specificity below.

In another approach to Research Question 1, Rasch analysis was carried out to see whether any items or sets of items showed evidence of bias against students who took modules which were not in their own subject area. For this study the students were divided into two groups only – BSS and Science. The only sets of items which showed consistent evidence of bias came from the two M(LMS) subtests which were later classified as highly specific (see Table 8.5). This suggests that only if modules contain highly specific subtests will students be at a disadvantage if they take a reading module in an inappropriate subject area. However, bias analyses are designed to identify variations in item performance caused by gross differences in the groups being compared, and students in the three broad subject areas of BSS, LMS and PST may not be sufficiently disparate. This possible lack of disparity among the subject areas raises important questions in this research, and will be discussed further below.

A study into whether students were able to complete the appropriate module more quickly than the inappropriate one (Chapter 7) showed that M(LMS), which contained the two most highly specific subtests, was finished by proportionately more LMS than BSS or PST students. However, LMS and PST students both finished M(BSS) as quickly as the BSS students. This suggests that science students may be able to process BSS tests as quickly as BSS students, whereas BSS students are unable to process LMS tests as quickly as LMS students. The PST results were inconclusive.

From these three approaches to Research Question 1, it appears that if subtests are sufficiently subject specific, students will do better at tests in their own subject area. However, subtests vary widely in their specificity. This variation is not necessarily obvious to test writers and users, and may well explain some of the contradictory results of earlier research into the effect of background knowledge on reading comprehension (see Chapter 1). This is discussed further below.

Research Question 2

> *Do students in the three broad subject areas of BSS, LMS and PST achieve significantly higher scores in a reading test within their future subject area than they do in a test in one of the other two subject areas?*

For this question, students who were in the process of changing their field of study were reclassified according to their future subject area (see Chapter 7). Since only a small proportion of students were making this move, no great changes in the results were expected, and indeed the results of the repeated measures analyses of variance were similar to those described above, except that in all cases the subject area effect was slightly reduced. This reduction suggests that students who were changing subject area were at a disadvantage when they took the module in their future rather than their past subject area, and leads me to query the validity of the IELTS examiners' decision to allocate students to subject modules according to their future field of study. Native speakers are expected to be able to learn the conventions and style of their new subject area once they have started their new course, and it is reasonable to expect L2 students to have a high enough level of English to do the same. The present practice ignores current research into the effect of background knowledge on reading comprehension (see Chapters 1, 2 and 3). If the examiners' aim is to encourage students to give as accurate a demonstration of their proficiency as is possible, then they should give students the test which best facilitates this, that is, the test in their own subject area. Otherwise they may find that students taking a test in an unfamiliar subject area may not do themselves justice, and may wrongly be excluded from a university course because of transitory ignorance of content and form which might be learnt in a few weeks.

Research Question 3

> *Do either undergraduates or postgraduates in the three broad subject areas of BSS, LMS and PST achieve significantly higher scores in a reading test within their own subject area than they do in a test in one of the other two subject areas? If there is a significant difference, on which of the two groups is the subject effect the stronger?*

The results of the repeated measures analyses of variance which were run separately for undergraduates and postgraduates are reported in Chapter 7. My expectation was that since most future undergraduates had not yet started to specialise in one discipline, and since they might not have been exposed to much academic writing, they would be less affected than postgraduates by the subject area of the reading modules. On the whole this expectation was confirmed. There was no significant subject area effect for the LMS/PST and BSS/PST undergraduate pairings, whilst for the postgraduate pairings the effects were significant. However, for the BSS/LMS pairing the subject area effect was significant for both undergraduates and postgraduates. M(LMS) contains the two highly specific LMS subtests which are based on scientific concepts which might be beyond the scope of many elementary science courses. It may be that these two subtests would put non-scientific undergraduates and postgraduates at an equal disadvantage; it is likely that once subtests reach a certain level of specificity, undergraduates too are affected by the subject area of the texts.

Research Question 4

Are the reading passages in the three reading modules specific to the appropriate subject area?

I have already, in the above sections, anticipated the answer to this question by referring to the variation in the specificity of the reading subtests. My suspicion that subtests varied in specificity was confirmed when I carried out repeated measures analyses of variance on pairs of subtests from different modules. The subject area effects of these pairs of subtests varied widely and ranged from the pairings with (LMS)Nitro which all showed significant effects (see Table 7.8), to those with (PST)Ship which only showed a significant effect when the subtest was paired with (LMS)Nitro. Although it was sometimes difficult to know whether the variation in subject area effect was due to one or both the subtests in a pair, it was generally possible to infer a subtest's level of specificity from a study of the whole set of pairings.

An investigation into the modules' Test Method Facets (Chapter 9) revealed no variation in the specificity of the test items, and it therefore seemed likely that any differences in the subtests' specificity were due to variations in the reading passages. Some of the reasons for this variation are given in response to Research Question 5 below.

Research Question 5

Is it possible to identify some characteristics of the reading passages which make them either more or less specific to their chosen subject areas?

The evidence from Chapters 8 and 9 shows that it is possible to identify some of the characteristics which lead to passages being more or less subject specific, but that these characteristics are not always immediately obvious. In Chapter 9 I showed that the source of a reading passage does not necessarily determine its level of specificity. The two texts which came from the most academic of the sources used for the reading modules turned out to be 'general' texts, but two passages from science popularisation journals were 'highly specific'. It seems in this case that it was the rhetorical function of the passages rather than the sources of the texts which affected their specificity: in the small selection of passages under discussion, the introductions to research papers were not subject specific, whereas the descriptions of processes were.

The subject specificity of a text is also likely to depend on the extent to which comprehension of that text requires knowledge of subject specific concepts which are not explained in the text. Unfortunately the TMC raters were not able to agree on the 'degree of contextualisation: topic specificity' of each text (see Chapter 9), so this could not be confirmed here.

As a check on the appropriacy of the passages for candidates in the relevant fields of study, students were asked to comment on the familiarity of the subject area of each text. On the whole these comments supported the results of the repeated analyses of variance (Chapter 7): the subject area of subject specific passages was more familiar to students within the subject area than to those outside it (see Chapter 8), and the subject area of the two highly specific texts became progressively less familiar to students the further their field of study was from the subject area. However, one text, which, against expectation, appeared to be general rather than specific according to the repeated measures analysis of variance results, was 'too specific' according to the students: it was on such a specialised aspect of engineering that only 10% of the students in the appropriate field of study said they were familiar with the subject area.

The students were also asked to say whether they were familiar with the text *topics*. The responses were similar to those relating to the texts' subject area, but students were in general less familiar with the topics than they were with the subject areas. However, in the case of two PST texts this trend was reversed: a higher proportion of BSS students were familiar with the topics than with the subject areas (see Appendices 8.1 and 8.2). This suggests that the topics were familiar to many people who were not familiar with the academic subject areas, and the fact that most BSS students said they were familiar with the topic of (PST)Fuel emphasises the lack of subject specificity of this supposedly PST text.

As a further check on the specificity of the passages, university lecturers were asked to comment on the appropriacy of the texts, and on the likelihood of the topics being familiar to their students (see Chapter 8). Their comments on the appropriacy of the texts were sometimes surprising (as in the case of a BSS text which was considered by the engineers to be appropriate for engineering students), and once again showed that some texts were not appropriate for the designated modules. Nevertheless, on the whole the lecturers' comments on the probable familiarity of the topics to their students agreed with those of the students.

Research Question 6

> *Do students with some familiarity with the subject area of an academic reading test, whether or not it is obtained from formal study in that area, achieve significantly higher scores on the test than students who do not have that knowledge?*

Students were asked in the questionnaire to say whether they read texts in disciplines both within and outside their field of academic study (see Chapter 10). I used these responses for a reclassification of the students according to both their field of study and their background reading so that the students could be grouped according to a more accurate index of their background knowledge. However, so many students read in all three subject areas, and so many science students read BSS materials, that the boundaries of the students' background knowledge proved to be too fuzzy for repeated measures analyses of variance to be calculated. It was not possible, therefore, to answer this question directly, but the research did show that classifying students by subject area alone was not satisfactory, and that science students, in particular, read widely outside their own subject area.

Research Question 7

Which contributes more to EAP students' reading proficiency scores – background knowledge or level of L2 reading proficiency?

To answer this question I ran two multiple regression analyses, one based on students' scores on the complete reading modules and one on the revised ones. The independent variables included language proficiency and variables relating to background knowledge (see Chapter 10). When scores on the complete modules were included in the analysis, language proficiency was shown to have by far the strongest relationship with the results: it accounted for 44% of the variance, whereas the addition of background knowledge variables only increased this figure to 45%. When the revised modules were used, however, the contribution of language proficiency, although still very strong, was less marked: it accounted for 26% of the variance, and the addition of background knowledge variables raised this to 38%. It seems likely that as the modules became more subject specific, background knowledge had a proportionately stronger effect on test scores. The comparative effects of language proficiency and background knowledge will be discussed further below.

Of the background knowledge variables included in this analysis, the two which were most strongly related to students' scores on their own module were background reading and subject area familiarity. The significant contribution of background reading emphasises how important areas outside the students' field of academic study are in increasing their background knowledge. The fact that subject area familiarity made a significant contribution to the test scores, whereas topic familiarity did not, suggests that knowledge of a subject area might have a greater effect than topic familiarity on the subject specificity of a reading passage.

Research Question 8

Does the effect of background knowledge on L2 reading comprehension vary according to the level of L2 proficiency of the reader?

Students were divided into three levels according to their scores on the Grammar test. Those with scores of less than 60% did not appear to profit from their background knowledge: there were no significant subject area effects. However, for students with scores above 60%, the subject area effects of the most subject specific pairing – BSS and LMS – were highly significant (see Tables 10.10 and 10.11). There was no steady increase in the effect of background knowledge as students' level of proficiency rose. Rather there seemed to be a threshold below which students did not make use of this knowledge, and above which they did.

Research Implications

Background Knowledge

It is clear from my own results and from the research described in the first three chapters of this book that background knowledge plays a key part in the reading process. Anyone who has studied the research by Bransford into the critical effect that awareness of context has on L1 reading (see my Chapter 2), or Schank and Abelson's (1977) description of the inferences required for the comprehension of even the simplest restaurant script, cannot be in any doubt about the importance of prior knowledge.

We do not yet know enough about the cognitive processes of the brain to be able to tell whether we store and retrieve knowledge in the ways proposed by schema theorists, and although explanations of the processes involved in simple events such as the SHIP'S CHRISTENING (Anderson and Pearson 1988) are very persuasive, they are too simplistic to take full account of all the multiple schemata that must be activated during a reading of any of the IELTS reading texts. For example, if we take a sentence from PST(Ship) we can see that it is not just one schema that must be activated but many.

> *Place the steel ULF (underwater lifting frame) on the seabed immediately above the hull, the four legs seated each in a steel foot which in turn is founded in an excavated pit in the seabed. (Appendix 6.1)*

At the very least the reader must activate schemata relating to the underwater lifting frame, steel, the seabed, a ship's hull and pit excavating, and these must interrelate in a very complex way. Schema theory cannot provide a complete explanation of how we store and retrieve knowledge because it does not account for this diversity of processes which must all take place at the same time. It might, however, as I mentioned in Chapter 2, account for one of the subprocesses involved in parallel distributed processing (McClelland, Rumelhart and Hinton 1986) or in mental modelling (Johnson-Laird, 1983). Whatever the actual processes involved in reading comprehension may be, the value of schema theory to applied linguists is that it proposes formal structures for the acquisition and retrieval of knowledge, and thus gives some form to the amorphous notion of background knowledge.

In Chapter 10, I showed how schema theory might be a useful foundation for further investigations into the reasons for the varying specificity of the IELTS reading subtests. I gave as an example some hypothetical students' approaches to answering one of the PST module test questions, and showed how their responses might be influenced by their content schemata. Another area of schema theory which might also produce useful insights into the findings of my study relates to 'formal' schemata (see Chapters 2 and 3). Rumelhart (1975), Thorndyke (1977), Meyer and Freedle (1984) and Carrell (1987a) believe that people have formal schemata of the structure of different kinds of writing, and that if they can activate the appropriate formal schemata when reading, their comprehension and memory of that text will be enhanced. If this is the case, a student who has the formal schemata relating to the reading passages in his or her IELTS reading module might find the passages easier to read than one who does not.

Another theory, which ties in with Bachman's notions of contextualisation (Bachman 1990 :132) although not explicitly related to it, and which would account well for some of my findings is Haviland and Clark's (1974) Given /New Strategy (see my Chapter 2). Applying Haviland and Clark's ideas to an IELTS passage, we can say that the author of the text would have written it for readers with an expected amount of Given knowledge in the chosen field, and that this knowledge would be necessary for a complete understanding of the passage. For a reader outside that field, without the required prior knowledge, the information would not be Given, and in cases where this reader had to make inferences and go beyond the passage for comprehension, he or she might fail to understand the text fully. If, however, the passage contained a high proportion of Given material, then prior knowledge about the subject area might not be necessary. Passages such as (LMS)Tooth and (LMS)Child, which, according to my research, are not subject specific, might contain a high proportion of Given material. It would be interesting to relate these ideas to students' reading test introspections.

Short Circuits and Threshold Levels

As we saw in Chapter 3, research into whether advanced L2 learners can make use of L1 reading processes (see, for example, Coady 1979 and Cziko 1980) has produced mixed results, possibly because of differences in the levels of proficiency of the students in the experiments. In spite of this, I accept McLeod and McLaughlin's proposal (1986) that once an advanced learner is sufficiently familiar with the vocabulary and linguistic rules of a language to be able to decode written text automatically, that learner will be able to make full use of L1 reading processes. It is therefore reasonable to think that for L2 readers there is a continuum from the novice who is learning elementary decoding skills, to the highly advanced learner who is capable of using the same reading processes in the first and the second languages.

From the results of my research it seems that the use of L1 reading processes may not gradually increase with L2 language proficiency, but that there may be stages that the learners have to reach before they can change their L2 reading processes. For students at the lower levels of L2 learning there appears to be a linguistic threshold that must be reached before they are able to make use of top-down processes to interpret reading texts. The ability of elementary learners to decode letters, words and phrases may be so weak that they are able to make little or no use of their background knowledge; their comprehension is 'short circuited' (Clarke, 1980). Perfetti and Lesgold (1977) suggest that 'bad' L1 readers are so slow at decoding symbols that they cannot retain material in their short-term memory long enough to be able to call up the relevant background knowledge. This is likely to be the case to an even greater extent with low proficiency L2 readers. This 'threshold hypothesis' (see Alderson, 1984) is supported in my research: the subject area of the reading modules had no effect on the performance of low proficiency students, and there was no gradual rise in the effect of background knowledge for students with increasingly higher scores, but for students with Grammar scores of above 60%, there was a highly significant subject effect. This suggests that low proficiency students do not have the linguistic capacity to make use of their background knowledge, but that higher proficiency students do. This does not mean that such learners do not attempt to use such processes, but, as Kozminsky and Graetz (1986) say, they are not able to integrate these processes effectively with bottom-up processes.

For advanced learners, too, there may be a further threshold above which they are able to use bottom-up reading processes as automatically as native speakers. Above this threshold readers are so proficient linguistically that they can compensate for a certain lack of background knowledge by making full use of their language resources. As Bernhardt (1991) says (see Chapter 3), linguistic knowledge begins to override knowledge-driven inferencing. This would account for the fact that ESP teachers are able to understand and teach texts outside their own subject area. Stanovich (1980) suggests that if L1 readers do not recognise a word or phrase, they use the top-down method of guessing, and that if the topic is unfamiliar, they make more use of bottom-up methods. This may apply to high level L2 learners as well, and may explain why in the BSS/LMS pairing, students with scores of above 80% on the Grammar test were less affected by subject area than were the intermediate level students.

Although the students in my study represented a wide range of language proficiency, wide enough for some research to be carried out into reading processes at different levels of ability, none of the students were beginners, and few, if any, were very highly proficient, since few of the most linguistically able students would be attending the English classes at which my tests were administered. This means that my findings were inevitably limited. It is possible that with a wider sample there would have been stronger evidence of these two threshold levels. The results would also have been more informative if the reading modules had been more subject specific. A study is needed in which the student sample covers the complete range of L2 speakers from new beginners to top proficiency learners.

The Comparative Importance of Language Proficiency and Background Knowledge

Although it may be accepted that readers' levels of L2 language proficiency will affect their use of prior knowledge, there has been disagreement about the relative importance of these two factors in reading comprehension. In some of the early research based on schema theory it was claimed that the content of a passage had more effect on learners' comprehension than the syntactic complexity of a text (for example, Johnson, 1981 and 1982). Although there is no direct relationship between syntactic complexity and level of proficiency, this could be understood to imply that background knowledge is more important than level of language proficiency level. However, this does not accord with the short circuit and threshold hypotheses; nor does it accord with my own findings. In a related study, I found that language proficiency appeared to have a much stronger effect on students' scores than did background knowledge. However, the comparative importance of the variables seemed to depend on the specificity of the tests. When the students' scores were based on the revised, more subject-specific subtests, background knowledge contributed to a higher proportion of the test score variance than it did when the students' scores were based on the compete modules. As we know, once a text is highly specialised, and is based on complex concepts which are familiar to only a limited group of readers, good language proficiency is no longer sufficient for text comprehension. The effect of background knowledge on reading comprehension, therefore, depends not only on the proficiency level of the students, but also on the specificity of the reading passages. This brings me directly to two central problems in research into the effect of background knowledge on reading: firstly, the difficulty in assessing that background knowledge, and secondly, the difficulty in ensuring that the reading passages are suitably specific for the purposes of the research. I shall first examine the question of how we assess background knowledge, and then discuss text specificity.

The Assessment of Background Knowledge

The difficulty in assessing the extent of people's background knowledge cannot be overemphasised. Unless the aims of the research are very restricted, as they were, for example, in Spilich *et al.'s* (1979) research into the effect of baseball knowledge on the comprehension of a baseball text (see Chapter 2), it is wrong to assume that because people belong to one group they share the same background knowledge. In the case of my research, it is wrong to assume that because students come from the same academic discipline, they share the same academic knowledge. Within one of the IELTS broad subject areas students may be learning subjects as diverse as computer studies and mineral resources, and they may also have acquired relevant knowledge outside their academic courses. To get more idea of the students' background knowledge, I looked at factors which would affect their academic subject knowledge, such as past schooling and background reading, and I also asked students whether they were familiar with the subject areas and topics of the texts they had just read. However, I made no attempt to measure other, possibly untestable, factors such as memory or general knowledge.

In research into the effect of background knowledge on reading performance, one of the most useful ways of assessing students' background knowledge is to give students knowledge tests in the subject areas of the proposed research (see Tan 1990). Unfortunately such tests are time-consuming to administer and in any case will not provide a complete picture of the student's range of information in the relevant area. It might be quicker to give one of Bernhardt's (1991) free association tests (see Chapter 3), but here again such a test will not elicit a complete impression of the students' knowledge. In academic reading studies, students' reading habits should certainly be taken into account. The question in my questionnaire was very general and did not distinguish between different kinds of reading (see Appendix 6.4). If there had been time it might have been more valuable to ask students separate questions about books, newspaper articles and academic papers. It might also have been useful to ask students about their viewing and listening habits, since much information can be absorbed from television and radio. However, because of the variation in people's memory and the range of their general knowledge, it has to be accepted that any survey, however thorough, would be incomplete.

Specificity of Texts

One of the most important findings in my research has been that the IELTS reading passages vary widely in their subject specificity. Although most of the texts come from academic publications, my analyses revealed that they ranged from 'general' texts which were equally understandable by students in any subject area, to 'highly specific' passages which could generally only be understood by readers with not only a knowledge of a particular subject area, but also a detailed knowledge of some specialised process within it. My own intuitions about the texts' specificity (see Table 7.6) proved to be only partly correct, and I had to adjust the 'general' and 'specific' labels that I had assigned to the reading passages once I had studied the results of repeated measures analyses of variance, the analysis of bias, and the students' comments on the familiarity of the reading passage subject areas. The fact that these passages were selected by experienced EAP teachers, and checked by members of the IELTS project committee, suggests that the test constructors were not aware that test specificity might pose a problem, or that it is difficult to know in advance how specific a passage will be. The fact that the three TMC raters could not agree on the specificity of some of the texts (see Chapter 9) suggests that for applied linguists with a BSS background the latter at least may be the case. This poses a real problem for ESAP test writers.

This variation in text specificity may prove to be a common weakness with research into the effect of background knowledge on academic reading comprehension. It may be the reason, for example, why Shoham, Peretz and Vorhaus (1987) (see my Chapter 1) found less subject area effect than they expected in their research into the effect of background knowledge on academic reading performance. The passages they used 'were excerpted from articles that had appeared in science, technology, biology and humanities and social science readers'(1987:84), and, as far as I know, were not assessed for specificity. Any future research of this kind should include some kind of assessment of the subject specificity of the reading passages, even if this is limited to a questionnaire to subject specialists in the appropriate subject areas.

ESAP or EGAP

I now return to the outstanding research issue which I formulated in Chapter 3:

> *If it is accepted that background knowledge does have some effect on reading comprehension, should this be explicitly taken into account when EAP proficiency tests are devised? Should students intending to study in different academic areas be given reading tests in these different subject areas, so that they are not disadvantaged by a lack of appropriate background knowledge?*

My answer to this question is naturally affected by the fact that my research was based on three subject areas which are so broad that they each include a wide range of disciplines. My results might have been different if the research had been based, for example, on the five subject areas of the original ELTS battery. However, as subject areas become narrower, it becomes more difficult to allocate students to the appropriate modules (see Criper and Davies 1988), and even with more specific tests such as, say, Engineering, there is a risk that some students will not find that the reading passages relate to their own branch of the subject (see Chapter 4). It might therefore be that any advantage that my research might have gained from separating out the disciplines might have been offset by the difficulty of placing students into the most appropriate fields of study.

The question of which subject areas should be tested in an ESAP test has not yet been satisfactorily answered, and if we look at the case of the two IELTS science modules we can see some of the problems. Because there were no research findings to help decide which subject areas the IELTS test battery should contain, the IELTS project committee decided that there should be two science modules – LMS and PST. This decision was made partly at the suggestion of some receiving institutions (see Chapter 4) and partly because approximately equal numbers of students were registering for courses in the three subject areas of BSS, LMS and PST. However, this division of science into LMS and PST has turned out to be neither necessary or advisable. The fact that many or most LMS students study PST subjects during their university course and are therefore familiar with the subject areas of many PST texts, and the fact that the analysis of the results of the revised modules showed no subject area effect between LMS and PST students suggests that a PST-based module would be appropriate for both LMS and PST students. It might therefore be possible to reduce the number of reading modules to two – BSS and Science. On the other hand, the PST module does not seem to be appropriate for both physical scientists and engineers. Its uneasy mixture of texts does not satisfy subject specialists in either area (see Chapter 8), and in the content validation of the first IELTS draft tests (see Chapter 4) there was a clear split between the views of the physical scientists and the engineers. This suggests that there should still be two science modules, but that one should cover the life, medical and physical sciences, and the other engineering and technology. Studies of the test performance of the physical scientists and the engineers in my sample might throw some light on this, but I suspect that any decision to change the subject areas in this way would alter the problem rather than remove it.

ESAP tests are not like tests for other easily defined groups such as doctors or pilots who will be using their English for clearly defined purposes: EAP students do not fit comfortably into separate distinct groups. As Weir's (1983) needs analysis and our content validation (Chapter 4) have shown, the enabling skills and tasks required for students in different disciplines overlap. Although tests in their own field of study may help those whose disciplines are central to the subject area of the module concerned, they may penalise those who for one reason or another are not sufficiently familiar with the appropriate subject area. As the responses to the question relating to familiarity of subject area have shown (see Appendix 8.1), although the majority of students were familiar with the subject area of specific texts in the appropriate field of study, a sizeable minority were not. For example, a third of the LMS students were not familiar with the subject areas of the 'highly specific' (LMS)Genes and (LMS)Nitro reading passages.

