

IDENTIFYING THE LANGUAGE PROBLEMS OF OVERSEAS STUDENTS
IN TERTIARY EDUCATION IN THE UNITED KINGDOM

CYRIL JAMES WEIR

Volume One of a thesis submitted for the
Ph.D. Degree of the University of London

Department of English for Speakers of Other Languages, Institute
of Education.

1983

ACKNOWLEDGEMENTS

A lot of people helped to make this project feasible. It would be impossible to mention them all by name, but I should like to thank all the students who filled out questionnaires and took the tests, plus the members of staff who co-operated in these exercises. Together they contributed substantially to the material success of the test by providing data on which to base the test, test results and external validation criteria. I am grateful to all who commented upon the tests and to the staff of the Associated Examining Board, especially those in the Research Unit, without whose support the project would not have been possible.

In particular I should like to thank Mr. C.P. Hill and Dr. J.G. Houston for their valued and generous supervision and support; Mr. J. Wilmut and Mr. K. Trinder for carefully taking a novice through the statistics; the other members of the Working Party: Dr. J.C. Alderson, Mr. A.M. Emmett, Mr. G.C.R. James, Mr. K. Morrow, Mr. G. Blue, Mr. D. Porter, Mr. H. Thomas, for their sound advice; Ms. D. Wall, Ms. S. Davis, Mr. T. Whelpton for marking the scripts; Ms. L. Byrne and Ms. J.M. Dray for typing the thesis; Ms. K. Rogers for her unstinting clerical assistance.

All of these people have contributed to whatever success this research has had; its faults are all my own.

ABSTRACT

The research and development leading to the pre-test version of the Associated Examining Board's Test in English for Academic Purposes (T.E.A.P.) is described and discussed in this work. The aim of this test is to provide, by means of individual profiles of competence in reading comprehension, listening comprehension and writing, information on overseas students' understanding and use of written and spoken English in academic situations. This will help institutions to make decisions on acceptance or rejection and indicate whether or not remedial language tuition might be necessary.

In the first part of this enquiry, the study levels, subject discipline areas and institutions where overseas students enrol in the further and higher education sectors in Britain are established and the problems overseas students encounter in these academic contexts are examined.

Next the basic concepts of test construction are discussed and the relative merits of various approaches to language testing are compared in the light of these; particular attention being paid to the a priori validation of test tasks.

The thesis details a methodological framework for establishing the specification of language tasks facing students at a variety of levels and in a variety of subject areas, together with the results of the investigations that were conducted. Two principal methods of data collection were employed in completing the specification; first by a series of observational visits to a variety of educational institutions and second by the circulation of a questionnaire to staff and students in a similar variety of educational institutions. A report is made on the extent of difficulty both overseas and British students encounter in coping with the tasks and attendant performance constraints in the academic contexts under review.

The realisation of this specification in a test battery designed to assess students' ability to perform language tasks relevant to the

academic context in which they have to operate is then described. This battery incorporated a variety of test formats of varying degrees of directness of fit to the specification outlined earlier. In this way it was hoped to establish the relative merits of various methods of assessing a student's performance on those tasks that the research indicated as important to overseas students following English-medium courses.

The pre-test battery is tested on a sample of students drawn from the population of overseas students for whom the test is intended and on a control group of English native speakers. The results are used in the internal and external validation of the test. The way in which various validation procedures helped determine the final format of the Test in English for Academic Purposes (T.E.A.P.) is examined.

The implications of the results are discussed and possible areas of further research are suggested at the end of the thesis.