I have shown how difficult it is to make sure that ESAP texts are specific to their designated subject area, and what complex procedures have to be followed to ensure that the texts are specific. The inclusion of a general reading passage in what is supposedly a subject specific test may not matter to the students as long as it does not place them at a disadvantage, but it destroys the whole purpose of having tests in different subject areas. All the effort of producing subject specific tests is wasted if such tests turn out to consist of general reading passages. However, even if the texts are specific or highly specific, it is not clear from my study how many students would profit or suffer from taking reading modules in different subject areas. It therefore seems advisable not to give academic students subject specific reading modules, but to give them an EGAP reading test instead.

If it is accepted that EAP students should all take the same reading module, a decision has to be made about the types of reading passages on which the test should be based. Should they be non-academic texts? Should they be academic texts written for BSS students, or should they be academic texts taken from a variety of disciplines? The answer to this may depend on whether the candidates are undergraduates or postgraduates. Future undergraduates might not be expected to be familiar with academic genres since they are in most cases coming fresh from school but postgraduates would be expected to be familiar with academic writing. Since I am not concerned in this study with differences in the needs of undergraduates and postgraduates, and since the majority of IELTS academic candidates are postgraduates, I shall confine my comments to tests which might be suitable for postgraduates.

Non-Academic Texts

The texts used in the present TOEFL 'are taken from general reading materials rather than specialised textbooks' (Hale 1988). Such texts, by definition, are not academic, and do not contain the variety of genres encountered in academic studies. For IELTS to include reading passages of this non-academic type would be to ignore the results of Weir's (1983) needs analysis and the IELTS content validation study (see Chapter 4), and would make any such test a test of general rather than university level reading. One of the strong points about the present IELTS is that the inclusion of academic texts encourages tutors to use study skills teaching methods when they are preparing students for the test (see Adams 1992). This gives students practice in the academic skills they will need in their further studies. The introduction of general non-academic texts might decrease this study skills element, and would also lessen the test's face validity. The passages should therefore come from authentic academic texts. (Since there is some uncertainty as to what an 'academic ' text is [see Chapter 9] the test specifications would need to follow the example of the present ones [see Appendix 4.5] and contain a detailed description of the kinds of texts that would be acceptable.)

BSS-Based Texts

If it is agreed that the reading comprehension passages should come from academic sources, it has to be decided on what subject areas these texts should be based, and there are some arguments in favour of BSS. Chi squared tests of the relationship between students' test scores and their familiarity with the reading passage subject areas showed that there was no correspondence between familiarity with the subject area and M(BSS) test scores, whereas for the two science modules there was (see Table 8.4). This suggests that most science students were not disadvantaged when they were given BSS texts, and agrees with the findings of much of the ESP testing research described in Chapter 1. Alderson and Urquhart (1985b), Shoham, Peretz and Vorhaus (1987), Jensen and Hansen (1995) and O'Neill, Steffen and Broch (1994) all found that while science students scored as highly as non-science students on non-science texts, non-scientists did not achieve equal scores on the science-based texts. However, the repeated measures analysis of variance studies in Chapter 7 showed that there were significant subject area effects when the three BSS subtests were paired with subtests from the science modules. Although it is possible that these subject effects were one sided, and that BSS students were being disadvantaged when they took the science-based tests, whilst science students were not disadvantaged when they took the BSS-based test, we do not know if this is the case. It must also be remembered that although most science students reported that they read BSS based material, 13% did not. It would therefore be inequitable to give science students a reading module based on BSS texts.

Academic 'General' Texts

If it is not reasonable to choose passages from one subject area, and if it is agreed that the texts should come from academic sources, the only remaining option is that they should come from a variety of disciplines within BSS, LMS and PST, but that they should be non-subject specific, that is 'general' in the sense that (LMS)Tooth and (LMS)Child are. They should be equally comprehensible for students in all fields. Such texts would have to be submitted to subject specialists (and possibly to students) to check that they were genuinely not subject specific, and the results of the trials analysed to see that no disciplines were advantaged over others. The difficulty here might be to ensure that a variety of text types was included: in Chapter 9, I showed that introductions to papers tended to be 'general' and processes 'specific'. Test writers would need to find passages that varied in presentation and genre, and yet did not favour students in one subject area over another.

Suggestions for Further Research

I have made many suggestions for further research throughout this book. I shall briefly now list those that I think are most important, and shall add to them some which I have not had cause to mention before. Although some of the suggestions are related to the effect of background knowledge on reading comprehension, and some to EAP testing, the findings of most of them are likely to be interrelated.

1 In this book I have made many general suggestions about the reading processes of students in the three IELTS subject areas. Some of these suggestions can only be verified by a qualitative study into students' reading processes. Students should be asked to introspect as they take the reading modules within and outside their own study area. They should focus on the use they make of background knowledge as they describe what goes through their mind as they answer each test item. Their answers could be related to the tenets of schema theory, or to Haviland and Clark's (1974) Given/New Strategy, or to Bachman's (1990) Levels of Contextualisation. However, since the TMC raters appeared to be uncertain of the meaning of the 'contextualisation–topic specificity' facet, and since this should be useful for interpreting the subject specificity of reading passages, it would be helpful if it could be more fully defined.

2 In my study the range of students was not wide enough for me to get a full idea of the thresholds or stages at which students at varying levels of proficiency start to use different reading processes. A large study should be undertaken in which L2 students over the whole spectrum of language ability from new beginners to those with the highest levels of proficiency are tested for differences in their use of top-down and bottom-up processes, and the use they make of background knowledge in particular. Such a study should perhaps be based on a homogeneous group of students sharing the same first language, since, as I suggested in Chapter 5, variations in the reading processes of different speech communities might otherwise confound the results.

3 My research into the comparative effects of background knowledge and language proficiency on reading comprehension suggests that when reading materials are not highly specific for their readers, language proficiency will be a better predictor of reading scores than background knowledge, but that as the texts become more specific, background knowledge becomes more important. It would be interesting to run multiple regression analyses on students' scores on a range of subtests of varying levels of specificity, to see whether my supposition is correct. If it is, regressions on the 'general' subtests might show that background knowledge failed to account for any of the score variance, whereas on the 'highly specific' subtests background knowledge might account for more of the variance than language proficiency.

4 It would have been interesting in this present study to reclassify the students according to narrower disciplines such as engineering, law and physics, or yet narrower ones such as civil engineering and maritime law, and to compare these results with those from the broader set of classifications. However, there were too few students in some of the disciplines for this to be worthwhile. With a larger and better balanced student sample, such a comparison would make a valuable study since broad groupings such as those in the present study

must mask important differences between the disciplines within a subject area.

5 The results of my investigations into the sources of the IELTS reading passages (Chapter 9), and the results of the TMC rating exercise (Chapter 9) showed that there is some uncertainty as to what an 'academic' text might be. If it is decided that an EGAP test should include reading passages based on academic material, more research is needed into what distinguishes an academic text from a non-academic one.

Summary of the Main Findings

It is clear that background knowledge does affect EAP reading comprehension. However, there are two important conclusions to be drawn from this study which must influence our interpretation of the other findings. The first is that there was no clear demarcation between the students in the three fields of study. In each of the three, students came from a wide range of academic disciplines, and their reports about their background reading showed that most read materials outside the IELTS subject area. From their questionnaire answers it also emerged that some students within a given field of study were not familiar with the subject areas of the subtests, although some students from outside that area were. It became plain that it is not possible to separate out EAP students into different subject areas in the way that it is possible, for example, to separate doctors from pilots.

The second finding is that the reading passages vary in their subject specificity from 'general' to 'highly specific'. These passages, although they were selected by informed EAP teachers and testers, are in some cases no more appropriate for students in the relevant subject area than for students in the other academic fields. If we look at the main research results we can see the effects of this variation in text specificity. Students on the whole achieved higher scores on the reading module within rather than outside their own subject area. However, because of the lack of specificity of the revised PST module, and the overlap between the LMS and PST study areas, there was no subject area effect for the LMS/PST pairing. Similarly, when undergraduates and postgraduates were compared to see which group was the more affected by the subject area of the reading modules, the most subject-specific pairing – BSS/LMS – showed no difference between the two academic levels: in both cases students did significantly better at the test within their own subject area. However, for the two other pairings, the undergraduates were not affected by the subject area, whereas the postgraduates were. It seems that the subtests had to reach a certain level of specificity before both undergraduates and postgraduates were affected by the subject area of the passages.

The specificity of the subtests also affected the results of the studies into background knowledge and reading comprehension. When the modules included 'general' passages, the level of language proficiency had markedly more effect on students' scores than did background knowledge. However, once the modules contained only 'specific' passages, background knowledge became proportionately more important. It might be hypothesised that if all the subtests had been 'highly specific', background knowledge might have made an equal or greater contribution to comprehension than language ability.

The importance of background knowledge was also shown to vary according to students' level of language proficiency. Students with low L2 language skills seemed to be unable to make full use of their background knowledge until they reached a certain proficiency threshold. However, once again the findings varied according to the specificity of the reading passages. The threshold effect which was demonstrated by the most specific pairing was not replicated by the other two: for those modules in which some of the passages were not specific the results were contradictory.

Variation in the specificity of reading passages may account for many of the anomalous results in the ESP testing literature, and may indeed influence the results of other areas of research into the effect of background knowledge on reading comprehension.

References

Adams, B. 1992. *Washback – An Investigation of the Influence of the IELTS (International English Language Testing System) Test on Teaching, as Perceived by Teachers.* MA Dissertation. Lancaster University: Department of Linguistics and Modern English Language.

Adams, M. J. and A. Collins. 1979. A Schema-Theoretic view of reading. In R. O. Freedle (ed.).

Adams, R. J. and S. Khoo. 1993. *Quest: The Interactive Test Analysis System.* Hawthorne, Victoria: The Australian Council for Educational Research.

Adams, R. J. and K. J. Rowe. 1988. Item bias. In J. P. Keeves (ed.).

Adams, S. J. 1982. Scripts and the recognition of unfamiliar vocabulary: Enhancing second language reading skills. *Modern Language Journal* 66: 155–9.

Alba, J. W., S. G. Alexander, L. Hasher and K. Camiglia. 1981. The role of context in the encoding of information, *Journal of Experimental Psychology: Human Learning and Memory* 7: 283–92.

Alba, J. W. and L. Hasher. 1983. Is memory schematic? *Psychological Bulletin* 93: 203–31.

Alderson, J. C. 1981. Report of the discussion on testing English for specific purposes. In J. C. Alderson and A. Hughes (eds.).

Alderson, J. C. 1983. The cloze procedure and proficiency in English as a foreign language. In J. W. Oller (ed.).

Alderson, J. C. 1984. Reading in a foreign language: a reading problem or a language problem? In J. C. Alderson and A. Urquhart (eds.).

Alderson, J. C. 1988a. Testing English for specific purposes – How specific can we get? In A. Hughes (ed.).

Alderson, J. C. 1988b. Testing and its administration in ESP. In D. Chamberlain and R. J. Baumgardner (eds.).

Alderson, J. C. 1988c. New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing* 5: 220–32.

Alderson, J. C. 1990a. Testing reading comprehension Skills (Part One). *Reading in a Foreign Language* 6: 425–38.

Alderson, J. C. 1990b. Testing reading comprehension skills (Part Two). *Reading in a Foreign Language* 7: 465–503.

Alderson, J. C. 1993a. The relationship between grammar and reading in an English for academic purposes test battery. In D. Douglas and C. Chapelle (eds.).

Alderson, J. C. 1993b. Judgements in language testing. In D. Douglas and C. Chapelle (eds.).

Alderson, J. C. and C. M. Clapham. 1992. Applied linguistics and language testing: a case study of the ELTS Test. *Applied Linguistics* 13: 149–67.

Alderson, J. C. and C. M. Clapham. 1993. *Examining the ELTS Test: an Account of the First Stage of the ELTS Revision Project.* IELTS Research Report 2. British Council, University of Cambridge Local Examinations Syndicate and International Development Program of Australian Universities and Colleges, London, Cambridge and Canberra.

Alderson, J. C. and A. Hughes (eds.). 1981. *Issues in Language Testing*, ELT Documents 111. Oxford: Pergamon Press.

Alderson, J. C. and Y. Lukmani. 1989. Cognition and reading: Cognitive levels as embodied in test questions. *Reading in a Foreign Language* 5: 253–70.

Alderson, J. C. and A. H. Urquhart. 1983. The effect of student background discipline on comprehension: a pilot study. In A. Hughes and D. Porter (eds.).

Alderson, J. C. and A. H. Urquhart (eds.). 1984. *Reading in a Foreign Language.* London: Longman.

Alderson, J. C. and A. H. Urquhart. 1985a. The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2: 192–204.

Alderson, J. C. and A. H. Urquhart. 1985b. This test is unfair: I'm not an economist. In P. C. Hauptman, R. Le Blanc and M. B. Wesche (eds.).

Anderson, R. C. and A. Ortony. 1975. On putting apples into bottles: a problem in polysemy. *Cognitive Psychology* 7: 167–80.

Anderson, R. C. and P. D. Pearson. 1988. A schema-theoretic view of basic processes in reading comprehension. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Anderson, R. C, J. W. Pichert, E. T. Goetz, D. L. Schallert, K. V. Stevens, and S. R. Trollop. 1976. Instantiation of general terms. *Journal of Verbal Learning and Verbal Behavior* 15: 667–79.

Anderson, R. C., Reynolds, R. E., Schallert, D. L. and Goetz, E. T. 1977a. Frameworks for comprehending discourse. *American Educational Research Journal* 14: 367–381.

Anderson, R. C., R. J. Spiro and W. E. Montague (eds.). 1977b. *Schooling and the Acquisition of Knowledge.* Hillsdale, NJ: Lawrence Erlbaum.

Andrich, D. 1988. *Rasch Models for Measurement.* Sage, Newbury Park, CA.

Ausubel, D. P. 1960. The use of advance organisers in the learning and retention of meaningful material. *Journal of Educational Psychology* 51: 267–72.

Ausubel, D. P. 1963. *The Psychology of Meaningful Verbal Learning.* New York, NY: Grune and Stratton.

Ausubel, D. P. and D. Fitzgerald. 1962. Organisers, general background and antecedent learning variables in sequential verbal learning. *Journal of Educational Psychology* 53: 243–9.

Bachman, L. F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F., F. Davidson, B. Lynch and K. Ryan. 1989. Content analysis and statistical modelling of EFL proficiency tests. Paper presented at the Language Testing Research Colloquium, San Antonio, TX.

Bachman, L. F., F. Davidson and M. Milanovic. 1991. The use of test method characteristics in the content analysis and design of English proficiency tests. Paper presented at the Language Testing Research Colloquium, Princeton, NJ.

Bachman, L. F., F. Davidson, K. Ryan and I. Choi. 1995. *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*. Cambridge: Cambridge University Press.

Baker, R. L. 1995. *Classical Test Theory and Item Response Theory in Test Analysis*. Special Report 2, Language Testing Update. Centre for Research in Language Education, Department of Linguistics, Lancaster University, Lancaster.

Barber, C. L. 1962/1985. Some measurable characteristics of modern scientific prose. In J. M. Swales, 1985.

Barnett, M. A. 1989. *More than Meets the Eye*. Englewood Cliffs, NJ: Prentice Hall Regents.

Barr, R., M. Kamil, P. Mosenthal and P. D. Pearson (eds.). 1991. *Handbook of Reading Research*, Vol 2. London: Longman.

Bartlett, F. C. 1932. *Remembering*. Cambridge: Cambridge University Press.

Bazerman, C. 1988. *Shaping Written Knowledge*, Madison, WI and London: The University of Wisconsin Press.

Beers, T. 1987. Schema-Theoretic models of reading: Humanising the machine. *Reading Research Quarterly* 2: 369–77.

Berman, R. 1984. Syntactic components of the foreign language reading process. In J. C. Alderson and A. Urquhart (eds.).

Bernhardt, E. B. 1991. *Reading Development in a Second Language*. Norwood, NJ: Ablex.

Block, E. 1986. The comprehension strategies of second language readers. *TESOL Quarterly* 20: 463–94.

Bloom, B. S. 1956. *Taxonomy of Educational Objectives. Book 1: Cognitive Domain*. London: Longman.

Blue, G. 1987. Report after an ELTS revision meeting of teachers of English for academic purposes. Department of Linguistics and Modern English Language, Lancaster University, Lancaster.

Blue, G. (ed.). 1993. *Language Learning and Success: Studying through English*. Modern English Publications, London.

Blue, G. 1993. Nothing succeeds like linguistic competence: The Role of Language in Academic Success. In G. Blue (ed.).

Bobrow, D. G. and A. Collins (eds.). 1975. *Representation and Understanding: Studies in Cognitive Science*. New York, NY: Academic Press.

Bossers, B. 1991. On thresholds, ceilings and short-circuits: The relation between L1 reading, L2 reading and L2 knowledge. In J. H. Hulstijn and J. F. Matter (eds.).

Bossers, B. 1992. *Reading in Two Languages*. Drukkerij Van Driel, Amsterdam.

Bransford, J. D. 1979. *Human Cognition: Learning, Understanding and Remembering*. Belmont, CA: Wadsworth.

Bransford, J. D. and M. K. Johnson 1972. Contextual prerequisites for understanding some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior* 11: 717–26.

Bransford, J. D. and M. K. Johnson. 1973. Considerations of some problems of comprehension. In W. D. Chase (ed.).

Bransford, J. D. and N. S. McCarrell. 1977. A sketch of a cognitive approach to comprehension: Some thoughts about understanding what it means to comprehend. In P. N. Johnson-Laird and P. C. Wason (eds.).

Bransford, J., B. Stein and T. Shelton. 1984. Learning from the perspective of the comprehender. In J. C. Alderson and A. H. Urquhart (eds.).

Bridgeman, B. and S. Carlson. 1983. *Survey of Academic Writing Tasks Required of Graduate and Undergraduate Foreign Students*. TOEFL Research Report 15, Princeton, NJ: Educational Testing Service.

Britton, B. K. and J. B. Black (eds.). 1985. *Understanding Expository Text*. Hillsdale, NJ: Lawrence Erlbaum.

Buck, G. 1990. *The Testing of Second Language Listening Comprehension*. Unpublished Ph. D. thesis, Lancaster University, Lancaster.

Burroughs, G. E. R. 1971. *Design and Analysis in Educational Research*. Educational Monograph 8, School of Education, University of Birmingham.

Candlin, C. N., J. H. Leather and C. J. Bruton. 1976. Doctors in casualty: Applying communicative competence to components of specialist course design. *International Review of Applied Linguistics* 14: 246–72.

Carrell, P. L. 1983. Three components of background knowledge in reading comprehension. *Language Learning* 33: 183–207.

Carrell, P. L. 1984a. Evidence of a formal schema in second language comprehension. *Language Learning* 34: 2, 87–112.

Carrell, P. L. 1984b. The effects of rhetorical organisation on ESL readers. *TESOL Quarterly* 18: 441–70.

Carrell, P. L. 1985. Facilitating ESL reading by teaching text structure. *TESOL Quarterly* 19: 727–52.

Carrell, P. L. 1987a. Content and formal schemata in ESL reading. *TESOL Quarterly* 21: 461–81.

Carrell, P. L. 1987b. Readability in ESL. *Reading in a Foreign Language* 4: 21–40.

Carrell, P. L. 1988. Some causes of text-boundedness and schema interference in ESL reading. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Carrell, P. L. 1991. Second language reading: Reading ability or language proficiency? *Applied Linguistics* 12: 159–79.

Carrell, P. L., J. Devine and D. E. Eskey (eds.). 1988. *Interactive Approaches to Second Language Reading.* Cambridge: Cambridge University Press.

Carrell, P. L., B. G. Pharis and J. C. Liberto. 1989. Metacognitive strategy training for ESL reading. *TESOL Quarterly* 23: 647–78.

Carroll, J. B. 1972. Defining language comprehension: Some speculations. In J. B. Carroll and R. O. Freedle (eds.).

Carroll, B. J. 1981. Specifications for an English language testing service. In J. C. Alderson and A. Hughes (eds.).

Carroll, J. B. and R. O. Freedle (eds.). 1972. *Language Comprehension and the Acquisition of Knowledge.* Washington, DC: Winston.

Carver, R. P. 1977. A theory of reading comprehension and rauding. *Reading Research Quarterly* 13: 8–64.

Chamberlain, D. and R. J. Baumgardner (eds.). 1988. *ESP in the Classroom: Practice and Evaluation.* London: Modern English Publications.

Chase, W. D. (ed.). 1973. *Visual Information Processing.* New York, NY: Academic Press.

Chiesi, H. L., G. J. Spilich and J. F. Voss. 1979. Acquisition of domain-related information in relation to high and low domain knowledge. *Journal of Verbal Learning and Verbal Behaviour* 18: 257–73.

Chomsky, N. 1957. *Syntactic Structures.* The Hague: Mouton.

Chomsky, N. 1959. Review of *Verbal Behavior* by B. F. Skinner. *Language* 35: 26–58.

Clapham, C. M. 1981. Reaction to the Carroll paper (1). In J. C. Alderson and A. Hughes (eds.).

Clapham, C.M. 1994. *The Effect of Background Knowledge on EAP Reading Test Performance.* Unpublished Ph.D. thesis, Lancaster University.

Clapham, C. M. and J. C. Alderson (forthcoming). *Constructing and Trialling the IELTS Test.* IELTS Research Report 3, British Council, University of Cambridge Local Examinations Syndicate and International Development Program of Australian Universities and Colleges, London, Cambridge and Canberra.

Clapham, C. M. and A. Hughes. 1988. Summary of discussion on predictive validity. In A. Hughes, D. Porter and C. J. Weir (eds.).

Clarke, M. A. 1980. The short circuit hypothesis of ESL reading – or when language competence interferes with reading performance. *Modern Languages Journal* 64: 203–9.

Clarke, M. A. and S. Silberstein. 1977. Toward a realisation of psycholinguistic principles in the ESL reading class. *Language Learning* 27: 135–54.

Clymer, T. 1972. What is reading? Some current concepts. In A. Melnik and J. Merritt (eds.).

Coady, J. 1979. A psycholinguistic model of the ESL reader. In R. Mackay, B. Barkman and R. R. Jordan (eds.).

Cohen, A., P. R. Glasman, P. R. Rosenbaum-Cohen, J. Ferrara and J. Fine. 1988. Reading English for specialised purposes: Discourse analysis and the use of student informants. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Coleman, H. 1988. Teaching and learning styles within and across cultures: Implications for language pedagogy. Paper presented at the Fourth ILE International Conference, Hong Kong.

Collins, A. M. and M. R. Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8: 240–8.

Corder, S. P. and E. Roulet (eds.). 1973. *Theoretical Linguistic Models in Applied Linguistics*, Paris: AIMAV and Didier.

Craik, F. I. and R. S. Lockhart. 1972. Levels of processing. A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11: 671–84.

Criper, C. 1981. Reaction to the Carroll paper (2). In J. C. Alderson and A. Hughes (eds.).

Criper, C. and A. Davies. 1988. *ELTS Validation Project Report.* ELTS Research Report 1 (i). London and Cambridge: British Council and University of Cambridge Local Examinations Syndicate.

Crocker, L. and J. Algina. 1986. *Introduction to Classical and Modern Test Theory.* London: Holt, Rinehart and Winston.

Cushing Weigle, S. 1994. *Effects of Training on Raters of ESL Compositions: an Investigation of Construct Validity.* Unpublished Ph. D. thesis, University of California at Los Angeles, CA.

Cziko, G. A. 1980. Language competence and reading strategies: a comparison of first- and second-language oral reading errors. *Language Learning* 30: 101–16.

Davies, A. 1965. *Proficiency in English as a Second Language.* Unpublished Ph.D. thesis, University of Birmingham, Birmingham.

Davies, A. 1981a. Review of J. Munby: *Communicative Syllabus Design. TESOL Quarterly* 15: 32–6.

Davies, A. 1981b. Reaction to the Palmer and Bachman and the Vollmer papers. In J. C. Alderson and A. Hughes (eds.).

Davies, A. and J. C. Alderson. 1977. *English Proficiency Test Battery (EPTB).* Department of Applied Linguistics, Edinburgh: Edinburgh University.

de Jong, J. H. A. L. and D. K. Stevenson (eds.). 1990. *Individualizing the Assessment of Language Abilities.* Multilingual Matters, Clevedon, PA.

Devine, J. 1987. General language competence and adult second language reading. In J. Devine, P. L. Carrell and D. E. Eskey (eds.).

Devine, J. 1988a. A case study of two readers: Models of reading and reading performance. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Devine, J. 1988b. The relationship between general language competence and second language reading proficiency: Implications for teaching. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Devine, J., P. L. Carrell and D. E. Eskey (eds.). 1987. *Research in Reading as English as a Second Language.* Washington, DC: TESOL.

van Dijk, T. A. and W. Kintsch. 1983. *Strategies of Discourse Comprehension.* Academic Press, London.

Donin, J. and M. Silva. 1993. The relationship between first- and second-language reading comprehension of occupation specific texts. *Language Learning* 43: 373–401.

Dornic, S. (ed.). 1977. *Attention and Performance VI.* Hillsdale, NJ: Lawrence Erlbaum.

Douglas, D. and C. Chapelle (eds.). 1993. *A New Decade of Language Testing Research: Collaboration and Cooperation.* Washington, DC: TESOL.

Dudley-Evans, T. 1987. Introduction to genre analysis and ESP. *ELR Journal* 1: 1–9.

Dudley-Evans, T. and W. Henderson. 1990. *The Language of Economics: The Analysis of Economics Discourse.* ELT Documents 134, London: Modern English Publications.

ELTSREV 1987a. Report on views of British Council headquarters staff. Department of Linguistics and Modern English Language. Lancaster University, Lancaster.

ELTSREV 1987b. Report on discussions with teachers of English for academic purposes. Department of Linguistics and Modern English Language, Lancaster University, Lancaster.

ELTSREV 1987c. Report on discussions with language testers. Department of Linguistics and Modern English Language, Lancaster University, Lancaster.

Eskey, D. E. 1988. Holding in the bottom: An interactive approach to the language problems of second language readers. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Everitt, B. S. 1968. Moments of the statistics Kappa and weighted Kappa. *The British Journal of Mathematical and Statistical Psychology* 21: 97–102.

Evinston, P. (ed.). 1975. *The Psychology of Computer Vision*, New York, NY: Winston.

Fahnestock, J. 1986. Accommodating science: The rhetorical life of scientific facts. *Written Communication* 3: 275–96.

Flesch, R. 1962. *The Art of Readable Writing*. New York, NY: Collier.

Floyd, P. and P. L. Carrell. 1987. Effects on ESL reading of teaching cultural content schemata. *Language Learning* 37: 89–108.

Foulkes, J. and P. Hargreaves. 1988. Subject specialisation and performance on ESP tests. Paper presented at the Language Testing Research Colloquium, Champaign, IL.

Francis, B., M. Green and C. Payne. 1993. *The GLIM System: Release 4 Manual*. Oxford: Oxford University Press.

Freebody, P. and R. C. Anderson. 1983. Effects of vocabulary difficulty, Text Cohesion and Schema Availability on Reading Comprehension. *Reading Research Quarterly* 18: 272–93.

Freedle, R. O. (ed.). 1979. *New Directions in Discourse Processing*. Vol 2, Norwood, NJ: Ablex.

Fries, C. C. 1963. *Linguistics and Reading*. Holt, New York, NY: Rinehart and Winston.

Garnham, A. 1985. *Psycholinguistics:Central Topics*. London: Methuen.

Garnham, A. and J. Mason. 1987. Episode structure in memory for narrative text. *Language and Cognitive Processes* 2: 133–44.

Geoghegan, G. 1983. *Non-Native Speakers of English at Cambridge University*. Cambridge: The Bell Educational Trust.

Gollasch, F. W. (ed.). 1975. *Language and Literacy: the Selected Writings of Kenneth Goodman*. London: Routledge and Kegan Paul.

Goodman, K. S. 1967/1970. Reading: a psycholinguistic guessing game. In D. V. Gunderson (ed.).

Goodman, K. S. 1973. Psycholinguistic universals in the reading process. In F. Smith, 1973.

Goodman, K. S. 1975. The reading process. In F. W. Gollasch (ed.).

Goodman, K. S. 1981. Letter to the Editors. *Reading Research Quarterly* 16: 477–8.

Gough, P. B. 1972. One second of reading. In J. F. Kavanagh and I. G. Mattingley (eds.).

Grabe, W. 1988. Reassessing the term 'Interactive'. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Gray, W. S. 1948. *On Their Own in Reading*. Glenview, IL: Scott Foresman.

Gray, W. S. 1960. The major aspects of reading. In H. M. Robinson (ed.) 1960.

Griffin, P. 1990. Characteristics of the test components of the IELTS battery: Australian trial data. Paper presented at the Regional English Language Centre, Singapore.

Gunderson, D. V. (ed.). 1970. *Language and Reading*. Center for Washington, DC: Applied Linguistics.

Hale, G. A. 1988. Student major field and text content: Interaction effects on reading comprehension in the test of English as a foreign language. *Language Testing* 5: 49–61.

Hambleton, R. K., H. Swaminathan and H. J. Rogers. 1991. *Fundamentals of Item Response Theory.* Newbury Park, CA: Sage.

Hamilton, J., M. Lopes, T. McNamara and E. Sheridan. 1993. Rating scales and native speaker performance on a communicatively oriented EAP test. *Melbourne Papers in Language Testing* 2: 1–24.

Hamp-Lyons, E. M. 1987. *Testing Second Language Writing in Academic Settings.* Unpublished Ph. D. thesis, University of Edinburgh, Edinburgh.

Hauptman, P. C., R. Le Blanc and M. B. Wesche (eds.). 1985. *Second Language Performance Testing.* Ottawa: University of Ottawa Press.

Haviland, S. E. and H. H. Clark. 1974. What's new? Acquiring new information as a process in comprehension. *Journal of Verbal Learning and Verbal Behavior* 13: 512–21.

Head, H. 1920. *Studies in Neurology.* Oxford: Oxford University Press.

Hieronymus, A. N., H. D. Hoover and E. F. Lindquist. 1986. *Preliminary Teachers Guide. Multilevel Battery Levels 9 – 14. Iowa Tests of Basic Skills,* Chicago, IL: Riverside Publishing Company.

Horn, E. 1937 *Methods of Instruction in the Social Studies.* New York, NY: Charles Scribner.

Huddleston, R.D. 1971 *The Sentence in Written English: A Syntactic Study Based on an Analysis of Scientific Texts.* Cambridge: Cambridge University Press.

Hudson, T. 1982. The effects of induced schemata on the 'Short Circuit' in L2 reading: Non-decoding factors in L2 reading performance. *Language Learning* 32: 3–31.

Huey, E. B. 1908/1968. *The Psychology and Pedagogy of Reading.* Cambridge, MA: MIT Press.

Hughes, A. (ed.). 1988. *Testing English for University Study.* London: Modern English Publications.

Hughes, A. and D. Porter (eds.). 1983. *Current Developments in Language Testing.* London: Academic Press.

Hughes, A., D. Porter and C. J. Weir (eds.). 1988. *ELTS Validation Project: Proceedings of a Conference Held to Consider the ELTS Validation Project Report.* ELTS Research Report 1(ii). London and Cambridge: British Council and University of Cambridge Local Examinations Syndicate.

Hulstijn, J. H. 1991. How is reading in a second language related to reading in a first language? In J. H. Hulstijn and J. F. Matter (eds.).

Hulstijn, J. H. and J. F. Matter (eds.). 1991. *Reading in Two Languages.* AILA Review 8, Amsterdam.

Hutchinson, T. and A. Waters 1987. *English for Specific Purposes.* Cambridge: Cambridge University Press.

Hymes, D. 1972. On communicative competence. In J. B. Pride and J. Holmes (eds.).

Jenkinson, M. D. 1972. Sources of knowledge for theories of reading. In A. Melnik and J. Merritt (eds.).

Jensen, C. and C. Hansen. 1995. The effect of prior knowledge on EAP listening-test performance. *Language Testing* 12: 99–119.

Johnson, P. 1981. Effects on reading comprehension of language complexity and cultural background of a text. *TESOL Quarterly*, 15:169–81.

Johnson, P. 1982. Effects on reading comprehension of building background knowledge. *TESOL Quarterly* 16: 503–16.

Johnson-Laird, P. N. 1983. *Mental Models*. Cambridge: Cambridge University Press.

Johnson-Laird, P. N. and P. C. Wason (eds.). 1977. *Thinking: Readings in Cognitive Science*. Cambridge: Cambridge University Press.

Jordan, R. R. 1989. English for academic purposes (EAP). *Language Teaching* 22: 150–164.

Just, M. A. and P. A. Carpenter. 1977. *Cognitive Processes in Comprehension*. Hillsdale, NJ.: Lawrence Erlbaum.

Kattan, J. 1990. *The Construction and Validation of an EAP Test for Second Year English and Nursing Majors at Bethlehem University*. Unpublished Ph.D. thesis. Lancaster University, Lancaster.

Kavanagh, J. F. and I. G. Mattingley (eds.). 1972. *Language by Ear and Eye*. Cambridge, MA: MIT Press.

Keeves, J. P. (ed.). 1988. *Educational Research Methodology and Measurement: an International Handbook*. Oxford: Pergamon Press.

Kerlinger, F. N. 1973. *Foundations of Behavioural Research*. London: Holt, Rinehart and Winston.

Kintsch, W. 1974. *The Representation of Meaning in Memory*. Hillsdale, NJ: Lawrence Erlbaum.

Kintsch, W. and E. Greene. 1978. The role of culture-specific schemata in the comprehension and recall of stories. *Discourse Processes* 1: 1–13.

Kintsch, W. and T. A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* 85: 363–394.

Klare, G. R. 1974. Assessing readability. *Reading Research Quarterly* 10: 62–102.

Koh, M. Y. 1985. The role of prior knowledge in reading comprehension. *Reading in a Foreign Language* 3: 375–80.

Kozminsky, E. and N. Graetz. 1986. First versus second language comprehension: Some evidence from text summarising. *Journal of Research in Reading* 9: 3–21.

LaBerge D. and S. J. Samuels. 1974. Toward a theory of automatic information processing in reading. *Cognitive Psychology* 6: 293–323.

Laufer, B. and D. D. Sim. 1985. Measuring and explaining the reading threshold needed for English for academic purposes texts. *Foreign Language Annals* 18: 405–11.

Lee, J. F. 1986. Background knowledge and L2 reading. *Modern Languages Journal* 70: 350–4.

Levine, M. G. and G. J. Haus. 1985. The effect of background knowledge on the reading comprehension of second language learners. *Foreign Language Annals* 18: 391–7.

Lewis-Beck, M. S. 1980. *Applied Regression: An Introduction*. Beverly Hills, CA: Sage.

Lipson, M. Y. 1984. Some unexpected issues in prior knowledge and comprehension. *The Reading Teacher*, April: 760–4.

Lord, F. M. 1980. *Applications of Item Response Theory to Practical Testing Problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lumley, T. 1993. The notion of subskills in reading comprehension tests: an EAP example. *Language Testing* 10: 211–34.

Lyons, J. (ed.). 1985. *New Horizons in Linguistics*. Harmondsworth: Penguin Books.

McClelland, J., D. E. Rumelhart and G. E. Hinton. 1986. Models and general issues in cognitive science. In D. E. Rumelhart, J. McClelland and the PDP Research Group (eds.).

McEldowney, P. L. 1976. *Test in English (Overseas): The Position after 10 Years*. Manchester: Joint Matriculation Board.

McLaughlin, B., T. Rossman and B. McLeod. 1983. Language learning: an information-processing perspective. *Language Learning* 33: 135–58.

McLeod, B. and B. McLaughlin. 1986. Restructuring or authenticity? Reading in a second language. *Language Learning* 36: 108–23.

McNamara, T. P., D. L. Miller and J. D. Bransford. 1991. Mental models and reading comprehension. In R. Barr, M. Kamil, P. Mosenthal and P. D. Pearson (eds.).

Mackay, R., B. Barkman and R. R. Jordan (eds.). 1979. *Reading in a Second Language: Hypotheses, Organization and Practice*. Rowley, MA: Newbury House.

Mandl, H., N. Stein and T. Trabasso (eds.). 1984. *Learning and Comprehension of Text*, Hillsdale, NJ.: Lawrence Erlbaum.

Mandler, J. M. 1984. *Stories, Scripts and Scenes: Aspects of Schema Theory*. Hillsdale, NJ.: Lawrence Erlbaum.

Manning, W. 1987. *Development of Cloze Elide Tests of English as a Second Language*. TOEFL Research Report 23, Princeton, NJ: Educational Testing Service.

Matthews, M. 1990. Skill taxonomies and problems for the testing of reading. *Reading in a Foreign Language* 7: 511–17.

Mead, R. 1982. Review of Munby: *Communicative Syllabus Design. Applied Linguistics* 3: 70–8.

Melnik, A. and J. Merritt (eds.). 1972. *Reading: Today and Tomorrow*. London: University of London Press.

Meyer, B. J. F. 1975. *The Organisation of Prose and Its Effect on Memory*. Amsterdam: North-Holland Publishing Company.

Meyer, B. J. F. 1984. Text dimensions and cognitive processing. In H. Mandl, N. Stein, and T. Trabasso (eds.).

Meyer, B. J. F. 1985. Prose analysis: Purposes, procedures, and problems. In B. K. Britton and J. B. Black (eds.).

Meyer, B. J. F. and R. O. Freedle. 1984. Effects of discourse type on recall. *American Educational Research Journal* 21: 121–43.

Microsoft. 1991. *Word for Windows Users' Guide*. Redmond, WA: Microsoft.

Miller, M. D. and R. L. Linn. 1988. Invariance of item characteristic functions with variations on instructional coverage. *Journal of Educational Measurement* 25: 205–19.

Minsky, M. 1975. A framework for presenting knowledge. In P. Evinston (ed.).

Minsky, M. 1977. Frame system theory. In P. N. Johnson-Laird and P. C. Wason (eds.).

Mitchell, D. C. 1982. *The Process of Reading: a Cognitive Analysis of Fluent Reading and Learning to Read*. Chichester: John Wiley.

Moller, A. 1981. *A Study in the Validation of Proficiency Tests of English as a Foreign Language*. Unpublished Ph. D. thesis, University of Edinburgh, Edinburgh.

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

Myers, G. 1991. Lexical cohesion and specialized knowledge in science and popular science tests. *Discourse Processes* 14: 1–25.

Namukwai, V. and R. Williams. 1988. The readability of government reports: A case study from Zambia. *Reading in a Foreign Language* 5: 205–27.

Nash-Weber, B. 1975. The role of semantics in automatic speech understanding. In D. G. Bobrow and A. Collins (eds.).

Nickel, G. (ed.). 1977. *Proceedings of the Fourth International Congress of Applied Linguistics*. Volume 3, Hochschulverlag, Stuttgart.

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

Oller, J. W. (ed.). 1983. *Issues in Language Testing Research*. Rowley, MA.: Newbury House.

Oller, J. W. and C. Conrad. 1971. The Cloze Procedure and ESL Proficiency. *Language Learning* 21: 183–96.

O'Neill, K., M. Steffen and E. Broch. 1994. Does undergraduate major affect DIF on reading comprehension items? Paper presented at the National Convention of the American Educational Research Association, New Orleans, LA.

Orasanu, J. (ed.). 1986. *Reading Comprehension: from Research to Practice*. Hillsdale, NJ.: Lawrence Erlbaum.

Otto, W. (ed.). 1977. *Reading Problems*. Boston, MA: Addison-Wesley.

Otto, W. and S. White. (eds.) 1982. *Reading Expository Material*. New York, NY: Academic Press.

Palmer, A. S. and L. F. Bachman. 1981. Basic concerns in test validation. In J. C. Alderson and A. Hughes (eds.).

Pearson, P. D. and L. Fielding. 1991. Comprehension instruction. In R. Barr, M. Kamil, P. Mosenthal and P. D. Pearson (eds.).

Peretz, A. S. and M. Shoham. 1990. Testing reading comprehension in LSP. *Reading in a Foreign Language* 7: 447–55.

Perfetti, C. A. and A. M. Lesgold. 1977. Discourse comprehension and sources of individual differences. In M. A. Just and P. A. Carpenter (eds.).

Pichert, J. W. and R. C. Anderson. 1977. Taking different perspectives on a story. *Journal of Educational Psychology* 69: 309–15.

Pollitt, A. 1988. Predictive validity (1). In A. Hughes, D. Porter and C. J. Weir (eds.).

Powers, D. E. 1985. *A Survey of Academic Demands Related to Listening Skills*. TOEFL Research Report 20. Princeton, NJ: Educational Testing Service.

Pride, J. B. and J. Holmes (eds.). 1972. *Sociolinguistics*. Harmondsworth: Penguin.

Reckase, M. D., T. A. Ackerman and J. E. Carlson. 1988. Building a unidimensional test using multidimensional items. *Journal of Educational Measurement* 25: 193–202.

Robinson, H. A. (ed.). 1966. *Reading: Seventy-Five Years of Progress*. Chicago, IL: University of Chicago Press.

Robinson, H. M. (ed.). 1960. *Sequential Developments of Reading Abilities*. Supplementary Educational Monographs 90, Chicago, IL: University of Chicago Press.

Robinson, H. M. 1966. The major aspects of reading. In H. A. Robinson (ed.).

Robinson, P. 1991. *ESP Today: A Practitioner's Guide*. New York, NY: Prentice Hall.

Roller, C. M. and A. R. Matambo. 1992. Bilingual readers' use of background knowledge in learning from text. *TESOL Quarterly* 26: 129–41.

Rumelhart, D. E. 1975. Notes on a schema for stories. In D. G. Bobrow and A. Collins (eds.).

Rumelhart, D. E. 1977a. Towards an interactive model of reading. In S. Dornic (ed.).

Rumelhart, D. E. 1977b. *An Introduction to Human Information Processing*. New York, NY: Wiley.

Rumelhart, D. E. 1980. Schemata: the building blocks of cognition. In R. J. Spiro, B. C. Bruce and W. F. Brewer (eds.).

Rumelhart, D. E., J. McClelland and the PDP Research Group, (eds.). 1986. *Parallel Distributed Processing*. Vol. 1, Massachusetts Institute of Technology, Cambridge, MA.

Rumelhart, D. E. and A. Ortony. 1977. The representation of knowledge in memory. In R. C. Anderson, R. J. Spiro and W. E. Montague (eds.).

Samuels, S. J. 1977. Introduction to theoretical models of reading. In W. Otto (ed.).

Samuels, S. J. and M. L. Kamil. 1988. Models of the reading process. In P. L. Carrell, J. Devine and D. E. Eskey (eds.).

Sarig, G. 1987. High level reading in the first and in the foreign language: Some comparative process data. In J. Devine, P. L. Carrell and D. E. Eskey (eds.).

Schallert, D. 1982. The significance of knowledge: A synthesis of research related to schema theory. In W. Otto and S. White (eds.).

Schank, R. C. 1972. Conceptual dependency: a theory of natural language understanding. *Cognitive Psychology* 3: 552–631.

Schank, R.C. 1980. What's in a schema anyway? Review of R. O. Freedle (ed.). *New Directions in Discourse Processing. Contemporary Psychology* 25: 814–6.

Schank, R. C. and R. Abelson. 1977. Scripts, plans and knowledge. In P. N. Johnson-Laird and P. C. Wason (eds.).

Schroeder, L. D., D. L. Sjoquist and P. E. Stephan. 1986. *Understanding Regression Analysis*. Beverly Hills, CA: Sage.

Seaton, I. 1981. Background to the specifications for an English language testing service and subsequent developments. In J. C. Alderson and A. Hughes (eds.).

Segalowitz, N., C. Poulsen and M. Komeda. 1991. Lower level components of reading skill in higher level bilinguals: Implications for reading instruction. In J. H. Hulstijn and J. F. Matter (eds.).

Shoham, M., A. S. Peretz and R. Vorhaus. 1987. Reading comprehension tests: General or subject specific? *System* 15: 81–8.

Skehan, P. 1984. Issues in the testing of English for specific purposes. *Language Testing* 1: 202–20.

Skinner, B. F. 1957. *Verbal Behavior*, New York, NY: Appleton-Century-Crofts.

Smith, F. 1973. *Psycholinguistics and Reading*. Holt, New York, NY: Rinehart and Winston.

Smith, F. 1988. *Understanding Reading*. 4th Edition, New York, NY: Holt, Rinehart and Winston.

Spilich, G. J., G. T. Vesonder, H. L. Chiesi and J. F. Voss. 1979. Text processing of domain-related information for individuals with high and low domain knowledge. *Journal of Verbal Learning and Verbal Behavior* 18: 275–90.

Spiro, R. J. 1980. Constructive processes in prose comprehension and recall. In R. J. Spiro, B. C. Bruce and W. F. Brewer (eds.).

Spiro, R. J., B. C. Bruce and W. F. Brewer (eds.). 1980. *Theoretical Issues in Reading Comprehension*. Hillsdale, NJ: Erlbaum.

Spolsky, B. 1977. Language testing: Art or Science. In G. Nickel (ed.).

SPSS. 1990. *Reference Guide for the Statistical Package for the Social Sciences*. Chicago, IL: SPSS.

Stanovich, K. E. 1980. Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly* 16: 32–71.

Stanovich, K. E. 1991. Word recognition: Changing perspectives. In R. Barr, M. Kamil, P. Mosenthal and P. D. Pearson (eds.).

Steffensen, M. S. and C. Joag-Dev. 1984. Cultural Knowledge and Reading. In J. C. Alderson and A. H. Urquhart (eds.).

Steffensen, M. S., C. Joag-Dev and R. C. Anderson. 1979. A cross cultural perspective on reading comprehension. *Reading Research Quarterly* 15: 10–29.

Strang, R. 1972. The nature of reading. In A. Melnik and J. Merritt (eds.).

Swaffar, J. K., K. M. Arens and H. Byrnes. 1991. *Reading for Meaning: an Integrated Approach to Language Learning*. Englewood Cliffs, NJ: Prentice Hall.

Swales, J. M. 1985. *Episodes in ESP*. Oxford: Pergamon.

Swales, J. M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Symons, S. and M. Pressley. 1993. Prior knowledge affects text search success and extraction of information. *Reading Research Quarterly* 28: 251–9.

Tan, S. H. 1990. The role of prior knowledge and language proficiency as predictors of reading comprehension among undergraduates. In J. H. A. L. de Jong and D. K. Stevenson (eds.).

Tannen, D. 1979. What's in a frame? In R. O. Freedle (ed.).

Tatsuoka, K. K., R. L. Linn, M. M. Tatsuoka and K. Yamamoto. 1988. Differential item functioning resulting in the use of different solution strategies. *Journal of Educational Measurement* 25: 301–19.

Taylor, I. and M. M. Taylor. 1983. *The Psychology of Reading*, New York, NY: Academic Press.

Thorndike, E. L. 1971/1972. Reading as reasoning: A study of mistakes in paragraph reading. In A. Melnik and J. Merritt (eds.) 1972.

Thorndyke, P. W. 1977. Cognitive structures in comprehension and memory of narrative discourse. *Cognitive Psychology* 9: 77–110.

Trimble, L. 1985. *English for Science and Technology*. Cambridge: Cambridge University Press.

Turner, C. E. 1988. *The Underlying Factor Structure of L2 Cloze Test Students: Causal Modeling as an Approach to Construct Validation*. Unpublished Ph.D. thesis, McGill University, Montreal.

Urquhart, A. H. 1984. The effect of rhetorical ordering on readability. In J. C. Alderson and A. H. Urquhart (eds.).

van den Wollenberg, A. L. 1982. Two new test statistics for the Rasch model. *Psychometrika* 47: 123–40.

van Ek, J. A. and J. L. M. Trim. 1975/1991. *Threshold Level 1990.* Strasbourg: Council of Europe Press.

Vesonder, G. T. 1979. *The Role of Knowledge in Processing Experimental Reports.* Unpublished Ph. D. thesis, University of Pittsburgh, PA.

Vollmer, H. J. 1981. Why are we interested in 'General Language Proficiency'? In J. C. Alderson and A. Hughes. (eds.).

Voss, J. F. and G. L. Bisanz. 1985. Knowledge and the processing of narrative and expository texts. In B. K. Britton and J. B. Black (eds.).

Walcutt, C. C. 1967. Reading – a professional definition. *Elementary School Journal* 67: 363–5.

Weir, C. J. 1983. *Identifying the Language Problems of Overseas Students in Tertiary Education in the United Kingdom.* Unpublished Ph. D. thesis, University of London, London.

Westaway, G., J. C. Alderson and C. M. Clapham. 1990. Directions in testing for specific purposes. In J. H. A. L. de Jong and D. K. Stevenson (eds.).

Widdowson, H. 1973. Directions in the teaching of discourse. In S. P. Corder and E. Roulet (eds.).

Widdowson, H. 1979. *Explorations in Applied Linguistics.* Oxford: Oxford University Press.

Widdowson, H. 1983. *Learning Purpose and Language Use.* Oxford: Oxford University Press.

Wilkins, D. 1976. *Notional Syllabuses.* Oxford: Oxford University Press.

Wilson, P. T. and R. C. Anderson. 1986. What they don't know will hurt them: The role of prior knowledge in comprehension. In J. Orasanu (ed.).

Winograd, T. 1972. *Understanding Natural Language.* New York, NY: Academic Press.

Winograd, T. 1975. Frame representations and the declarative-procedural controversy. In D. G. Bobrow and A. M. Collins (eds.).

Winograd, T. 1983. *Language as a Cognitive Process: Syntax.* Reading, MA: Addison-Wesley.

Wolff, D. 1987. Some assumptions about second language text comprehension.

Appendices

Chapter Four
The Development of the IELTS Reading Modules

4.1 Main Disciplines Covered by the IELTS Reading Modules
4.2 Questionnaire for Subject Specialists. (All questionnaires and test papers have been reduced in size to save space.)
4.3 Subject Specialists' Disciplines
4.4 Questionnaire for Language Teachers, Testers and Applied Linguists
4.5 Extracts from Draft Specifications – LMS Reading Module

Chapter Six
The Main Study – Research Questions and Data Collection

6.1 The Reading Modules and Answer Keys
6.2 Clausal Analysis Graphs
6.3 The Grammar Test
6.4 Students' Questionnaire
6.5 Distribution of Scores on the Grammar Test and the Reading Modules

Chapter Eight
Reasons for the Variation in Subtest Specificity

8.1 Subject Area Familiarity
8.2 Topic Familiarity
8.3 Questions to Subject Specialists
8.4 Subject Specialists – Familiarity with Topic

Chapter Nine
Text Source and the TMC Rating Instrument

9.1 Test Method Characteristics Instrument

Chapter 10
Language Proficiency and Background Knowledge

10.1 Distribution of the Scores on the Revised LMS and PST Modules

Appendix 4.1 Main Disciplines Covered by the IELTS Reading Modules

Business Studies and Social Science

Archaeology, Art History, Business Studies, Development Studies, Economics, Education, English Language Teaching, English Literature, Estate Management, Finance, Geography, History, Languages, Law, Linguistics, Music, Philosophy, Political Science, Population Studies, Psychology, Public Administration, Sociology, Social Anthropology, Theatre Studies, Theology, Urban Development

Life and Medical Sciences

Agriculture, Agronomy, Animal Nutrition, Bacteriology, Biology, Clinical Tropical Medicine, Community Health, Dentistry, Ecology, Environmental Science, Epidemiology, Forestry, Genetics, Immunology, Land Protection, Medicine, Microbiology, Nutrient Enrichment, Obstetrics and Gynaecology, Ophthalmology, Paediatric Medicine, Parasitology, Pathology, Physiology, Plant Physiology, Toxicology, Veterinary Science, Virology, Zoology

Physical Science and Technology

Architecture, Chemistry, Civil Engineering, Computer Studies, Construction Management, Control Systems, Electrical Measurement, Electronics Engineering, Energy, Exploration Geophysics, Fermentation Technology, Fluid Mechanics, Geochemistry, Geology, Hydrology, Laser Spectroscopy, Mathematics, Mechanical Engineering, Metallurgy, Meteorology, Petroleum Geology, Physics, Soil Mechanics, Water Resources

Appendix 4.2 Questionnaire for Subject Specialists

I. **Specifications**

1.a. Do the specifications take into account what is known about reading and writing in your discipline?
If no,
b. What additional information do the item writing teams need to consider?
2. Are the "Purposes" sufficiently detailed to cover the kinds of reading and writing done by your students?
3. Are the criteria for selecting texts specified in such a way that the most appropriate texts will be chosen?

II. **Texts**

4.a. Are the texts used for the sample items the sorts of texts your *undergraduate* students have to read?
If no,
b. How do they differ from the ones they will have to deal with in their studies?
5. Would students wishing to enter your undergraduate courses be disadvantaged if they took a test using these texts?
6.a. Are the texts used for the sample items the sorts of texts your *postgraduate* students have to read?
If no,
b. How do they differ from the ones they will have to deal with in their studies?
7. Would students wishing to enter your postgraduate courses be disadvantaged if they took a test using these texts?
8. Does the language used in these texts differ from that which your undergraduates and/or postgraduates meet in their studies?
9. Can you rank the texts in order of their suitability for your students? (You may want to rank the texts differently for undergraduates and postgraduates.)
10. Could you give us samples of the course materials and texts (references will be fine here) that your students work with in their studies?

III. **Tasks**

11.a. Does the sample test contain the sorts of reading tasks your *undergraduate* students have to do?
If no,
b. Would students wishing to enter your undergraduate courses be disadvantaged if they were asked to do these tasks in a language test?
12. What other tasks might your undergraduate students be asked to do with these texts?
13.a. Does the sample test contain the sorts of reading tasks your *postgraduate* students have to do?
If no,
b. Would students wishing to enter your postgraduate courses be disadvantaged if they were asked to do these tasks in a language test?
14. What other tasks might your postgraduate students be asked to do with these texts?

Appendix 4.3 Subject Specialists' Disciplines

M (Business & Social Sciences)

Accountancy	Education
Commerce	Education
Business Studies	Educational Research
Business Analysis	Police Studies
Business Administration	ELT
Public Administration	English Literature
Management Studies	Theatre Studies
Information Technology	Religious Studies
Politics	Philosophy
Politics	Law
Political Economy	Law
Economics	Population Studies
Sociology	Geography
Psychology	Earth Sciences

M (Physical Science & Technology)

Chemical Engineering	Computer Science
Civil Engineering	Computer Studies
Civil Engineering	Computer Studies
Civil Engineering	Computer Studies
Civil Engineering	Physics
Mechanical Engineering	Physics
Mechanical Engineering	Physics
Engineering Geology	Physics
Geology	Physics
Electronics	Mathematics
Oceanography (Chemical)	Mathematics
Ship Sciences	Mathematics
Transportation	Mathematical Statistics
Transportation	Mathematics & Computer Science
Basic Science	Information Technology
Chemistry	Physical Science
Chemistry	

M (Life & Medical Sciences)

Agriculture	Plant Biology
Agriculture	Plant Biology
Agriculture	Genetics
Agronomy	Physiology & Pharmacology
Agronomy	Psychology
Animal Nutrition	Psychology
Ecology	Medicine
Environmental Science	Psychiatry
Forestry & National Resources	Veterinary Science
Biochemistry	Tropical Veterinary Medicine
Biology	Biotechnology

Appendix 4.4 Questionnaire for Language Teachers, Testers and Applied Linguists

(Questions 1–19 related to the Grammar, Listening and Speaking Tests.)

General Questions

Because different students will take different subject modules, we are concerned that no student be disadvantaged because of the module he/she takes, yet we also want the modules to reflect differences among the fields of study. The following 3 questions relate to the comparability and distinctiveness of the modules:

20. Is it acceptable that the modules differ considerably in format and structure? Is it desirable that they do so?
21. Do the modules measure abilities that are not measured by the general (G) components?
22. Do the modular components overlap inappropriately with the general components?
23. Are the modules appropriate for students entering either undergraduate or graduate courses?

Specific Questions

We should be grateful if you would answer the following questions about <u>each</u> of the three discipline-related modular components.

M (Physical Science and Technology)

I. **Specifications**

24. a. Do the specifications take into account what is known about reading and writing in this subject area?
    If no,
    b. What additional information do the item writing teams need to consider?
25. Are the "Purposes" sufficiently detailed to cover the kinds of reading and writing done by students in this subject area?
26. Are the criteria for selecting texts specified in such a way that the most appropriate texts will be chosen?

II. **Texts**

27. a. Are the texts used for the sample items the sorts of texts students have to read in this subject area?

If no,

b. How do they differ from the ones they will have to deal with in their studies?

28. Would students in this subject area be disadvantaged if they took a test which contained these texts?

29. Can you rank the texts in order of their suitability for students in this subject area?

30. Does the language used in these texts differ from that which students in this subject area meet in their studies?

III. **Tasks**

31. a. Does the sample test contain the sorts of reading tasks students have to do in this subject area?

If no,

b. Do you think these students would be disadvantaged if they were asked to do these tasks in a language test?

32. What other tasks might they be asked to do with these texts?

(Questions 33–35 relate to the Writing test.)

M (Life and Medical Sciences)

I. **Specifications**

36. a. Do the specifications take into account what is known about reading and writing in this subject area?

If no,

b. What additional information do the item writing teams need to consider?

37. Are the "Purposes" sufficiently detailed to cover the kinds of reading and writing done by students in this subject area?

38. Are the criteria for selecting texts specified in such a way that the most appropriate texts will be chosen?

II. **Texts**

39. a. Are the texts used for the sample items the sorts of texts students have to read in this subject area?

If no,

b. How do they differ from the ones they will have to deal with in their studies?

40. Would students in this subject area be disadvantaged if they took a test which contained these texts?

41. Can you rank the texts in order of their suitability for students in this subject area?

42. Does the language used in these texts differ from that which students in this subject area meet in their studies?

III. **Tasks**

43. a. Does the sample test contain the sorts of reading tasks students have to do in this subject area?
If no,
b. Do you think these students would be disadvantaged if they were asked to do these tasks in a language test?

44. What other tasks might they be asked to do with these texts?

(Questions 45–47 relate to the Writing test.)

M (Arts and Social Sciences)

I. **Specifications**

48. a. Do the specifications take into account what is known about reading and writing in this subject area?
If no,
b. What additional information do the item writing teams need to consider?

49. The authors describe students' written answers at five levels of sophistication. Do these cover the kinds of writing your students produce?

50. Are the criteria for selecting texts specified in such a way that the most appropriate texts will be chosen?

II. **Texts**

51. a. Are the texts used for the sample items the sorts of texts students have to read in this subject area?
If no,
b. How do they differ from the ones they will have to deal with in their studies?

52. Would students in this subject area be disadvantaged if they took a test which contained these texts?

53. Can you rank the texts in order of their suitability for students in this subject area?

54. Does the language used in these texts differ from that which students in this subject area meet in their studies?

III. **Tasks**

55. a. Does the sample test contain the sorts of reading tasks students have to do in this subject area?
If no,
b. Do you think these students would be disadvantaged if they were asked to do these tasks in a language test?

56. What other tasks might they be asked to do with these texts?

Appendix 4.5 Extracts from the Draft Specifications – LMS Reading Module

1. INTRODUCTION

....

General Test Focus

The focus of the test is the range of proficiency in reading and writing as outlined in the descriptors for IELTS Bands 5, 6 and 7. The tasks set in the test must focus on the Academic Tasks described below, which are relevant to undergraduate and postgraduate students. Item writers should address these Academic Tasks, developing items which allow candidates to demonstrate levels of proficiency across Band Levels 5, 6 and 7 without excluding the possibility that some candidates may demonstrate higher or lower levels. Materials used and the tasks set should arise from appropriate sources and be relevant to, or clearly directed at, appropriate audiences (as specified overleaf). Item writers should concentrate on developing tasks which address the Academic Tasks listed and not the formal knowledge of grammar.

....

2. SECTION 1: READING (55 minutes)

Test Focus

a) Band Levels

The primary focus for reading in this test should be in the Range of Bands 5, 6 and 7.

b) Academic Tasks

The test should sample the candidates' ability to perform the following tasks. (It is not implied or assumed that these can or must be tested in isolation or independently of each other.)

(i) Identifying structure, content, sequence of events and procedures
(ii) Following instructions
(iii) Finding main ideas which the writer has attempted to make salient
(iv) Identifying the underlying theme or concept

(v) Identifying ideas in the text, and relationships between them e.g. probability, solution, cause, effect

(vi) Identifying, distinguishing and comparing facts, evidence, opinions, implications, definitions and hypotheses

(vii) Evaluating and challenging evidence

(viii) Formulating an hypothesis from underlying theme, concept and evidence

(ix) Reaching a conclusion by relating supporting evidence to the main idea

(x) Drawing logical inferences

c) Source and Audience

Scientific magazines, books, academic papers and well-written newspaper articles relating to the Life and Medical Sciences written by scientists for the informed lay person and for scientists in other fields. (Textbooks and journal articles for specialised audiences within the field of Life and Medical Science should be avoided because they may be too field-specific.)

If the texts contain technical terms which might not be understood by candidates in all subject areas covered by the test (see Target Population), a glossary should be provided; the definitions should require a reading ability of no more than Band 5.

Stimulus Materials

a) Level

A range of materials should be selected with difficulty levels suitable for candidates whose reading abilities lie within Bands 5–7.

b) Texts

There must be at least three reading passages. These should be drawn from at least two different texts (i.e. one text may, for example, be divided into two reading passages) and should present different styles of writing. At least one text should be in the Life Sciences and one in the Medical Sciences. (One way of utilising different styles in an original text might be to divide an academic paper into a review, a description and a discussion.) There may be a thematic link between the texts, but not at the cost of biasing the test in favour of one academic or vocational area, nor at the cost of other test design issues, for example, choice of item types.

The texts used and their topics should be scientific but "neutral"; i.e. they should not be highly discipline-specific nor biased for or against any of the discipline areas covered by the test. Texts from the more serious scientific journals directed at the general public are more likely to be suitable than those from popular newspapers although scientific articles from serious newspapers could be used. Understanding of the texts should not depend on knowledge of any particular branch of science beyond that which might be expected of a candidate entering any of the subject areas listed. The texts should deal with issues which are interesting, recognisably appropriate and accessible to candidates entering courses in the Life and Medical Sciences. They should be at a level of sophistication appropriate for undergraduate students.

Suitable types of text include:

(i) the introductory section from an experimental report
(ii) reviews or reports of research or other projects
iii) analyses of "problems" with suggested explanations or courses of action
(iv) discussion and argument presenting the interpretations, views and opinions of the author or others
(v) speculative discussions of issues
(vi) case studies
(vii) public inquiries

At least one text should contain detailed logical argument and/or text-embedded definitions, and at least one should contain non-verbal materials such as diagrams, graphs, tables, mathematical formulae or illustrations. Texts must be realistic and in modern English and must appear to be authentic, even if the original texts have been modified.

Item writers must submit their texts for approval before proceeding to write items on them.

c) <u>Length</u>

1,500 or 2,500 words in total, depending on the number of figures and diagrams embedded in the text.

d) <u>Structure</u>

Where possible, the reading passages should be sequenced in order of increasing level of reading ability required.

Test Tasks

a) <u>Tasks</u>

The test should sample widely, but not necessarily exhaustively, the Academic Tasks listed on the previous page, where possible testing several Academic Tasks within one test component. When submitting items, item writers should say which of Tasks (i)–(x) each item is intended to test.

First-term undergraduate native speakers must be able to complete all tasks successfully within the time allowed.

b) <u>Item Types</u>

(i) Although the actual number of items will depend on the item types chosen, it is envisaged that the typical test will consist of approximately 40 items. The procedure for assessing comprehension may include the following item types:

Choosing from a "heading bank" a heading appropriate to identified sections of the text
Copying words, phrases, etc. from the text
Information transfer
Labelling or completing diagrams, tables, charts, graphs or illustrations
Listing items or ideas from the text relevant to a given topic or concern
Matching
Multiple-choice
Short-answer questions, up to 3 words only
Sorting events into order
Sorting names/objects into sets
Summary completion (gap-filling)

(ii) Items must be designed so that they can be marked objectively by people who are neither teachers nor language testers and who will not be required to evaluate the quality of an answer.

(iii) Candidates will be awarded one mark for each correct item.

(iv) Item writers should use a variety of item types.

(v) Multiple-choice questions should be used only if that format seems the most suitable way of testing something. They must have four options.

(vi) True/False questions should be avoided as should any item type which has only two possible answers, e.g. short-answer questions to which the response is "Yes" or "No".

(vii) Short-answer questions should be used only if the range of possible answers is small and can be specified exhaustively in the key.

(viii) Item writers may propose alternative test items to those included in these specifications, but these will need to be approved by the International Editing Committee.

....

Appendix 6.1 The Reading Modules and Answer Keys

Name: ....

Institution: ....

Date: ....

INTERNATIONAL ENGLISH LANGUAGE
TESTING SYSTEM

MODULE A

SPECIMEN VERSION

Time allowed: 55 minutes (Reading)

SECTION 1: READING

In this section you will find 3 reading passages. Each of these will be accompanied by some questions. Some of the questions will come before the relevant reading passage; some will come after the passage.

Start at the beginning of the section. If you cannot do one part of the test in the suggested time, leave it and start on the next.

University of Cambridge Local Examinations Syndicate
British Council
International Development Program

SECTION 1: READING

PART 1: LIFE WITHOUT A SUNSCREEN

You are advised to spend about 15 minutes on Questions 1 - 11.

Questions 1 - 3

Answer these questions, using Reading Passage 1, "Life Without a Sunscreen", on Pages 3 - 4.

Make sure your answers reflect the author's views, not your own.

Write your answers beside the questions. The first one has been done as an example.

example: CFC stands for ... chlorofluorocarbon'

1. CFCs break down into .. chlorine compounds

2. Where is ozone formed? atmosphere/stratosphere

3. What does ozone do?

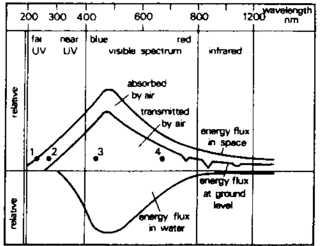
absorbs far UV/absorbs UV between 200 and 300nm/
absorbs (sun's) ultraviolet radiation/protects life
from harmful UV radiation/blocks the sun's ultraviolet
radiation

READING PASSAGE 1: LIFE WITHOUT A SUNSCREEN

Harmful ultraviolet radiation may reach the surface of the Earth in increasing quantities in the 1990s as the ozone layer thins - yet biologists still cannot predict the effects. Scientists have now established that the springtime "hole" in the ozone layer above Antarctica is a direct effect of chlorine compounds produced by the breakdown of chlorofluorocarbons (CFCs), compounds that do not occur naturally and have lifetimes of scores of years in the atmosphere. Some evidence suggests that a smaller ozone hole may be developing over the Arctic, and that CFCs have depleted the ozone in the stratosphere by a few per cent over the past decade.

In the next two or three years, as the Sun builds towards a peak of activity in 1991 to 1992, an increase in the number of charged particles reaching the upper atmosphere will stimulate chemical reactions that counteract the reduction of ozone concentrations caused by CFCs. But atmospheric scientists anticipate that afterwards there will be a sharp decline in the concentration of ozone, and a corresponding increase in the amount of ultraviolet radiation penetrating the atmosphere. This is because as solar activity declines these natural processes will work in step with the influence of CFCs and other "anthropogenic" compounds.

This causes concern for two reasons, both linked to the fact that the ozone in the stratosphere absorbs the Sun's ultraviolet radiation. If there is less ozone, and UV penetrates deeper in the Earth's atmosphere before being absorbed, there will be a change in the heat balance of the atmosphere, with lower layers becoming warmer and higher layers becoming cooler. Secondly, there is a direct threat to life on the surface of the Earth, because the ozone layer absorbs UV with wavelengths that are damaging to living cells. It is this form of UV that is implicated in most skin cancers in people. Although no one now disputes that stratospheric ozone will be depleted in the 1990s, ecologists have yet to define even a hazy outline of the implications for life on Earth.



peak absorbance for -
• 1 germicidal lamps: 254 nm • 2 sunburn (human): 297 nm
• 3 chlorophyll a: 440 nm • 4 chlorophyll a: 720 nm

All the UV in the environment comes from the Sun. The spectrum of energy that the Sun emits is determined almost entirely by the surface temperature, about 600 Kelvin. Most of the energy is in the visible band, but about 6 per cent is radiated in the ultraviolet, at wavelengths shorter than 400 nanometres (nm). The Earth's atmosphere absorbs almost all radiation with wavelengths less than about 300 nm before it can reach the ground. "Extreme" UV (also known as shortwave or vacuum UV) has wave-lengths shorter than 200 nm, and is totally absorbed by both nitrogen and oxygen in the atmosphere. One result of this is that this radiation breaks apart some diatomic molecules of oxygen and the single oxygen atoms join up with other diatomic oxygen molecules to form triatomic oxygen molecules (ozone). The ozone then absorbs ultraviolet in the band between 200 nm and 300 nm, known as "far" UV (or hard UV). The radiation in the band from 300 nm to 400 nm, which does reach the ground, is known as near UV.

Far UV is potentially the most damaging for life, because it has a strong effect on both nucleic acids and proteins. But the Sun produces much less far ultraviolet radiation than it does near UV; at present only a little of this radiation, in the band from 290 nm to 300 nm, ever reaches the ground and then only in strong sunlight. Radiation from 290 nm to 320 nm, a range that overlaps the arbitrary boundary between far and near UV, is sometimes known as biologically active UV. It is the radiation that causes most sunburn and most skin cancers. This is the radiation that will penetrate to the ground more easily as the ozone layer thins. Even a small decrease in the concentration of ozone could double the amount of biologically active UV that reaches the ground; if ozone depletion continues, there could be a severalfold increase in biologically active UV at ground level.

In contrast, most of the ultraviolet reaching the ground today is in the near UV. About half of the Sun's energy output in this waveband reaches the ground already; so whatever happens to the atmosphere, no more than twice the present quantity could ever reach the surface of the Earth. The energy in the upper range of this waveband also has so little effect on living things that it is commonly regarded as "harmless". Some experiments do suggest that very high doses can harm people, but the doses were much higher than we could ever receive from the Sun.

4

Questions 4 - 11

Use Reading Passage 1, "Life without a Sunscreen", on Pages 3 - 4 to help you fill in the following table. Some of the boxes have been filled in for you.

Write your anwers in the column underneath the table. The first one has been done as an example.

	Near UV	Far UV	Extreme UV	Biologically Active UV
wavelengths (nm)	*(example)*	4.	8.	10.
effect on atmosphere	no information given	5.	absorbed by nitrogen and oxygen	absorbed by ozone
effect on humans/life	harmless	6.	none	11.
Proportion of all UV reaching earth	over 50%	7.	9.	little

Your answers

example: 300-400

4. 200-300
5. absorbed (by O_3)/absorbed (by ozone)/the ozone absorbs the far UV
6. strong/damages nucleic acids and proteins/damaging/(has strong) effect on both nucleic acids and proteins
7. small/little/far less than 50%/less than 30%/much less than near UV
8. less than 200/shorter than 200
9. none/nil
10. 290-320
11. causes skin cancer/causes sunburn

5

PART 2: ENERGY FROM FUELS

You are advised to spend about 10 minutes on Questions 12 - 18.

Questions 12 - 18

Use Reading Passage 2 "Energy from Fuels", on Pages 7 - 8, together with the graphs, to answer the following questions.

12. What kind of fuel was predominantly used before coal? wood

13. Which fuel source was most important for industrial development? coal

14. Why are only a proportion of fossil fuels extracted? economic reasons/ cost/ cheaper

Name the four sources of energy termed "natural forces".

15. tidal/sea/rivers/water

16. wind

17. sun/solar

18. geothermal/natural heat (sources from the earth's crust)

READING PASSAGE 2: ENERGY FROM FUELS

Energy Demand and Supply

From the turn of the century up to 1975 energy production in the United States increased by 900%. During the same period, world demand for energy rocketed, and it is only in the last few years that the rate of increase has steadied (see Figure 1). While it is extremely difficult to predict future patterns of demand, most experts agree that requirements for energy are likely to increase. Clearly, the relation between energy demand and production is of vital concern, and a failure to achieve a proper balance could lead to disastrous consequences.

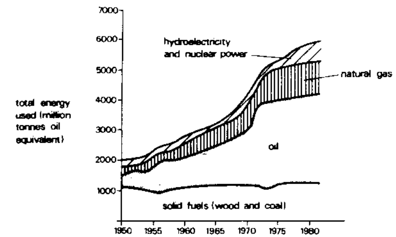


Figure 1

The relative importance of different types of energy sources varies in different parts of the world and is tied to a number of complex and interrelated factors, the most important of which are local availability and access to imported fuels. Each of these is in turn dependent on a range of economic and political factors. However, taking a global perspective, there have been recognisable trends over the last hundred and fifty years. From the middle of the nineteenth century and for the next hundred years or so, coal was the dominant world fuel and easy access to large coal reserves was an important factor in the development of industrial bases in many parts of the world. Prior to this time, the primary source of energy worldwide had been wood. More recently, in the last few decades, technological advances in the exploration, extraction and processing of oil and natural gas have led to these fuels achieving prominence in many countries. Oil and natural gas have the added advantages of being cheaper to extract and transport than coal, and create less environmental pollution.

Non-Renewable Sources

All the energy sources referred to so far are fossil fuels, and as such, reserves are limited. Quite how long they will last is difficult to predict as new reserves are being discovered and future demand is uncertain. Figure 2 shows the number of years known reserves of these fuels can be expected to last at current production rates. However, at present, it is economic considerations that restrict the comprehensive extraction of fossil fuels from existing wells and mines. For example, extraction of oil from existing wells is only about 40% at present. It may be that changing economic circumstances will make it viable to exploit these existing wells and mines more fully. However, sooner or later, reserves will be exhausted, and the world will be unable to depend on fossil fuels as sources of much-needed energy.

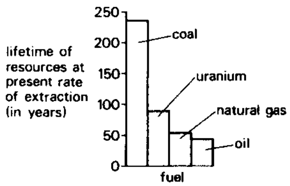


Figure 2

An alternative source of energy is nuclear power but this also has its drawbacks. Firstly, uranium, although not a fossil fuel, is nevertheless non-renewable. In addition, there is considerable public opposition to the growth of the nuclear power industry arising from concern about possible health hazards associated with working with and disposing of radioactive materials.

Renewable Sources

The distinction between non-renewable and renewable sources of energy is a fundamental one. If we are to have any hope of satisfying future energy demand on a global scale, it is essential that we investigate sources of energy that make use of what can be termed 'natural forces'. Four particular areas have been the subject of considerable attention recently.

The harnessing of tidal power is one such area. In France a power station built across the River Rance takes its energy from the ebb and flow of the tide. Another method exploiting the power of the sea involves the placing of huge floats on the surface of the water which then convert the movement of the waves into power to drive generators.

Experiments to convert wind energy are also being carried out. However, it seems that harnessing wind energy is likely to be expensive and unreliable. Between fifty and a hundred large windmills would be needed to replace one small conventional power station.

Energy reaching the earth from the sun is many thousands of times greater than the world's total energy requirement. This can be harnessed directly to provide energy for individual buildings and houses.

Another interesting possibility for helping to satisfy future energy requirements involves the harnessing of natural heat sources beneath the earth's crust - geothermal energy.

8

NO TEST MATERIAL ON THIS PAGE

9

PART 3: THE RECOVERY OF THE MARY ROSE

You are advised to spend about 30 minutes on Questions 19 - 40.

Questions 19 - 27

Study the section "Method of Recovery" in Reading Passage 3, "The Recovery of the Mary Rose", on Pages 13 - 14. This describes the proposed steps in the raising of the hull of the Mary Rose. It also describes two stages when things might go wrong. These are listed as "Danger No 1" and "Danger No 2" on the opposite page.

Match the steps and dangers listed opposite with sentences A - L below. Some of the steps have been filled in for you.

A. Crane moves Mary Rose and ULF over cradle.

B. Transfer weight of hull from ULF to cradle.

C. Drain rubber bags and inflate with air.

D. Fit ULF into cradle and lower hull.

E. Hull collapses because of water pressure.

F. Position lifting frame over Mary Rose.

G. Transfer hull to barge.

H. Fix wires to hull and lifting frame.

I. Mary Rose and cradle raised by crane.

J. Hull does not fit shape of cradle.

K. Raise hull from seabed by jacking up ULF.

L. Fill protective mattress with water.

Method of Recovery

			Your answers
Stage 1			
19.	Step 1		F
	Step 2	*example:*	H
20.	Step 3		K
21.	Danger No 1		E
Stage 2			
22.	Step 4		A
	Step 5	*example:*	D
23.	Danger No 2		J
Stage 3			
24.	Step 6		L
	Step 7	*example:*	B
25.	Step 8		I
26.	Step 9		C
27.	Final recovery operation		G

NO TEST MATERIAL
ON THIS PAGE

READING PASSAGE 3: THE RECOVERY OF THE MARY ROSE

The Mary Rose was a warship constructed in 1511 for King Henry VIII of England. She sank off the English coast in 1545. For many years she lay forgotten on the seabed, but in 1967 her hull was discovered, buried in mud. In 1982, amid worldwide publicity, she was raised to the surface and taken safely to harbour in Portsmouth.

The passage below is taken from an article describing the planning and the process of the recovery of the Mary Rose. You will not need to understand the meaning of every technical term to complete the questions.

Method of Recovery

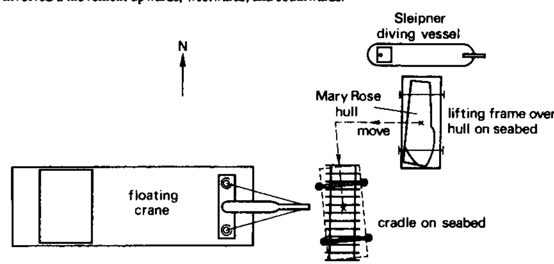
The operation had the following sequence, as planned:

1. Place the steel ULF (underwater lifting frame) on the seabed immediately above the hull, the four legs seated each in a steel foot which in turn is founded in an excavated pit in the seabed.

2. Attach the hull to the horizontal transom members by a diagonal system of 12mm diameter steel wires.

3. Raise the horizontal frame by jacklifting against the legs and through the tension in the suspension wires 'lift off' the Mary Rose clear of the seabed. This would be the first moment of truth because, at this point, inspection would reveal whether or not the Mary Rose was capable of carrying its own weight under water.

4. Attach slings from the giant crane to the ULF, lift ULF, with hull slung underneath, clear of its feet and the seabed and move the whole mass to a position where the four legs of the ULF were immediately above the stabbing guides of the steel cradle. This manoeuvre, in simple terms, involved a movement upwards, westwards, and southwards.



5. Lower the legs within the stabbing guides into the sockets provided. Once it was established that all four legs of the ULF were bearing on the cradle the horizontal table of the ULF could be slowly jacked down to lower the hull of the Mary Rose to a position as close as practical to the cradle. This was the second moment of truth because, at this point, inspection would reveal to what degree of accuracy the shape of the cradle had been designed to receive the hull.

6. The mattress of rubber bags lining the cradle would then be inflated with water to provide a positive cushion between the cradle and the hull, following which the ULF would continue to be jacked down until the slackness in the suspension system indicated that the weight of the hull was bearing on the cradle.

7. The main slings from the crane could now be rigged to the slings previously fitted to the cradle and the final lift commenced.

8. As the hull broke the surface it was intended to pump out the water in the mattress under the hull and replace it with air so that the load on the crane (and the cradle) would be at a practical minimum, i.e. a water-filled mattress in water and an air-filled mattress in air.

9. The cradle would be raised above sea level sufficient to manoeuvre a barge underneath which would carry the Mary Rose on the final leg of her journey to Portsmouth.

Underwater Transfer

The first attempt to transfer the hull into the cradle started on the morning of Saturday 9 October 1982.

The operation was controlled and directed by the Mary Rose Trust officers operating from the diving ship Sleipner, who were in direct and secure contact with the crane master and were relying solely on a technique of remote control and sonar docking, with inspections from time to time by Mary Rose Trust divers.

During this operation, violent underwater collisions between ULF and cradle were observed by eye witnesses, and on other occasions the attitude of the ULF was observed to both tilt and yawl indicating that the ULF (or more likely its legs) was trapped in some way underwater.

It was abundantly clear that the technique of sonar docking and remote control from Sleipner would have to be abandoned. Furthermore, the command and control of the operation would have to be changed, and the alternative (and preferred) method of effecting the transfer employed.

This method was to have one diver in the water at each leg position monitoring the dimensional relationship between ULF and cradle and reporting this information by diver's telephone to Captain Brannam on the crane. From this information, and his own visual observation, the crane master would be able to dock the ULF into the cradle.

It is significant that the docking was thus effected with comparative ease, although it was obvious that only three legs were in the stabbing guides, the fourth being severely bent as a result of the underwater collisions.

Subsequent inspection by divers revealed that two legs of the ULF had been damaged. The SE leg slightly, but the NE leg really demanded that the recovery operation be delayed until the severe damage could be properly repaired or the leg replaced. But the weather forecast equally demanded a solution to the problem without further delay, not to mention the influence of world publicity.

At an impromptu meeting held on board the crane at 11 a.m. on Sunday 10 October it was decided that the most expedient solution to the problem in the circumstances was to cut off the NE leg and stabilise that corner of the frame by introducing an additional sling rigged to the main lifting strops.

Unfortunately, the removal of the NE leg meant that the jacking down operation to lower the hull into the cradle could no longer be employed.

The next dilemma was to accept the necessity, due solely to time constraints, and the consequences of the amputated NE leg, to lower the hull into the cradle without filling the mattress with water.

Time was pressing - the start of the final lift was literally hours away, so the engineers contrived a plan to inflate the mattress with some air as soon as could be effected, and complete the inflation, as previously envisaged, as the hull broke surface.

The Final Lift

The final lift operation started at 8 a.m. on Monday 11 October 1982, with the Mary Rose still hanging from the ULF, and, as a result of the ULF having only three legs, the distribution of load was indeterminate.

During the morning the air supply failed on several occasions, which meant that the air mattress was inflating very slowly and not transferring the load of the hull to the cradle as the hull came through the surface.

The weather forecast was not encouraging, and there was an urgency abroad to complete the lift. Although some air was getting through into the mattress, it was obvious that the ULF in the SE corner was still carrying the hull - which by now was partially out of the sea - a sevenfold increase in hull weight.

It was clear that the lift could not proceed much further unless the Mary Rose was properly settled into the cradle. Either the operation had to wait until the air mattress did the job or the hull had to be induced into the cradle.

The only practical way, at this juncture, to achieve this would be to cut the wires connecting the Mary Rose to the ULF (by now considerably distorted) and let the hull drop into its designed position in the cradle. However, this decision was never taken because, as the hull was raised higher, albeit slowly, the load on the ULF overtook the buoyancy in the air mattress and the 'safety pin' in the leg at the SE corner failed.

A dramatic moment, perhaps the best remembered event of the whole operation (certainly, the most recorded), yet, in fact, 'the safety pin' failing served successfully to induce the Mary Rose into her rightful and designed place - in the cradle. The air supply was no longer critical. The ULF, relieved of its load, was the passenger it was always intended to be. The safety pins, smaller than provided, were no longer inadequate. The lift could proceed.

Glossary

hull:	the body or frame of a ship
crane:	machine for moving heavy weights
cradle:	framework on which a boat can rest during repairs
sling:	belt or strap used to support hanging weight
mattress:	case stuffed with feathers, straw etc., or filled with air or water as bed or support for bed
barge:	flat bottomed boat

Questions 28 - 40

The passage opposite summarises the sections "Underwater Transfer" and "The Final Lift" on Pages 14 - 15.

Decide which phrase should go in each gap and then write the letter in the space provided. You may use any phrase more than once if you like. Note that there are more phrases than gaps.

The first one has been done as an example.

A. underwater telephone

B. adverse weather conditions

C. jacks

D. remote control and sonar docking

E. design concept

F. morse code

G. monitoring the position of the frame

H. extra stabilising

I. safety pin

J. giant crane

K. north-east

L. cradle

M. coming into violent collision

N. sevenfold increase in hull weight

O. its intended position

P. filling the protective mattress with water

Q. failure of the air supply

R. south-east

S. stabbing guides

16

Lifting the Mary Rose

During the first attempt to raise the Mary Rose, the operation was controlled by officers operating from a diving ship and relying on ..*(example)*.. techniques. The decision to change tactics was forced by the fact that the lifting frame and the cradle were ..28.. .

The operation proceeded with divers ..29.. and reporting back to the crane operator by means of ..30.. . The docking was accomplished successfully, but it became clear that not all the legs of the lifting frame had been engaged in the ..31.. of the cradle: the earlier underwater collisions had severely bent the ..32.. leg.

Because of the threat of ..33.. there was no time to effect the proper repairs, so the damaged leg was removed and ..34.. introduced to the frame. However, this meant that it was no longer possible to lower the lifting frame by the ..35.. and thus transfer the hull to the ..36.. . The problem was compounded by the impossibility of ..37.. , and an alternative plan of inflating it with air while still under water was adopted. The ..38.. during the lifting operation meant that, as the hull emerged from the sea, its weight had not been transferred to the cradle but was still being carried by one part of the lifting frame. In a dramatic moment the ..39.. in this part of the frame failed and the hull dropped, but fortunately it fell into ..40.. in the cradle and the task of recovering the Mary Rose could be continued as planned.

Your answers

example: ..D..

28. ..M..
29. ..G..
30. ..A..
31. ..S..
32. ..K..
33. ..B..
34. ..H..
35. ..C..
36. ..L..
37. ..P..
38. ..Q..
39. ..I..
40. ..O..

17

Name: ....

Institution: ....

Date: ....

INTERNATIONAL ENGLISH LANGUAGE
TESTING SYSTEM

MODULE B

SPECIMEN VERSION

Time allowed: 55 minutes (Reading)

SECTION 1: READING

In this section you will find 4 reading passages. Each of these will be accompanied by some questions. Some of the questions will come before the relevant reading passage; some will come after the passage.

Start at the beginning of the section. If you cannot do one part of the test in the suggested time, leave it and start on the next.

University of Cambridge Local Examinations Syndicate
British Council
International Development Program

SECTION 1: READING

READING PASSAGE 1: THE MYSTERY OF DECLINING TOOTH DECAY

Fluoridation consists of raising the concentration of the fluoride ion F^- in water supplies to about 1 part per million with the aim of reducing dental caries (tooth decay) in children. In fluoridated areas, there are now many longitudinal studies (studies over time) which record large reductions in the incidence of caries.The results of these and of fixed time surveys have led to the 'fluoridation hypothesis', namely that the principal cause of these reductions is fluoridation.

Until the early 1980s, there had been comparatively few longitudinal studies of caries in unfluoridated communities. Only a small minority of the studies in fluoridated areas had regularly examined control populations, and there seemed to be little motivation to study other unfluoridated communities. But during the period 1979-81, especially in western Europe where there is little fluoridation, a number of dental examinations were made and compared with surveys carried out a decade or so before. It soon became clear that large reductions in caries had been occurring in unfluoridated areas also.

Caries in unfluoridated areas

Over 20 studies report substantial temporal reductions in caries in children's permanent teeth in unfluoridated areas of the developed world. In many of these cases, the magnitudes of these reductions are comparable with those observed in fluoridated areas and attributed to fluoridation.

Several of these studies give clues as to factors which are unlikely to be the main causes of the reductions. A comparison of dental health surveys in Brisbane (1954 and 1977) indicates a reduction of about 50% in caries, as measured by the number of decayed, missing and filled permanent teeth per child and averaged over the age groups, in the 23-year period. The 1977 survey distinguished between children who took fluoride tablets regularly, irregularly or not at all. Although there were differences in caries incidences between the three categories (which could reflect factors unrelated to fluoride levels), even the "no tablet" group had on average 40% less caries experience than that recorded in 1954. So fluoride tablets were not the principal cause of the reductions observed in Brisbane.

A study in the partly fluoridated city of Auckland, New Zealand, examined the influence of social class (which reflects environmental and lifestyle factors, such as diet) as well as fluoridation on dental health as measured by the levels of dental treatment received by children. The paper showed that treatment levels have continued to decline in both fluoridated and unfluoridated parts of the city and that these reductions are related strongly to social class, there being less caries in the "above average social rank" group than in other children. Thus the main ethical argument for fluoridation, that it should assist the disadvantaged, is not borne out by this study.

The problem of explaining the decline in teeth decay goes well beyond the field of dentistry: contributions from nutritionists, immunologists, bacteriologists, epidemiologists and mathematical statisticians, amongst others, may be required.

PART 1. THE MYSTERY OF DECLINING TOOTH DECAY

You are advised to spend about 15 minutes on Questions 1 - 18.

Questions 1 - 18

The paragraph below is a SUMMARY of Reading Passage 1 on the opposite page. Complete the summary by writing ONE or TWO words in each space. These must be taken from the reading passage. The first one has been done as an example.

	Your answers
Fluoridation is the addition of ..(*example*).. to	*example* : fluoride
water supplies. The ..1.. that fluoridation is the	1. hypothesis
..2.. of declining ..3.. in children was supported	2. (main)(principal) cause
by many early ..4.. which found substantial ..5..	3. tooth decay/ (dental) caries
of caries in ..6.. communities. Later studies,	4. (longitudinal) studies/surveys
however, revealed a ..7.. pattern in ..8.. areas	5. reduction(s)/ decline
of the ..9.. . Other studies gave ..10.. as to	6. fluoridated
the ..11.. which may or may not be involved.	7. similar/ comparable
A Brisbane study of fluoride given in tablet form	8. unfluoridated
reported that over a period of 23 years there was	9. developed world/ western world/ western Europe
..12.. per cent less caries in all children, but in	10. clues
the group of children taking ..13.. fluoride the	11. factors
reductions were 40 per cent, indicating that ..14..	12. 50
could not be the main cause. A New Zealand study	13. no (tablet)
found that in all areas there was a decrease	14. (lack of) fluoride tablets/fluoridation
in levels of ..15.. which correlated highly with	15. (dental)treatment
..16.. . Explaining the complex causes of the	16. social class/rank contributions
decline in caries will require ..17.. from academic	17. studies
disciplines other than ..18.. .	18. dentistry

3

NO TEST MATERIAL ON THIS PAGE

4

PART 2: OUR CHILDREN'S TEETH

You are advised to spend about 15 minutes on Questions 19 - 27.

Questions 19 - 23

Look at Reading Passage 2, "Our Children's Teeth", on Page 7 and answer the following questions.

Choose which of the alternatives A, B, C, or D is the correct answer and put the appropriate letter in the space provided. The first one has been done as an example.

Your answers

example: Compared to 25 years ago British children now have

A. much healthier teeth.
B. slightly healthier teeth.
C. much less healthy teeth.
D. slightly less healthy teeth.

example: ...A...

19. The author finds it "disturbing" (line 8) that

A. not enough data are being collected.
B. tooth decay is commoner in some social groups.
C. advantaged children receive more fillings.
D. the treatment of tooth decay is varied.

19. ...B...

20. "Dental caries has become a class related disease" (line 9). The best evidence for this given by the author is the high rate of decay in

A. Scotland and Northern Ireland.
B. the north of England.
C. districts in the north west.
D. poorer areas of London.

20. ...D...

Your answers

21. "The changed pattern in the prevalence of caries" (lines 21 - 22) refers especially to the difference between

A. treated and untreated groups.
B. younger children and older children.
C. poor and affluent communities.
D. fluoridated and unfluoridated areas.

21. ...C...

22. The author believes that the social gradients in the prevalence of dental caries are due to differences in

A. geographical location.
B. water supplies.
C. preventive dental care.
D. general health.

22. ...C...

23. The Health Education Council recommended fluoride toothpaste for

A. adults eating foods containing sugar.
B. children not taking fluoride drops or tablets.
C. children in unfluoridated areas.
D. adults and children in general.

23. ...D...

READING PASSAGE 2 : OUR CHILDREN'S TEETH

The dental health of British children has improved dramatically in the past 25 years. In England and Wales the prevalence of dental caries fell in all age groups and in all geographic regions between 1973 and 1983. For example, among 5 year olds the mean number of decayed teeth was 3.4 in 1973 and 1.7 in 1983, while over the same decade the percentage of children in this age group completely free from decay rose from 29% to 52%. A further survey is expected in 1993, and data on caries are currently being collected routinely by health districts in England and Wales. All these surveys will, it is hoped, show further improvement in the dental health of children, but there are some disturbing trends.

Dental caries has become a class related disease that is commoner in poorer communities. This is a recent change as before 1970 the prevalence of decay was similar in all social groups and only its treatment varied: advantaged children received more fillings and fewer extractions than their less fortunate contemporaries. Now children from affluent backgrounds have less decay, producing appreciable national, regional, and local variations.

In 1983 the mean number of decayed teeth in 15 year olds was 5.6 in England, 6.7 in Wales, 8.4 in Scotland, and 9.2 in Northern Ireland. Within England regional differences were striking, with the lowest prevalence in the south and the highest in the north. During 1985-6, 5 year olds in West Essex had only 0.75 decayed teeth on average, whereas in several districts in the north west this value exceeded 3. That such differences are socially determined is shown most particularly by the high percentage of decay in several less affluent inner city areas in London.

Changed practices of preventive dental care probably explain this changed pattern in the prevalence of caries. The "rules" for dental health are now well established, and the Health Education Council summarised them in four simple statements: food and drink containing sugar should be restricted to mealtimes; the teeth and gums should be cleaned daily with a fluoride toothpaste; children in areas with suboptimal concentrations of fluoride in the drinking water should take fluoride drops or tablets daily; and everybody should attend the dentist regularly. These rules are now widely used by health educators, and the social gradients that exist in the prevalence of dental caries almost certainly arise because the rules are better understood and more widely followed in affluent communities.

Attempts to improve further the dental health of children must take account of these socially determined differences. Water fluoridation must remain the cornerstone of policy because it benefits all consumers regardless of social constraints on health behaviour.

7

A. "the rules are better understood and more widely followed in affluent communities" (lines 28 - 29)

B. "among 5 year olds the mean number of decayed teeth was 3.4 in 1973 and 1.7 in 1983" (lines 3 - 4)

C. "5 year olds in West Essex had only 0.75 decayed teeth on average, whereas in several districts in the north west this value exceeded 3" (lines 17 - 18)

D. "advantaged children received more fillings and fewer extractions than their less fortunate contemporaries" (lines 11 - 12)

E. "data on caries are currently being collected routinely by health districts in England and Wales" (lines 6 - 7)

F. "children in areas with suboptimal concentrations of fluoride in the drinking water should take fluoride drops or tablets daily" (lines 25 - 26)

G. "the mean number of decayed teeth in 15 year olds was 5.6 in England, 6.7 in Wales, 8.4 in Scotland, and 9.2 in Northern Ireland" (lines 14 - 15)

8

Questions 24 - 27

For each of the phrases below an example is provided in Reading Passage 2, "Our Children's Teeth", on Page 7. Decide which statement from the list A - G opposite provides an example of each phrase, and write the letter in the space provided.

The first one has been done as an example.

Your answers

example: the fall in the prevalence of dental caries in all age groups in England and Wales between 1973 and 1983 — *example:* ..B...

24. the variation in treatment in different social groups before 1970 — 24. ..D...

25. national variation in the prevalence of tooth decay in 1983 — 25. ..G...

26. regional differences in dental decay in England — 26. ..C...

27. rules used widely by health educators — 27. ..F...

PART 3: THREE WAYS TO MAKE A TRANSGENIC BEAST

You are advised to spend about 15 minutes on Questions 28 - 36.

Questions 28 - 32

Reading Passage 3, "Three Ways to Make a Transgenic Beast", on Pages 11 - 12 describes three methods of transferring genes in mice.

Question 28 below refers to the first method of transferring genes (injection).

Some of the steps in the first method are listed below (A - E). Write the steps in the correct order. Put the appropriate letters in the spaces provided. The first and the last ones have been done for you.

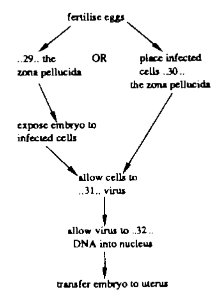
Your answers

28.

A. DNA injected	First step	..C....
B. eggs fertilised	Second step	..B....
C. hormones fed to donor animals	Third step	..E....
D. eggs transferred into oviduct	Fourth step	..A....
E. eggs harvested	Fifth step	..D....

Questions 29 - 32 refer to the second method of transferring genes in mice (retrovirus).

Complete the diagram by finding the correct word from the text to fill each numbered space. (You will find Figure A helps you understand the text.)



Your answers

29. ......remove.........

30. ......inside/in(to).........

31. ......release/spread.........

32 insert/enter/ferry/carry/transfer

READING PASSAGE 3: THREE WAYS TO MAKE A TRANSGENIC BEAST

Researchers have devised three methods of transferring genes in mice, and animal breeders are now seeking to apply these techniques to livestock.

Method 1: Injection

The most common approach is to inject DNA directly into fertilised eggs. Before harvesting the eggs, researchers feed hormones to the donor animals. This causes them to ovulate several eggs at a predictable time so that researchers can time matings and recover the fertilised eggs a few hours later. When egg and sperm meet, each carries only one member of each pair of chromosomes present in the adult animal, and the chromosomes are tightly condensed. Soon after the sperm penetrates the egg, the chromosomes disperse and are contained within small bodies known as pronuclei. The pronuclei swell and migrate toward the centre of the egg.

The goal is to inject genes - sequences of DNA - into a pronucleus. To accomplish this delicate task, a researcher holds the egg on one pipette by gentle suction, and with another fine pipette places copies of the gene in one pronucleus. Approximately 30 per cent of injected embryos degenerate within a few hours. Researchers transfer surviving eggs into the oviduct of recipient females. This is the only proven route for gene transfer in livestock.

Method 2: Retrovirus

But other techniques could work in farm animals. Certain viruses known as retroviruses can ferry genes into cells (see Figure A). They infect cells and insert their DNA into the chromosomes. They replicate in the infected cell and then spread to other cells in the animal. Genetic engineers can disable a retrovirus so that it can still incorporate into the chromosome but cannot replicate and spread.

But retroviruses cannot pass through the coat that surrounds the embryo - the zona pellucida. To overcome this barrier, researchers may remove the zona to expose the cells of the embryo to fibroblasts, which release viruses into the medium. Alternatively, researchers can place cells that release viruses inside the zona, exposing embryos to the virus over a longer period. After either treatment, researchers transfer embryos to the uterus of recipient females.

Using retroviral vectors in livestock has disadvantages. The technique creates off-spring that are usually "mosaic" - where some cells contain the transgene while others do not. Secondly, there is a limit to the size of the foreign DNA that a retrovirus can carry. Finally, we need to be sure that the modified viruses cannot leave the transgenic animals, perhaps regaining the ability to replicate through recombination with wild viruses. Given these difficulties retroviruses may never be used routinely to transfer genes into livestock.

Method 3: Stem Cell

The third way of producing transgenic mice uses special cells from the embryo (see Figure B). The fetus develops from an inner part of a very young embryo (often referred to as a pre-embryo), called the inner cell mass. Martin Evans and Matt Kaufman at the University of Cambridge developed a method to isolate these cells and grow them in such a way that they continue to divide, but do not become specialised or "differentiated". When such "stem cells" are injected into the hollow cavity of another embryo that has reached the blastocyst stage - when the embryo is a hollow ball of cells

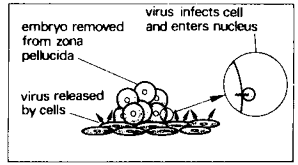


Figure A *Retroviruses can act as carriers of foreign genes. Researchers grow the viruses in fibroblasts, which then shed viruses that infect the cells of the embryo*

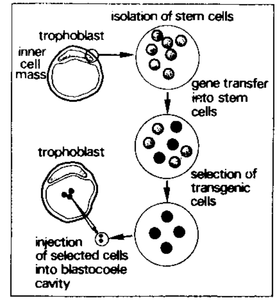


Figure B *Another technique has worked in mice and golden hamsters, and could be applied to livestock. Infected stem cells can ferry genes into embryos*

- the injected cells colonise the embryo and participate in the formation of all of the tissues, including the germ cells from which sperm or eggs develop. Several groups have transferred genes into mice using stem cells.

Researchers use a number of methods to introduce foreign genes into the embryonic stem cells in the first place. Sometimes it is possible to detect transgenic cells in culture before the cells are transferred to another embryo. Because some of these cells retain the ability to colonise the embryo and to form germ cells, they provide a means of introducing foreign DNA into the mouse germline. As with retroviral infection of embryos, transgenic mice produced by this route are mosaic. So far, researchers have grown stem cells only from mouse and golden hamster embryos, but there is no reason to suppose that we cannot develop techniques for farm animals.

Glossary

ferry: transport, carry
fibroblast: type of cell
embryo: fertilised egg, which develops into a young animal
harvest: collect
livestock: farm animals
transgenic: having genes from another animal

Questions 33 - 36

Answer these questions using the whole of Reading Passage 3, "Three Ways to Make a Transgenic Beast", on Pages 11 - 12.

33. Which of the three methods of transferring genes is illustrated in Figure A?

2/retrovirus

34. Which of the methods is illustrated in Figure B?

3/stem cell

35. What is the definition of a mosaic animal?

some cells contain the transgene, others do not

36. According to the author, which of the 3 methods is or are unlikely to be used in farm animals?

A. the injection method
B. the retrovirus method
C. the stem cell method
D. all 3 methods

Your answer

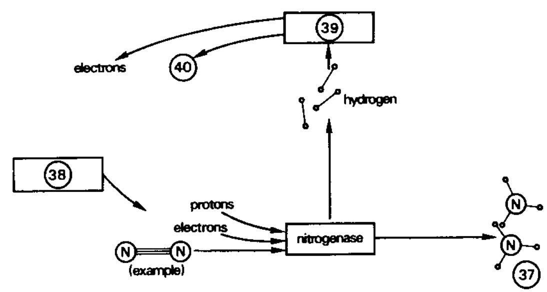
36. B

READING PASSAGE 4: NITROGEN FIXATION

In all nitrogen-fixing organisms the agent responsible for fixation is the enzyme nitrogenase, which catalyzes the conversion of molecular nitrogen into ammonia. In the reaction a transport protein donates electrons to the nitrogenase, which in turn transfers them to the diatomic molecule of nitrogen by a mechanism that is not yet fully understood. Three negatively charged electrons come to be associated with each atom of nitrogen; thereafter three protons (hydrogen nuclei) are withdrawn from the intracellular medium to neutralize the charge. Hence each diatomic molecule of nitrogen yields two molecules of ammonia.

The transfer of electrons from nitrogenase to molecular nitrogen has an energetically wasteful side reaction. Many of the electrons recombine with protons before they reach the nitrogen; the re-combined electrons and protons are released as molecular hydrogen gas (H_2). Certain strains of *rhizobium* synthesize hydrogenase, an enzyme that converts molecular hydrogen back into electrons and protons for re-use by nitrogenase.

NITROGEN FIXATION IN HYDROGENASE-SYNTHESISING STRAINS OF RHIZOBIUM



Glossary

Nitrogen fixation: the process in which nitrogen in the atmosphere is converted into a form that can be used by plants

Rhizobium: a genus of bacterium which can fix atmospheric nitrogen

PART 4: NITROGEN FIXATION

You are advised to spend about 10 minutes on Questions 37 - 40.

Questions 37-40

The diagram opposite illustrates Reading Passage 4, "Nitrogen Fixation". Complete the diagram by finding the correct word or words **FROM THE TEXT** to fill each numbered space.

Write your answers in the space provided below.

The first one has been done as an example.

Your answers

example: Nitrogen

37. ammonia/nh3/two (diatomic molecules) molecules of ammonia

38. (transport) protein

39. (rhizobium synthesise) hydrogenase

40. (3) protons

Name: ....

Institution: ....

Date: ....

INTERNATIONAL ENGLISH LANGUAGE
TESTING SYSTEM

MODULE C

SPECIMEN VERSION

Time allowed: 55 minutes (Reading)

SECTION 1: READING

In this section you will find 3 reading passages. Each of these will be accompanied by some questions. Some of the questions will come before the relevant reading passage; some will come after the passage.

Start at the beginning of the section. If you cannot do one part of the test in the suggested time, leave it and start on the next.

SECTION 1: READING

PART 1: QUALITY CIRCLES

Questions 1 - 12

You are advised to spend about 15 minutes on Questions 1 - 12.

Refer to Reading Passage 1, ''Quality Circles'', on Pages 4 - 5. Answer the following questions by copying the relevant words or phrases from the text.

The first one has been done as an example.

example: What is the underlying reason for the concern with quality?

Relevant words from the text : increasingly competitive market

What are the FOUR stages in quality circle programmes in which employees are involved? Give the relevant words from the text.

1. analyse
2. solve
3. make recommendations/presents solutions to management
4. implementing

5. If you wanted to examine examples of quality circles, where would you look for information?

Relevant words from the text : record section/Appendix 1/5 case studies

6. Was the idea of quality circles immediately popular when it was first introduced to Britain?

Answer 'Yes' or 'No' No

What words in the text helped you to come to this answer?

Relevant words from the text once dismissed as flavour of the month/ that would never catch on in Britain/by the late 1970's some British companies were starting to take notice

According to the text, quality circles are different from task forces in TWO ways. What are they? Give the relevant words from the text.

(For Questions 7 and 8, write ONE or TWO words only.)

7. One difference: permanent existence/regular meetings

8. Another difference: autonomy

9. Where did the ideas behind quality circles first come from?

Relevant words from the text : the United Sates/U.S./Deeming and Juran

10. When were quality circles first introduced in the United Kingdom?

Relevant words from the text : 1978/(the late) (19)70's

11. What, according to the text, is an important ingredient for the success of quality circles?

Relevant words from the text : the strongest backing of senior management

12. Of the aims of quality circles which are mentioned, TWO seem to form a separate category from the others. Which ones are they?

Relevant words from the text:

1. to produce a higher quality product
2. to reduce costs

READING PASSAGE 1: QUALITY CIRCLES

In an increasingly competitive market, companies are paying more attention to the quality of the products or services they provide. Some organisations are seeking ways of encouraging employee involvement in the company, increasing job satisfaction and motivating staff.

Quality circle programmes are being increasingly adopted by British management as one method of meeting these objectives. Quality circles are small groups of employees from the same work area who meet regularly in order to analyse their work-related problems, and to find solutions to them. They make recommendations to management and, if these are approved, are usually involved in implementing their proposals.

Once dismissed as a 'flavour of the month' management technique or as a Japanese concept that would never catch on in Britain, quality circles now exist in over 400 companies throughout the UK. And they are no longer confined to the manufacturing sector - service organisations have also begun to introduce them.

In this Study we examine how a quality circle programme is initiated and organised. Practical aspects (such as whether circle meetings are held in company time, communications with non-circle members) are considered, and we describe the quality circle problem-solving process. We also look at trade union views. The Record Section (Appendix 1) contains five case studies of companies which have introduced quality circles: three from the manufacturing sector - Black & Decker, Jaguar and Tioxide; and two from the service sector - Honeywell Control Systems and the London Life Association.

What is a Quality Circle?

The National Society of Quality Circles describes a quality circle as 'a group of four to twelve people coming from the same work area, performing similar work, who voluntarily meet on a regular basis in order to analyse and solve their own work-related problems. The circle presents solutions to management and is usually involved in implementing them'.

Unlike other problem-solving groups such as task forces, quality circles have a permanent existence (i.e. they have regular meetings) and they have considerable autonomy as to what problems or ideas they consider.

Where Did They Come From?

Quality circles have gained a reputation as an important ingredient of the modern Japanese economic miracle. But the quality control ideas on which circles are based were developed originally in the United States.

It was in Japan, however, in the early 1960s that the ideas of Deeming and Juran were first put into practice, with Professor Ishikawa, from Tokyo, becoming known as the 'father' of quality circles. Then, in the mid-1970s, the quality circle concept recrossed the Pacific when such American companies as Lockheed, General Motors and Honeywell led the first experiments in the Western world.

By the late 1970s some British companies were starting to take notice of the quality circle phenomenon. Rolls-Royce Aero Division lays claim to being the British pioneer in 1978. By 1981 it was estimated that around 100 companies in the UK had introduced quality circles, and the number today is thought to be at least 400.

THE REASONS WHY

Quality circles are only likely to succeed in companies where the programme has the strongest backing of senior management. So what do companies see as the benefits of a quality circle programme? The aims would certainly include some or perhaps all of the following:

- to produce a higher quality product or service (increasingly important in highly competitive markets);
- to create a greater degree of quality consciousness among all employees;
- to reduce costs;
- to involve employees more in the organisation;
- to increase employees' job satisfaction;
- to develop and motivate staff;
- to achieve better two-way communications;
- to encourage a more open, participative management style;
- to encourage team-working;
- to help create a better quality of working life.

Aims of continuing education

A. Staying fresh

B. Being prepared to carry out one's professional responsibilities

C. Enlarging the range of treatments available

D. Collaborating with other professions

E. Staying in touch with theoretical developments

F. Becoming expert in theoretical knowledge

G. Preparing to earn more money

H. Staying in touch with developments in the nature of the profession

I. Developing as an individual

J. Using knowledge regularly

K. Maintaining the ability to acquire new knowledge

L. Contemplating theory

M. Preparing for career change

N. Staying in touch with developments in professional knowledge

6

PART 2: THE PURPOSES OF CONTINUING EDUCATION

You are advised to spend about 20 minutes on Questions 13 - 27.

Questions 13 - 19

Reading Passage 2, "The Purposes of Continuing Education", on Pages 8- 9, discusses eight different aims of continuing career education. From the list of fourteen titles (A - N) opposite, choose the most suitable title for each of these aims.

The first one has been done as an example.

Your answers

example: Aim 1 ..N..

13. Aim 2 ..E..

14. Aim 3 ..H..

15. Aim 4 ..M..

16. Aim 5 ..A..

17. Aim 6 ..I..

18. Aim 7 ..K..

19. Aim 8 ..B..

7

READING PASSAGE 2: THE PURPOSES OF CONTINUING EDUCATION

What are the purposes of continuing career education? Some people think about the matter very simplistically, but most people realize that several goals must be sought simultaneously and that the process of doing this is difficult. We may identify at least eight purposes of continuing professional education.

The first of these is to keep up with the new knowledge required to perform responsibly in the chosen career. Just think, for example, how much has happened in the various professions in ten years. At the start of the decade, we had just learned how to keep a man in orbit around the earth; at its end, we had sent many men to the moon. At the same time, physicians and surgeons learned to save the lives of people who in previous times would have died. Patterns of resource allocation applied by public administrators are changing as a result of new doctrines. And pharmacists - perhaps more than the members of any other profession - have had to keep up with a changing scene, as new products and new therapies succeeded one another or enlarged the range of treatments available.

Practical careers rest upon theoretical bodies of knowledge. The health professions, including pharmacy, depend upon anatomy, physiology, pathology and biochemistry; agriculture is based upon soil chemistry, meteorology and plant pathology; and social work finds its roots in sociology, psychiatry and political science. The professional does not need to become expert in these underlying bodies of knowledge, but he or she does need to learn about developments in these fields. Keeping up with changes in the relevant basic disciplines, therefore, constitutes the next aim.

The third aim is to master new ideas about the nature of the profession itself. A recent study of dentistry noted that "25 or 30 years ago, dental practice was limited to relieving pain. Today it is concerned with the overall management of oral, facial and speech defects, and with the structures and tissues of the mouth as they relate to the total health of the individual." In other occupations, an equally profound change has occurred and the modern practitioner who does not understand that fact is obsolete.

The fourth purpose of career-oriented continuing education is to prepare the individual for changes in his or her career. A person may move in one of many directions, such as from generalist to specialist, from one speciality to another, upward in a hierarchy from a lesser to a more responsible job of the same sort, or into a completely new career.

In every kind of career, people go stale after a while. Consequently, maintaining freshness of outlook on one's work, so that detail is not neglected, is also important. This forms the next aim. Perhaps education is not the only way to achieve this, but it can at least be aided by educational means : for example, by supervisory training, by self-appraisal and by peer review. It can also be achieved by putting oneself in a new work or study situation which demands attention to detail.

Continuing to grow as a well-rounded person is essential as well. The mind should never be fully engaged in the practice of work, to the exclusion of all other matters. One needs to withdraw from that practice occasionally so that one can be stimulated by contemplating theory or by seeking understanding and skill in different aspects of life. Otherwise, one becomes dogmatic, not only about a single speciality but about all aspects of life. This, then, is another aim of continuing professional education.

8

The seventh objective is to retain the power to learn. This objective is an important adjunct to all the others. The skills of mastering knowledge are like other skills : they are learned by practice, they are lost if they are not used regularly, and they can later be regained only with difficulty.

Finally, it is essential to ensure that the individual performs effectively the role imposed by membership in a profession. The professional must learn how to take collective responsibility, to make right choices on issues, to improve and extend the delivery of service, to collaborate with other professions, and to help monitor the actions of fellow professionals.

These eight purposes suggest that a fully fledged program of continuing professional education cannot be developed either easily or rapidly.

9

Questions 20 - 27

The passage opposite is a summary of "The Purposes of Continuing Education" on Pages 8 - 9.

Decide which word or phrase should go in each gap and then write the letter in the space provided. Note that there are more phrases than gaps.

The first one has been done as an example.

A. an example

B. theoretical knowledge

C. a thorough program

D. educational in nature

E. modern technology

F. dismissed

G. out of date

H. continuing education

I. career

J. personal growth

K. awareness

L. professional knowledge

Summary of "The Purposes of Continuing Education"

Your answers

Continuing professional education probably has at least eight different purposes. Achieving a balance between these is not easy.

It is important for a member of a profession to keep his or her ..*(example)*.. up to date. The need for this can be seen from space technology to medicine, from public administration to pharmacy. In all of these fields, rapid changes are taking place. Keeping one's ..20.. up to date is also important. Although professionals do not need to understand in detail the developments which are taking place in the theories which underlie their profession, it is important that they should have some ..21.. of what is happening. Keeping in touch with developments in the fundamental character of a profession constitutes the third objective of continuing education. In some occupations, considerable changes have occurred and a professional who is not able to adapt to such changes will soon be ..22.. .

example: ....L....

20. ....B....

21. ....K....

22. ....G....

Professionals frequently need to make moves in their careers. Taking on a position with greater responsibility is ..23.. . Continuing education can prepare people to make such moves.

23. ....A....

The next aim is concerned with enabling people to stay professionally fresh. There are several ways of achieving this aim, some of which are ..24.. and some of which are not. A related aim is to encourage ..25.. . There are dangers in becoming too engrossed in one's work and, from time to time, the professional needs to stand back from routine work. Continuing professional education can facilitate this process.

24. ....D....

25. ....J....

The seventh objective is to help the individual to continue to be able to learn. This is important because, if the skill of acquiring new knowledge is not practised regularly, it may be lost. Clearly, this particular aim of ..26.. is related to all the others which we have been discussing. The final aim concerns the role which a professional person has to play in society at large. The individual has to learn to fulfil all those social functions which are expected of members of the profession.

26. ....H....

It is not a simple task to design ..27.. of continuing professional education. Furthermore, such a program cannot be carried out in a short space of time.

27. ....C....

PART 3: ACCESS TO HIGHER EDUCATION

You are advised to spend about 20 minutes on Questions 28 - 35.

Questions 28 - 31

Answer these questions using Reading Passage 3, "Access to Higher Education", on Pages 14 - 17. Choose which of the alternatives A, B, C, or D is the correct answer and put the appropriate letter in the space provided.

The first one has been done as an example.

Your answers

example: Since 1979 the number of full-time home students in higher education in Great Britain

A. has nearly trebled.

B. has increased more than in the 1970s.

C. has reached 85,000.

D. has grown by 25%.

example: ...B...

28. What is meant by "both sectors" (paragraph 1.2)?

A. "Polytechnics and Universities" and "Colleges"

B. "Polytechnics" and "Colleges"

C. "Polytechnics" and "Universities"

D. "Polytechnics and Colleges" and "Universities"

28. ...D...

29. It is not easy to compare participation rates among different countries, because

A. there are too many differences in the systems.

B. pupils enter higher education at different ages in different countries.

C. comparisons cannot be based on reputation.

D. Britain and France are ahead of the rest of the European Community.

29. ...A...

Your answers

30. The British government believes that in the future

A. there will be no increase in numbers of students.

B. the number of women students will continue to rise.

C. its educational policies to increase student numbers will succeed.

D. progress in Scotland may be more rapid.

30. ...C...

31. According to Projection Q, an increasing number of young people

A. will definitely be more qualified.

B. will have the intellectual competence to benefit from higher education.

C. will have only technical or vocational qualifications.

D. are probably going to be more qualified.

31. ...D...

READING PASSAGE 3: ACCESS TO HIGHER EDUCATION

1 The Record

1.1 Since 1979 the number of full-time home students in higher education in Great Britain has risen by more than 85,000 - almost three times the increase achieved during the 1970s. The size of the 18-19 year-old age group peaked in 1982; the continuing increase in student numbers reflects higher rates of participation in higher education both by 18-19 year-olds - for whom the Age Participation Index increased from 12.4 in 1979 to an estimated 14.2 in 1986 - and by mature entrants (aged 21 and over) whose numbers have grown by a quarter since 1979. The increases in participation have been particularly marked for women, who now account for about 44% of full-time students in higher education compared with less than 40% seven years ago.

1.2 Virtually all of this major increase in full-time student numbers has taken place in the polytechnics and colleges sector of higher education, more than making good the reduction in this sector in the late 1970s. The polytechnics and colleges have also accommodated over three-quarters of the 72,000 (27%) increase since 1979 in part-time student numbers. In both sectors the proportion of students on full-time science-related courses has increased - from 50% to 53% for universities and from 36% to 41% for the polytechnics and colleges.

1.3 Comparisons with higher education participation rates in other countries are not straightforward. There are differences in structures, course lengths, wastage rates and ways of counting part-time students. The best like-for-like comparison is probably of the proportion of the relevant age groups gaining degrees and higher diplomas. On that basis Britain is on a par with attainments in France and ahead of the rest of the European Community, though achievements here are not as good as those in Japan and the USA. Moreover, such comparisons take no account of quality - for which the reputation of higher education in this country is high.

2 The Future

2.1 In November 1986 the Government published *Projections of Demand for Higher Education in Great Britain 1986-2000*, which displays two projections of future student numbers. Projection P, the lower of the two, is based on the assumptions that the numbers of young people entering full-time higher education will remain a constant proportion of those gaining the traditional qualifications for entry (two or more GCE A levels or three or more Scottish Highers), and that the entry rates for mature students will also remain constant. For the higher Projection Q, it is assumed that those proportions will increase - particularly amongst young women - in part to reflect the success of the Government's policies for schools and non-advanced further education.

2.2 Under both projections the Age Participation Index (API) for young entrants is envisaged to increase by the end of the century : to 15.8 under Projection P and to 18.5 under Projection Q. Subsequent reports from research commissioned by the Department of Education and Science have confirmed that young people whose parents hold higher-level qualifications are proportionately more likely to apply for and obtain places in higher education and that this factor is additional to the differential demand for higher education by social class which is already allowed for in the projections. This may have a significant influence on demand for higher education in the 1990s when many of the children of those who benefited from the big increase in higher education opportunities in the 1960s will reach age 18. If so, the API for young entrants is likely to be nearer to that underlying Projection Q.

2.3 Both Projections P and Q assume that higher proportions of young people will gain the traditional qualifications for entry to higher education as a result of girls achieving parity with boys' A level attainments, and changes in the social/occupational mix of the population. Additionally, Projection Q assumes a further increase in the potential number of traditionally qualified entrants to higher education, dependent primarily on the success of the policies presented in the Government's *Better Schools* document which imply that a significantly larger proportion of pupils at age 16 will reach the standards necessary to continue on to and to succeed in A level courses. The Government believes that by the end of the century improvements in the schools will be such that the proportion of young people in Great Britain as a whole gaining two or more A levels, or three or more Scottish Highers, will have reached 20% - the proportion already attained in Scotland.

2.4 Of potentially greater impact, however, is the assumption underlying Projection Q that there will be a significant increase in the proportion of qualified young people who enter higher education. This will depend very largely on continuing growth in the demand for higher education from young women, alongside increases in the proportion of higher education entrants with vocational and technical qualifications, for example those validated by the Business and Technician Education Council (BTEC) and the Scottish Vocational Education Council (SCOTVEC). The Government is committed to ensuring that girls have equal opportunities, throughout the education system, to develop their talents to the full. The development of the Technical and Vocational Education Initiative (TVEI) and the two-year Youth Training Scheme (YTS), and the streamlining associated with the National Council for Vocational Qualifications (NCVQ) and the Scottish Action Plan, should increase the proportion of young people gaining vocational qualifications and is likely to motivate more of them to seek entry to higher education.

2.5 The Government remains committed to the modified form of the Robbins Principle. Places should be available for all who have the necessary intellectual competence, motivation and maturity to benefit from higher education and who wish to do so. Planning of higher education will need to take account, inter alia, of regular monitoring of actual demand for places and of the effects of the Government's policies to improve performance in schools and non-advanced further education on the numbers of potential entrants to higher education.

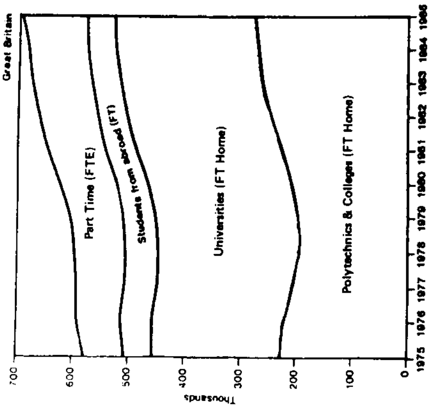


FIGURE A

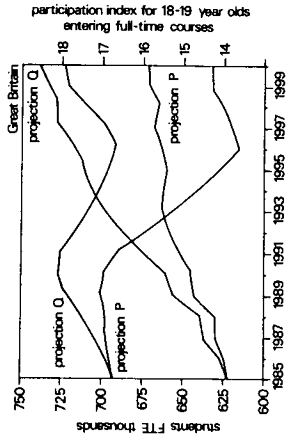


FIGURE B

16

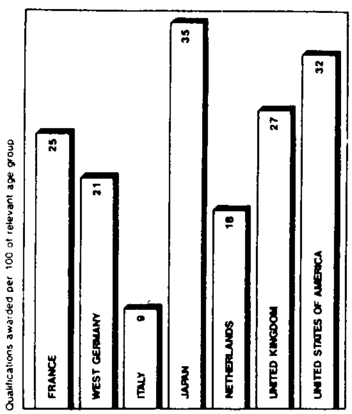


FIGURE C

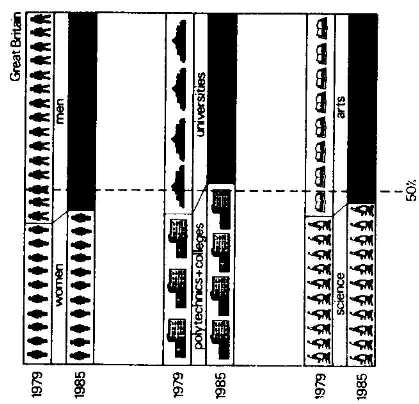


FIGURE D

17

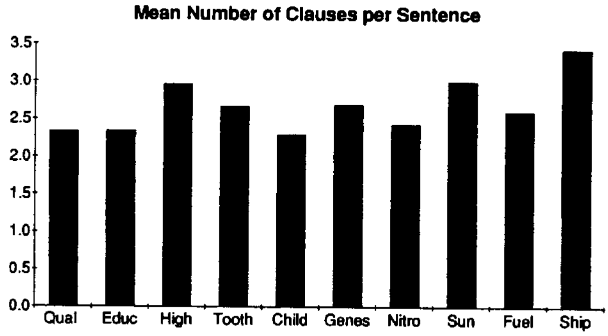
Questions 32-35

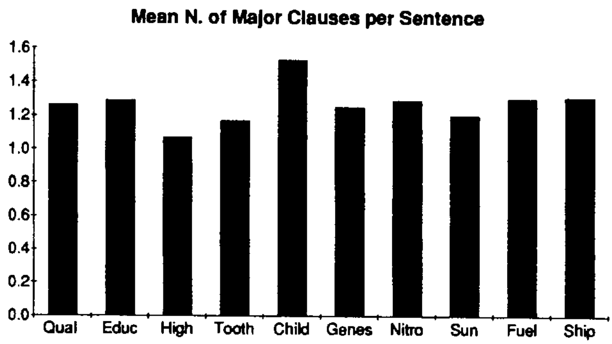
Refer to Reading Passage 3, "Access to Higher Education", on Pages 14 - 15. Look also at Figures A - D, which follow the Reading Passage on Pages 16 - 17. The figures can be matched with the appropriate paragraphs of the reading passage. Note that the figures may refer to more than one paragraph, and that one paragraph may be referred to in more than one figure.

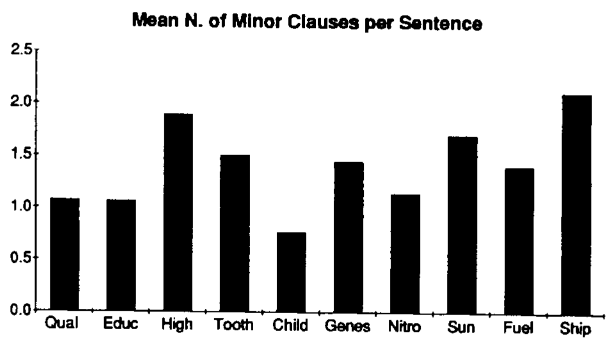
Fill in the table below, inserting an X under each figure where it refers to a given paragraph. In the example below, you will see that Figures A and D both refer to Paragraph 1.1. Therefore an X has been put in the appropriate boxes. Figures B and C do not refer to Paragraph 1.1, so the appropriate boxes have been left empty.

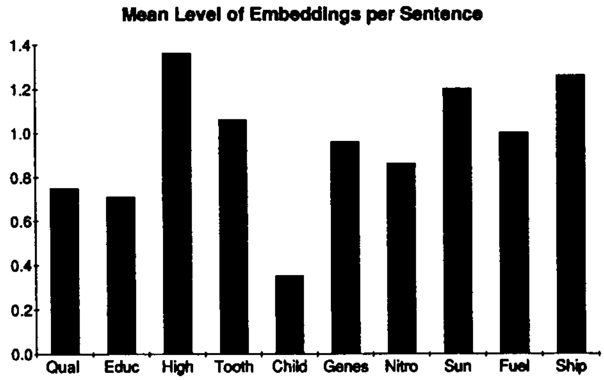
	Paragraph	Figure A	Figure B	Figure C	Figure D
example:	1.1	X			X
32.	1.2				
33.	1.3				
34.	2.1				
35.	2.5				

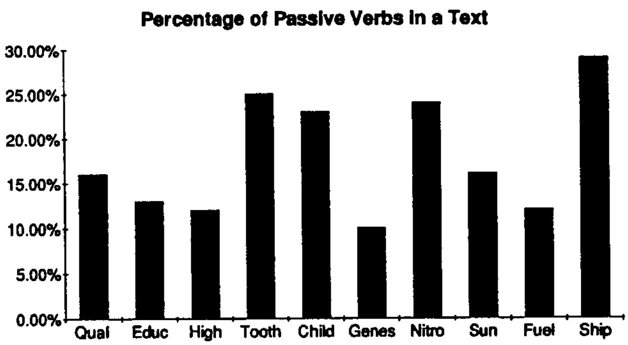
Appendix 6.2 Clausal Analysis Graphs








Mean Level of Embeddings per Sentence
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Qual
Educ
High
Tooth
Child
Genes
Nitro
Sun
Fuel
Ship


Percentage of Passive Verbs in a Text
30.00%
25.00%
20.00%
15.00%
10.00%
5.00%
0.00%
Qual
Educ
High
Tooth
Child
Genes
Nitro
Sun
Fuel
Ship

Appendix 6.3 The Grammar Test

Name: ....

Institution: ....

Date: ....

G1 (GRAMMAR)

TRIAL TEST VERSION 3

Time allowed: 20 minutes

Instructions

In this booklet you will find 6 exercises.

Start at the beginning of the test and work through it. If you find an exercise too difficult, go on to the next one. You can return to the difficult one later if you have time.

Please read the instructions for each exercise carefully.

University of Cambridge Local Examinations Syndicate
British Council
International Development Program

1

EXERCISE 1

Questions 1 - 4

In Paragraph 1 below, some words are in bold type (**like this**). If you looked them up in the dictionary, you would find that each word has several possible meanings. Choose the closest meaning for each <u>as it is used in the sentence in Paragraph 1</u>.

Write the letter in the space pr·· ided. The first one is done for you.

Unidentified Flying Objects

<u>Paragraph 1</u>

One day in the summer of 1947, an American pilot, Kenneth Arnold, was flying his plane over some mountains in the USA. Suddenly he looked up and **noticed** some lights to his left. The **lights** came from nine **objects** flying in a straight line **like** a **chain**. Each one was round, shining and as big as a plane. He thought that their flying speed was between 1,600 and 2,700 kilometers per hour. The objects seemed to be able to move up and down as they travelled. No aeroplane could do that in 1947.

example: **notice** — **Your answers**

A. an announcement or warning in advance
B. a general announcement giving information
C. to become aware of something
D. to be informed that you must leave your job

example: ......C......

1. **light**

A. opposite of dark
B. something that does not weigh very much
C. something that produces or reflects light
D. a place that has a lot of natural light

1. ......C......

2. **object**

A. aim or purpose
B. something that has a fixed shape or form
C. oppose or disapprove
D. trying to be fair

2. ......B......

3. **like**

A. showing similar qualities; comparing
B. something you enjoy or find attractive
C. prefer something
D. something that is satisfactory

3. ......A......

4. **chain**

A. a range of mountains
B. fasten something up
C. a number of similar shops or hotels
D. a number of things arranged in a row

4. ......D......

EXERCISE 2

Questions 5 - 9

Look at Paragraph 2. The words in bold type (**like this**) are not in their correct form. Write the correct form in the spaces provided.

One of them has been done for you.

Unidentified Flying Objects

<u>Paragraph 2</u>

When some **reports** asked him questions about what he had seen he **say** that the objects moved like flat stones or saucers thrown across water. The next day the newspapers **report** that Kenneth Arnold had seen "flying saucers" over the mountains. He had been **describe** their movements, but the journalists **misunderstand** and thought he was describing their **appear**. Ever since then the phrase "flying saucer" has come into use as a part of the language. Many strange objects have been seen since that time. So now they are called "Unidentified Flying Objects" and people simply treat the short form as a new word - "UFO".

		Your answers
5.	**reports**	reporters
example:	**say**	said
6.	**report**	reported
7.	**describe**	describing
8.	**misunderstand**	(had) misunderstood
9.	**appear**	appearance

EXERCISE 3

Questions 10 - 24

Read the following newspaper article. Decide which of the words or phrases marked A, B, C and D best fits into the sentence, and write that letter in the space provided. The first one has been done for you.

A Travel Host's Fear of Flying

Your answers

Yolanta Novak is a co-host of a new TV travel

example:	A. books	C. brochure
	B. series	D. interviews

. She readily — example: ......B......

admits she is the world's worst traveller. Yolanta has a terror of

planes, boats,

10.	A. skiing	C. cars
	B. dogs	D. races

and trains — 10. ......C......

and needs to be sedated before starting a journey. It's all the

11.	A. cause of	C. result in
	B. result of	D. due to

a nightmarish hot air — 11. ......B......

balloon accident

12.	A. who	C. where
	B. when	D. which

could easily — 12. ......D......

have left her crippled.

Yolanta will co-host Channel 7's new

travel program, "Escape", with her husband, Evan Green.

13.	A. Acting	C. Making
	B. Working	D. Travelling

the series, which starts on — 13. ......C......

Tuesday night at 9.30, was an act of great

14.	A. consternation	C. expectation
	B. astonishment	D. determination

and courage — 14. ......D......

on Yolanta's part. She admitted she broke down

15.	A. much	C. several
	B. little	D. plenty

times while on — 15. ......C......

16.	A. assigned	C. assigns
	B. assignments	D. assignable

for "Escape". Evan — 16. ......B......

recalled the horror of

17.	A. have been trapped	C. being trapped
	B. to be trapped	D. had been trapped

17. ......C......

in an out-of-control hot air balloon.

18.	A. "Suddenly	C. "Simultaneously
	B. "Meanwhile	D. "Afterwards

we were caught — 18. ......A......

in a strong wind. I looked at the shadow and knew the balloon

was moving very quickly. We tried to put the balloon

19.	A. on	C. into
	B. under	D. down

just beyond some power lines and — 19. ......D......

clear of trees.

20.	A. Although	C. But
	B. Besides	D. On the other hand

20. ......C......

we crashed into a concrete trough on a nearby stud farm." When

Evan, who was lucky to escape injury, climbed from the wrecked

basket he found

21.	A. her	C. their
	B. a	D. his

wife in agonising pain. — 21. ......D......

Evan said : "Since the

22.	A. hurt	C. accident
	B. time	D. situation

Yolanta has — 22. ......C......

found it difficult to fly and travel in boats and cars. It has been

Your answers

an enormous [23. A. question / B. puzzle / C. problem / D. idea] for her to — 23. ...C...

overcome this fear so that we [24. A. could do / B. should do / C. will do / D. would do] — 24. ...A...

the TV series."

EXERCISE 4

Questions 25 - 32

Read the passage below. It is the start of a mystery story. Fill in the gaps using only one word in each. Write your answers in the column to the right of the passage.

Your answers

My troubles started on the third morning of my holiday. I ..25.. standing outside the hotel waiting for Henry. He was late again. I looked impatiently at my watch. He wasn't ..26.. late, he was very late. In fact, if he was much later, he ..27.. not find me here at all.	25. was 26. slightly/simply/just/merely/only 27. would/might
Suddenly, out of nowhere, a girl rushed up to me. "Michael," she exclaimed, "how lovely to see you again!" My mouth fell open in astonishment. I had certainly never seen ..28.. in my life before.	28. her
"I'm sorry," I stuttered, "I'm afraid . . ."	
"Be quiet," she interrupted, rather rudely, I thought. "Start walking to the corner as ..29.. nothing had happened."	29. if/though
She grabbed my arm, and ..30.. I tried to pull myself away, something cold and hard pressed against my side. I looked down, and saw that she was pointing a gun at me.	30. as/when
My first reaction was to stop dead in my tracks ..31.. a jab in my side got me moving again. "Don't stop," she said. "Just keep on walking ..32.. for goodness sake look natural!"	31. but 32. and/but

I forced my face into a fixed smile, but my legs started to tremble. No one watching attentively would have failed to realize that something was wrong. However, no one even glanced at us.

EXERCISE 5

Questions 33 - 36

Later that day the girl proudly told her colleague what had happened. Read the following passage. For each question, decide which of the phrases best fits the story. (The girl's story should agree with Michael's story on Page 7.)

Write A, B, C or D in the space to the right of the passage.

Your answers

"Well, I chose a moment when he was looking at his watch.

Then I just rushed up to him and exclaimed that

33.	A. that's lovely to see her again.	C. it was lovely to see him again.	33. ....C....
	B. how lovely to see you again.	D. he was lovely to see again.	

Of course he stared at me in astonishment and started to protest but

34.	A. we were interrupted.	C. I interrupted him.	34. ....C....
	B. he interrupted me.	D. I was interrupted.	

When he felt my gun, he stood quite still but I ordered him

35.	A. keeping on walking	C. just keep on walking	, and I told him	35. ....B....
	B. to keep on walking	D. be kept on walking		

to look natural. In fact he looked terrified. Anyone watching carefully

36.	A. would not have failed to realize	C. did not fail to realize	36. ....A....
	B. failed to realize	D. could have failed to realize	

something was wrong."

EXERCISE 6

Questions 37 - 38

37. Put the following sentences in the correct order to tell the story from Henry's point of view. Write the letter of the sentence in the space on the right.

Sentence D is the first one in the correct order so D has been written beside the number 1.

Your answers

A.	He was walking away down the street with a girl.	1.	....D....
B.	But he wasn't waiting.	2.	....C....
C.	Still, I was only 20 minutes late.	3.	....F....
D.	Punctuality is not my strong point.	4.	....E....
E.	After all, he was my best friend.	5.	....B....
F.	Michael would surely wait for me.	6.	....A....

38. Later Michael told his story to a policeman. Put the following sentences and phrases in the correct order.

A.	I'd never ever seen her before	1.	....C....
B.	A complete stranger came up to me	2.	....B....
C.	The most terrifying thing has just happened to me	3.	....E....
D.	and threatening me with a gun	4.	....D....
E.	claiming she knew me	5.	....A....

Appendix 6.4 Students' Questionnaire

1. Name:____
   (Family) (Given)
2. Age:____________

	Male	Female
3. Sex (Circle 1 or 2):	1	2

4. Nationality:____
5. Country of Birth:______________________________________
6. What is your first language (the main language you speak at home)?________________

PREVIOUS EDUCATION

School Education

7. How many years of school education have you had altogether?____________________
8. Which of the following subjects did you study during the final two years of secondary school? Please tick all the appropriate boxes.

Foreign Languages ☐	Biology ☐	Chemistry ☐
Physics ☐	Mathematics ☐	Geography ☐
History ☐	Literature ☐	Economics ☐
Religion ☐	Social Studies ☐	Computer Science ☐

Other Subjects:____

____

University or College Education

9. Please fill in details of all courses attended since leaving school.

Name of University or College	Subjects studied (e.g. Geography, Physics, Accountancy)	Qualifications obtained

PRESENT COURSE OF STUDY

10. What subject(s) are you studying at the moment? (e.g. engineering, nursing, English Language only)________________________
11. Name of Institution:____
12. Level of course (circle appropriate number):

Secondary School	1
Undergraduate	2
Postgraduate	3
Technical and Professional Training	4
English Language Training	5

Other:____________________________

FUTURE COURSE OF STUDY

13. What subject(s) will you study next?
(e.g. engineering, nursing)______________________________
14. Name of Institution (if known):______________________________
15. Level of course (circle appropriate number):

Secondary School	1
Undergraduate	2
Postgraduate	3
Technical and Professional Training	4
Other:______________________	

TESTS

16. If you have taken other tests of English during the last 12 months, please give the scores and the approximate dates at which they were taken.

(I)ELTS______ ______ TOEFL______ ______ Michigan______ ______
(Score) (Date) (Score) (Date) (Score) (Date)

Cambridge First
Certificate______ ______ Other:____________ ______ ______
(Score) (Date) (Name) (Score) (Date)

BACKGROUND KNOWLEDGE

17. Think about the reading you do for your work or during your spare time. Do you read books, magazines, academic papers or newspaper articles on any of the following subjects?
(Please circle 1, 2 or 3 for each subject.)

	Often	Sometimes	Never
Physical Science	1	2	3
Natural (Life) Science	1	2	3
Technology	1	2	3
Medicine	1	2	3
Computer Science	1	2	3
Social Science	1	2	3
Education	1	2	3
Business	1	2	3
Economics	1	2	3
Geography	1	2	3
History	1	2	3
Literature	1	2	3
Languages	1	2	3
<u>Others</u> (Please name subjects)			
____________________	1	2	3
____________________	1	2	3
____________________	1	2	3

READING MODULES A, B AND C

You have just taken two reading tests. Think about these two tests and answer the questions relating to them. (Do not answer the questions about the module you did not take.)

	Module A	Module B	Module C
18. Which module do you think you did best? (Circle the appropriate number.)	1	2	3

Module A (If you did not do Module A, go to Question 34a.)

Reading Passage 1: "Life without a Sunscreen"

	Very Easy					Very Difficult
19. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
20. Were you familiar with this general area of physics before you read the passage?	1	2	3

(If you circled 3, do not answer Question 21.)

21. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
22. Were you familiar with this particular topic, i.e. the effects of the sun's rays?	1	2	3

23. If you were familiar with this topic, did this help you answer the questions?	Yes	No

Reading Passage 2: "Energy from Fuels"

	Very Easy					Very Difficult
24. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
25. Were you familiar with this general area of physics before you read the passage?	1	2	3

26. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
27. Were you familiar with this particular topic, i.e. sources of energy?	1	2	3

28. If you were familiar with this topic, did this help you answer the questions?	Yes	No

Reading Passage 3: "The Recovery of the Mary Rose"

	Very Easy					Very Difficult
29. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
30. Were you familiar with this general area of engineering before you read the passage?	1	2	3

31. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
32. Were you familiar with this particular topic, i.e. the raising of the Mary Rose?	1	2	3

33. If you were familiar with this topic, did this help you answer the questions?	Yes	No

Module B (If you did not do Module B, go to Question 49.)

Reading Passages 1 and 2: "The Mystery of Declining Tooth Decay" and "Our Children's Teeth"

	Very Easy					Very Difficult
34a. Was Reading Passage 1, "The Mystery of Declining Tooth Decay", easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Easy					Very Difficult
34b. Was Reading Passage 2, "Our Children's Teeth", easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
35. Were you familiar with this general area of dental health before you read the passage?	1	2	3

(If you circled 3, do not answer Question 36.)

36. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
37. Were you familiar with this particular topic, i.e. tooth decay in children?	1	2	3

38. If you were familiar with this topic, did this help you answer the questions?	Yes	No

Reading Passage 3: "Three Ways to Make a Transgenic Beast"

	Very Easy					Very Difficult
39. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
40. Were you familiar with this general area of genetics before you read the passage?	1	2	3

41. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
42. Were you familiar with this particular topic, i.e. methods of transferring genes in mice?	1	2	3

43. If you were familiar with this topic, did this help you answer the questions?	Yes	No

Reading Passage 4: "Nitrogen Fixation"

	Very Easy					Very Difficult
44. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
45. Were you familiar with this general area of biology before you read the passage?	1	2	3

46. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
47. Were you familiar with this particular topic, i.e. nitrogen fixation?	1	2	3

48. If you were familiar with this topic, did this help you answer the questions?	Yes	No

Module C

Reading Passage 1: "Quality Circles"

	Very Easy					Very Difficult
49. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
50. Were you familiar with this general area of business studies before you read the passage?	1	2	3

51. If you were familiar with this general area, did this help you answer the questions?	Yes	No

	Very Familiar	Familiar	Not Familiar
52. Were you familiar with this particular topic, i.e. quality circles?	1	2	3
53. If you were familiar with this topic, did this help you answer the questions?	Yes	No	

Reading Passage 2: "The Purposes of Continuing Education"

	Very Easy					Very Difficult
54. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
55. Were you familiar with this general area of education before you read the passage?	1	2	3
56. If you were familiar with this general area, did this help you answer the questions?	Yes	No	

	Very Familiar	Familiar	Not Familiar
57. Were you familiar with this particular topic, i.e. the reasons for continuing education?	1	2	3
58. If you were familiar with this topic, did this help you answer the questions?	Yes	No	

Reading Passage 3: "Access to Higher Education"

	Very Easy					Very Difficult
59. Was this reading passage easy or difficult? (Circle the appropriate number.)	1	2	3	4	5	6

	Very Familiar	Familiar	Not Familiar
60. Were you familiar with this general area of education before you read the passage?	1	2	3
61. If you were familiar with this general area, did this help you answer the questions?	Yes	No	

	Very Familiar	Familiar	Not Familiar
62. Were you familiar with this particular topic, i.e.higher education in Great Britain?	1	2	3
63. If you were familiar with this topic, did this help you answer the questions?	Yes	No	

Appendix 6.5 Distribution of Scores on the Grammar Test and the Reading Modules

BSS Module – All Students

Test Statistics

N of Items	35
N of Examinees	634
Mean	16.874
Variance	43.533
Std. Dev.	6.598
Skew	0.209
Kurtosis	-0.575
Minimum	2.000
Maximum	34.000
Median	16.000
Alpha	0.847
SEM	2.583
Mean Pcnt Cor	48
Mean Item-Tot.	0.400
Mean Biserial	0.521

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	2	1	0	|
3 – 4	9	2	1	| #
5 – 6	20	5	3	| ###
7 – 8	30	10	5	| #####
9 – 10	46	17	7	+ #######
11 – 12	76	29	12	| ############
13 – 14	56	38	9	| #########
15 – 16	89	52	14	| ##############
17 – 18	58	61	9	| #########
19 – 20	59	70	9	+ #########
21 – 22	55	79	9	| #########
23 – 24	43	86	7	| #######
25 – 26	30	90	5	| #####
27 – 28	31	95	5	| #####
29 – 30	16	98	3	+ ###
31 – 32	12	99	2	| ##
33 – 34	2	99	0	|

No examinees above this score

5 10 15

Percentage of Examinees

BSS Module – BSS Students

Test Statistics

N of Items	35
N of Examinees	341
Mean	17.806
Variance	46.948
Std. Dev.	6.852
Skew	0.113
Kurtosis	-0.726
Minimum	3.000
Maximum	34.000
Median	17.000
Alpha	0.860
SEM	2.560
Mean Pcnt Cor	51
Mean Item-Tot.	0.417
Mean Biserial	0.544

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	0	1	0	|
3 – 4	4	1	1	| #
5 – 6	11	4	3	| ###
7 – 8	17	9	5	| #####
9 – 10	21	16	6	+ ######
11 – 12	30	24	9	| #########
13 – 14	26	32	8	| ########
15 – 16	49	46	14	| ##############
17 – 18	29	55	9	| #########
19 – 20	36	65	11	+ ###########
21 – 22	29	74	9	| #########
23 – 24	25	81	7	| #######
25 – 26	17	86	5	| #####
27 – 28	23	93	7	| #######
29 – 30	14	97	4	+ ####
31 – 32	8	99	2	| ##
33 – 34	2	99	1	| #

No examinees above this score

5 10 15

Percentage of Examinees

LMS Module – All Students

Test Statistics

N of Items	40
N of Examinees	513
Mean	18.587
Variance	62.586
Std. Dev.	7.911
Skew	0.291
Kurtosis	-0.523
Minimum	1.000
Maximum	39.000
Median	18.000
Alpha	0.882
SEM	2.715
Mean Pcnt Cor	46
Mean Item-Tot.	0.421
Mean Biserial	0.549

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	2	1	0	
3 – 4	9	2	2	#
5 – 6	17	5	3	###
7 – 8	21	10	4	####
9 – 10	28	15	5	+ #####
11 – 12	51	25	10	##########
13 – 14	47	34	9	#########
15 – 16	52	44	10	##########
17 – 18	44	53	9	#########
19 – 20	42	61	8	+ ########
21 – 22	39	69	8	########
23 – 24	32	75	6	######
25 – 26	41	83	8	########
27 – 28	27	88	5	######
29 – 30	22	92	4	+ ####
31 – 32	11	95	2	##
33 – 34	12	97	2	##
35 – 36	10	99	2	##
37 – 38	3	99	1	#
39 – 40	3	99	1	+ #

5 10 15

Percentage of Examinees

LMS Module – LMS Students

Test Statistics

N of Items	40
N of Examinees	161
Mean	20.702
Variance	77.812
Std. Dev.	9.821
Skew	0.150
Kurtosis	-0.892
Minimum	3.000
Maximum	39.000
Median	20.000
Alpha	0.908
SEM	2.671
Mean Pcnt Cor	52
Mean Item-Tot.	0.466
Mean Biserial	0.605

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	0	1	0	
3 – 4	3	2	2	#
5 – 6	3	4	2	##
7 – 8	5	7	3	###
9 – 10	9	12	6	+ ######
11 – 12	17	23	11	###########
13 – 14	7	27	4	####
15 – 16	14	36	9	#########
17 – 18	13	44	8	########
19 – 20	13	52	8	+ ########
21 – 22	11	59	7	#######
23 – 24	7	63	4	####
25 – 26	13	71	8	########
27 – 28	13	80	8	########
29 – 30	7	84	4	+ ####
31 – 32	6	88	4	####
33 – 34	9	93	6	######
35 – 36	6	97	4	####
37 – 38	2	98	1	#
39 – 40	3	99	2	+ ##

5 10 15

Percentage of Examinees

PST Module – All Students

Test Statistics

N of Items	40
N of Examinees	527
Mean	18.250
Variance	56.245
Std. Dev.	7.500
Skew	0.511
Kurtosis	-0.202
Minimum	2.000
Maximum	40.000
Median	17.000
Alpha	0.882
SEM	2.579
Mean Pcnt Cor	46
Mean Item-Tot.	0.419
Mean Biserial	0.565

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	1	1	0	
3 – 4	2	1	0	
5 – 6	17	4	3	###
7 – 8	20	8	4	####
9 – 10	28	13	5	+ #####
11 – 12	57	24	11	###########
13 – 14	64	36	12	############
15 – 16	51	46	10	##########
17 – 18	58	57	11	###########
19 – 20	46	65	9	+ #########
21 – 22	44	74	8	########
23 – 24	34	80	6	######
25 – 26	27	85	5	#####
27 – 28	21	89	4	####
29 – 30	20	93	4	+ ####
31 – 32	9	95	2	##
33 – 34	12	97	2	##
35 – 36	8	98	2	##
37 – 38	7	99	1	#
39 – 40	1	99	0	+

5 10 15

Percentage of Examinees

PST Module – PST Students

Test Statistics

N of Items	40
N of Examinees	200
Mean	18.490
Variance	51.540
Std. Dev.	7.179
Skew	0.536
Kurtosis	-0.180
Minimum	4.000
Maximum	40.000
Median	18.000
Alpha	0.877
SEM	2.522
Mean Pcnt Cor	46
Mean Item-Tot.	0.409
Mean Biserial	0.559

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	0	1	0	
3 – 4	2	1	1	#
5 – 6	3	3	2	##
7 – 8	3	4	2	##
9 – 10	11	10	6	+ ######
11 – 12	23	21	12	############
13 – 14	32	37	16	################
15 – 16	18	46	9	#########
17 – 18	15	53	8	########
19 – 20	19	63	10	+ ##########
21 – 22	19	73	10	##########
23 – 24	11	78	6	######
25 – 26	15	85	8	########
27 – 28	10	90	5	#####
29 – 30	7	94	4	+ ####
31 – 32	3	95	2	##
33 – 34	4	98	2	##
35 – 36	3	99	2	##
37 – 38	1	99	0	
39 – 40	1	99	0	+

5 10 15

Percentage of Examinees

Grammar

Test Statistics

N of Items	38
N of Examinees	833
Mean	26.242
Variance	38.911
Std. Dev.	6.238
Skew	-0.562
Kurtosis	0.044
Minimum	5.000
Maximum	38.000
Median	27.000
Alpha	0.859
SEM	2.345
Mean Pcnt Cor	69
Mean Item-Tot.	0.399
Mean Biserial	0.576

Score Distribution Table

Score Interval	Frequency	PR	PCT	
1 – 2	0	1	0	|
3 – 4	0	1	0	|
5 – 6	3	1	0	|
7 – 8	3	1	0	|
9 – 10	9	2	1	+ #
11 – 12	6	3	1	| #
13 – 14	14	4	2	| ##
15 – 16	27	7	3	| ###
17 – 18	44	13	5	| #####
19 – 20	40	18	5	+ #####
21 – 22	69	26	8	| ########
23 – 24	87	36	10	| ##########
25 – 26	96	48	12	| ############
27 – 28	106	61	13	| #############
29 – 30	98	72	12	+ ############
31 – 32	95	84	11	| ###########
33 – 34	70	92	8	| ########
35 – 36	50	98	6	| ######
37 – 38	16	99	2	| ##

5 10 15

Percentage of Examinees

Appendix 8.1 Subject Area Familiarity

		BSS Students N = 282	LMS Students N = 60	PST Students N = 85
	Qual Circles	1. 50% 2. 50%	1. 27% 2. 73%	1. 32% 2. 68%
M(BSS)	Education	1. 64% 2. 36%	1. 39% 2. 61%	1. 47% 2. 53%
	Higher Educ	1. 54% 2. 46%	1. 33% 2. 67%	1. 33% 2. 67%
	Whole Test	1. 57% 2. 43%	1. 29% 2. 71%	1. 34% 2. 66%

		N = 160	N = 112	N = 90
	Teeth	1. 29% 2. 71%	1. 59% 2. 41%	1. 31% 2. 69%
M(LMS)	Genes	1. 11% 2. 89%	1. 66% 2. 34%	1. 19% 2. 81%
	Nitro	1. 9% 2. 91%	1. 64% 2. 36%	1. 41% 2. 59%
	Whole Test	1. 9% 2. 91%	1. 68% 2. 32%	1. 19% 2. 81%

		N = 147	N = 60	N = 140
	Sun	1. 33% 2. 67%	1. 77% 2. 23%	1. 69% 2. 31%
M(PST)	Fuel	1. 55% 2. 45%	1. 68% 2. 32%	1. 83% 2. 17%
	Ship	1. 4% 2. 96%	1. 11% 2. 89%	1. 10% 2. 90%
	Whole Test	1. 34% 2. 66%	1. 61% 2. 39%	1. 60% 2. 40%

Key
1 = familiar
2 = not familiar

Appendix 8.2 Topic Familiarity

		BSS Students	LMS Students	PST Students
		N = 290	N = 59	N = 82
	Qual Circles	1. 36% 2. 64%	1. 22% 2. 78%	1. 24% 2. 76%
M(BSS)	Education	1. 57% 2. 43%	1. 43% 2. 57%	1. 44% 2. 56%
	Higher Educ	1. 29% 2. 71%	1. 17% 2. 83%	1. 20% 2. 80%
	Whole Test	1. 32% 2. 68%	1. 21% 2. 79%	1. 20% 2. 80%
		N = 158	N = 109	N = 90
	Teeth	1. 28% 2. 72%	1. 43% 2. 57%	1. 26% 2. 74%
M(LMS)	Genes	1. 6% 2. 94%	1. 37% 2. 63%	1. 10% 2. 90%
	Nitro	1. 6% 2. 94%	1. 45% 2. 55%	1. 36% 2. 64%
	Whole Test	1. 5% 2. 95%	1. 40% 2. 60%	1. 12% 2. 88%
		N = 143	N = 51	N = 140
	Sun	1. 49% 2. 51%	1. 76% 2. 24%	1. 70% 2. 30%
M(PST)	Fuel	1. 71% 2. 29%	1. 75% 2. 25%	1. 78% 2. 22%
	Ship	1. 6% 2. 94%	1. 6% 2. 94%	1. 5% 2. 95%
	Whole Test	1. 44% 2. 56%	1. 57% 2. 43%	1. 56% 2. 44%

Key
1 = familiar
2 = not familiar

Appendix 8.3 Questions to Subject Specialists

Test Method Characteristics

The IELTS Test: Content Validation of the Reading Passages

3 February 1992

IELTS is a test of English for Academic Purposes which is taken by overseas students who need to show that their English is adequate for study at British universities. There are three versions of the test: one is for students in the social sciences, one for students in the physical sciences and technology, and one for students in the life sciences and medicine.

The purpose of these questions is to check whether the reading passages in the three tests are appropriate for students in the relevant subject areas.

1. Appropriacy of Text

Is the text the sort of text your undergraduate and postgraduate students have to read?

Highly Appropriate		Not at All Appropriate
0	1	2

2. Familiarity with Topic

Would students in your department be familiar with this particular topic?

Very Familiar		Not at All Familiar
0	1	2

3. Specialisation of Topic

Is the topic of this passage highly specialised?

Not at All Specialised		Highly Specialised
0	1	2

4(a) New Knowledge

To what extent does the passage introduce knowledge (including new vocabulary) that you would not expect your students to know?

No New Knowledge		A Lot of New Knowledge
0	1	2

4(b) Explanation of New Knowledge

If there is new knowledge in the passage, to what extent is it explained (through definition, example, paraphrase, etc.)?

Fully Explained		Not at All Explained
0	1	2

5. Academic Specificity

Is this passage specifically aimed at academic readers?

Not at All Specific		Highly Specific
0	1	2

Appendix 8.4 Subject Specialists – Familiarity with Topic

		BSS	LMS	PST
Qual	1.	100%	14%	40%
	2.	-	86%	60%
Educ.	1.	71%	57%	40%
	2.	29%	43%	60%
Higher	1.	.29%	14%	40%
	2.	71%	86%	60%
Teeth	1.	43%	86%	80%
	2.	57%	14%	20%
Genes	1.	-	86%	80%
	2.	100%	14%	80%
Nitro	1.	-	100%	60%
	2.	100%	-	40%
Sun	1.	29%	86%	100%
	2.	71%	14%	-
Fuel	1.	86%	86%	80%
	2.	14%	14%	20%
Ship	1.	29%	29%	60%
	2.	71%	71%	40%

Key

1 = Familiar
2 = Not Familiar

Appendix 9.1 Test Method Characteristics Instrument

Notes to Accompany the TMC Rating Sheets (20 January 1992)

Adapted from Bachman's Test Method Rating Instrument (4 March 1991)

[Note: References to pages in Bachman, 1990, relate to L. F. Bachman, *Fundamental Considerations in Language Testing*, Oxford University Press, 1990.]

In general, all scales are intended to be directional, so that a larger value would be expected to make an item more difficult. However, it should be remembered that these are not direct ratings of difficulty, but ratings of the content of the test itself in terms of the specific facets. That is, the ratings on these facets will enable us to look at how these characteristics of test tasks and content relate to difficulty, but they are not in themselves ratings of difficulty.

The level at which the ratings are to be given (passage, item) are indicated beside each facet. Ratings for items should only be applied to the stimulus – not to the answer (except in the case of multiple-choice items, where ratings should relate to the stimulus together with the possible answers).

Task

Rubric (Items only)

This facet relates to the instructions to test takers about how they should proceed in each part of the test. Ratings should on the following scale:

RUBRIC 0 = clear for unprepared test takers

1 = possibly unclear for unprepared test takers

2 = unclear for unprepared test takers

Item Type (Items only)

The following items illustrate different items types that may differ in their familiarity to test takers. Only individual items should be rated on this facet, and ratings should be made with the "unprepared" test taker in mind.

1. Amish settlers who arrived in Iowa on 1825 constructed the sod house in America.

   A B C D

2. As you see / address above, / we move house.

3. He did his best to ________ that the instructions for rating were clear.

   a. manage
   b. ensure
   c. convince
   d. produce

	Very familiar		Very unfamiliar
ITEM TYPE	0	1	2

Propositional Content

Vocabulary (Passages and items) (Bachman 1990, page 131)

Answer this in relation to the specific group of test takers for whom the test is intended. In the case of IELTS the test takers are ESL students who are attending or are hoping to attend undergraduate or postgraduate courses at English medium universities.

NB: These facets apply not only to words but also to fixed phrases and idiomatic expressions that may be relatively infrequent, specialised or ambiguous.

INFREQUENT	(Frequent) 0	1	2 (Infrequent)
SPECIALISED (e.g., technical, jargon, slang)	(General) 0	1	2 (Specialised)
AMBIGUOUS	(Clear) 0	1	2 (Ambiguous)

(Ambiguity refers to the possibility of more than one reading, or interpretation, of a phrase, sentence or text. For multiple-choice items, this could arise if the keyed response results in an ambiguous sentence, or if there is more than one possible answer.)

Degree of Contextualisation (Passages) (Bachman 1990, page 131)

In rating this facet, consider the relative proportion of "new" to "contextual" information. "New information" (which includes new vocabulary) is that which is not known to the test taker and cannot be predicted from the context. "Contextual information" is that which is developed in the passage itself. Thus, a passage is "not at all contextualised" if there is a lot of new information in the passage that is not explained through definition, example, paraphrase, etc. The passage is "highly contextualised" if there is no new information, or if the new information is explained. If the reader has prior knowledge that will help comprehension, then the text is contextualised. If the reader does not have relevant prior knowledge, the discourse is context reduced.

Input can be contextualised in terms of two types of information: cultural and that which is topic specific. Cultural content relates to national (general) culture such as national habits, customs and beliefs. Ratings on this facet should be as follows:

	Highly contextualised		Not at all contextualised
With respect to CULTURAL CONTENT	0	1	2
With respect to TOPIC SPECIFICITY	0	1	2

[Give separate ratings for BSS (Business and Social Science), LMS (Life and Medical Science), and PST (Physical Science and Technology) students.]

Example of rating:

0 = No new information, or new information is explained in text
1 = Little new information, not explained in the text
2 = New information that may be unfamiliar and is not explained in text

<u>Relationship of Item to Passage</u> (Items only)

This facet should be rated in terms of: 1) the extent of text to which the item relates, and 2) whether the item requires the test taker to relate information in the passage to the real world (i.e. to the test taker's knowledge schemata). For purposes of this rating, "specific part" means "one sentence or several contiguous sentences". Items should be rated on the following scale:

0 = No relationship to the passage; items can be answered without reference to the passage, OR relationship of item to passage is not clear.

1 = Item relates to one specific part of the passage, and requires only localised understanding of that part. If this is the case, write "1" even if it is also possible to reach the answer by referring to more than one part of the passage.

2 = Item relates to more than one specific part of the passage, or requires the test taker to relate one specific part to one or more others.

3 = Item relates to the entire passage, and requires an understanding of the entire passage.

4 = Item relates to one specific part of the passage, requires only localised understanding of that part, and requires the test taker to relate information in that part to the real world.

5 = Item relates to more than one specific part of the passage, or requires the test taker to relate one specific part to one or more others, and requires the test taker to relate the information in those parts to the real world.

6 = Item relates to the entire passage, requires an understanding of the entire passage, and requires the test taker to relate information in the passage to the real world.

<u>Topic</u> (Passages only) (Bachman 1990, page 137)

This facet has to do with the topic, or "subject", of the text, and not whether the test taker is British, or an academic, or in a specialised area. Thus, for example, a text that has a great deal of specific American, Australian, British or Canadian cultural content is highly specific to this category, and would be rated "2", irrespective of whether a given test taker is of that background or orientation.

Note, therefore, that for this facet the test taker should <u>not</u> be taken into account.

	Not at all Specific		Highly Specific
CULTURE SPECIFIC	0	1	2
ACADEMIC SPECIFIC	0	1	2
SPECIALISED TOPIC	0	1	2

Organisational Characteristics (Bachman 1990, page 139)

Grammar (Passages and items)

This relates to the complexity of sentence types and embeddings, and the frequency of the passive voice.

	Very simple		Very complex
GRAMMAR	0	1	2

Cohesion (Passages only)

This relates to the use of cohesive devices such as Reference, Substitution, Adversatives, Causals, Temporals and Lexical Cohesion as in M. A. K. Halliday and Ruqaiya Hasan, *Cohesion in English*, Longman, 1976.

	Not at all complex		Highly complex
COMPEXITY OF COHESIVE DEVICES	0	1	2

Rhetorical Complexity (Passages only)

This facet should be rated in terms of how complex the rhetorical organisation is, not on how familiar test takers are with it. RHETORICAL ORGANISATION should be rated according to how complex the text is in the classical rhetorical sense, e.g., instruction < description < comparison & contrast < argumentation. In general, instruction and description should be rated 0, comparison and contrast 1, and argumentation 2.

	Very simple		Very complex
RHETORICAL ORGANISATION	0	1	2

	One	Two	Three or more
NUMBER OF SPECIFIC TYPES OF RHETORICAL ORGANISATION	0	1	2

Sociolinguistic Characteristics

(The following explanation comes from Bachman 1990, page 97.)

Many cultural references and figures of speech will be incorporated, with set meanings, into the lexicon of any language. Nevertheless, knowledge of the extended meanings given by a specific culture to particular events, places, institutions, or people is required whenever these meanings are referred to in language use. For example, to interpret the following exchange, the language user would have to know that 'Waterloo' is used linguistically to symbolise a major and final defeat with awful consequences for the defeated:

> A: I hear John didn't do too well on his final exam.
> B: Yeah, it turned out to be his Waterloo.

Similarly, interpreting figurative language involves more than simply knowledge of referential meaning. For example, the correct interpretation of hyperboles such as "I can think of a million good reasons for not smoking" and clichés like "It's a jungle out there" require more than a knowledge of the signification of the words and grammatical structures involved.

	No occurrences	One occurrences	Two or more occurrences
CULTURAL REFERENCES (Passages only)	0	1	2
FIGURATIVE LANGAUAGE (Clichés, metaphors, etc.) (Passages only)	0	1	2

Appendix 10.1 Distribution of the Scores on the Revised LMS and PST Modules

LMS Module – Items 28–40

Test Statistics

N of Items	13
N of Examinees	513
Mean	4.536
Variance	10.592
Std. Dev.	3.255
Skew	0.306
Kurtosis	-0.866
Minimum	0.000
Maximum	13.000
Median	4.000
Alpha	0.809
SEM	1.423
Mean Pcnt Cor	35
Mean Item-Tot.	0.542
Mean Biserial	0.716

Score Distribution Table

```
Number   Fre-
Correct  quency PR  PCT
  0        68   13   13  + #############
  1        47   22    9  | #########
  2        52   33   10  | ##########
  3        51   42   10  | ##########
  4        50   52   10  | ##########
  5        47   61    9  + #########
  6        45   70    9  | #########
  7        41   78    8  | ########
  8        44   87    9  | #########
  9        28   92    5  | #####
 10        22   96    4  + ####
 11        10   98    2  | ##
 12         7   99    1  | #
 13         1   99    0  |
                         | ----+----+----+
                             5   10   15
                         Percentage of Examinees
```

PST Module – Items 1–11

Test Statistics

N of Items	11
N of Examinees	527
Mean	6.750
Variance	7.554
Std. Dev.	2.748
Skew	-0.473
Kurtosis	-0.678
Minimum	0.000
Maximum	11.000
Median	7.000
Alpha	0.771
SEM	1.316
Mean Pcnt Cor	61
Mean Item-Tot.	0.552
Mean Biserial	0.735

Score Distribution Table

```
Number   Fre-
Correct  quency PR  PCT
  0         3    1   13  + #############
  1        22    5    9  | #########
  2        26   10   10  | ##########
  3        28   15   10  | ##########
  4        40   23   10  | ##########
  5        44   31    9  + #########
  6        54   41    9  | #########
  7        73   55    8  | ########
  8        64   67    9  | #########
  9        86   83    5  | #####
 10        59   95    4  + ####
 11        28   99    2  | ##
                         | ----+----+----+
                             5   10   15
                         Percentage of Examinees
```

Subject Index

Author Index

K

L

M