V O L U M E O N E C O N T E N T S

CHAPTER ONE

	<u>Page</u>
1. THE BACKGROUND: THE PROBLEM TO BE STUDIED AND THE PURPOSE OF THE RESEARCH	4
1.1 RESEARCH PURPOSE: The Need for a Suitable Test	4
1.2 OVERSEAS STUDENT NUMBERS: Defining the Population for the Test	7
1.2.1 The Overseas Student	7
1.2.2 The Number of Overseas Students Enrolled on Courses in Tertiary Education	8
1.2.3 The Number of Overseas Students Accepted Annually for Courses in Tertiary Education	11
1.2.4 Survey of Individual Institutions in Tertiary Education	20
1.2.5 Future Trends	22
1.2.6 Conclusions	24
1.3 THE PROBLEM FOR STUDENTS AND INSTITUTIONS	25
1.3.1 The Nature and Level of Proficiency in English Required in an Academic Context	25
1.3.1.1 Present Language Entry Requirements of Institutions in the Tertiary Sector	27
1.3.1.2 How much is Enough? Selection and Cut-off Points	31
1.3.1.3 The Native Speaker Norm	32
1.3.2 The Need for Early Identification of a Shortfall in Language Proficiency	34
1.3.3 A Brief Survey of the Literature on the Problems Overseas Students Encounter in English Medium Study	37
1.3.3.1 Is it Just a Language Problem?	37
1.3.3.2 Establishing a Framework for Enquiry: a brief review of the literature concerning the problems encountered in the various study modes students have to operate in at the tertiary level	41
1.3.3.2.1 Lectures	42
1.3.3.2.2 Seminars	48
1.3.3.2.3 Practicals	50
1.3.3.2.4 Reading	50
1.3.3.2.5 Essays, report writing, dissertation	52

	<u>Page</u>
1.3.3.3 The Need for Remedial English - S.E.L.M.O.U.S.	53
1.3.3.4 Is Remedial Assistance Enough?	54
1.3.3.5 How Important is Language Proficiency for Academic Progress?	54
1.3.4 Conclusions	58

CHAPTER TWO

2. PRINCIPLES AND APPROACHES IN LANGUAGE TESTING	63
2.1 BASIC CONSIDERATIONS IN TEST DESIGN	63
2.1.1 The Concept of Validity	63
2.1.1.1 Construct Validity	64
2.1.1.2 Content Validity	70
2.1.1.3 Face Validity	72
2.1.1.4 Criterion-Related Validity	73
2.1.2 The Concept of Reliability	75
2.1.3 Test Efficiency	76
2.1.4 Validity and Reliability - An Inevitable Tension?	77
2.2 A CRITICAL REVIEW OF APPROACHES TO LANGUAGE TEST DESIGN	80
2.2.1 Introduction	80
2.2.2 The Psychometric-Structuralist Era	81
2.2.3 The Psycholinguistic-Sociolinguistic Era	83
2.2.4 The Communicative Paradigm	88
2.2.5 The Problem of Extrapolation	92

CHAPTER THREE

3. IN PURSUIT OF A NEW PARADIGM	105
3.1 TOWARDS A FRAMEWORK FOR DESCRIPTION OF COMMUNICATIVE TASKS	105

	<u>Page</u>
3.1.1 The English for Specific Purposes (E.S.P.) Approach	105
3.1.2 Establishing the General Descriptive Parameters of Communication	114
3.1.3 Establishing the Dynamic Communicative Characteristics	116
3.1.4 Establishing the Test Task Dimensions	120
3.2 METHODS OF DATA COLLECTION: Filling Out the Framework	122
3.2.1 The Observations	122
3.2.1.1 Introduction	122
3.2.1.2 The Observation Method of Data Collection	125
3.2.1.2.1 Events and activities	127
3.2.1.2.2 Setting	130
3.2.1.2.3 Interactions	131
3.2.1.2.4 Instrumentality	131
3.2.1.2.5 Dialect	132
3.2.1.2.6 Target level dimensions	132
3.2.1.2.7 Target level tolerance conditions	136
3.2.1.2.8 Communicative key	137
3.2.1.2.9 Enabling skills	138
3.2.1.2.10 Conclusions	140
3.2.2 The Follow Up Interviews	140
3.2.3 The Questionnaire Method of Data Collection	142
3.3 THE PURPOSE OF LECTURES, SEMINARS/TUTORIALS AND PRACTICAL CLASSES	145
3.3.1 The Lecture	145
3.3.2 The Seminar and the Tutorial	148
3.3.3 The Practical Class	151
3.4 ACTIVITIES IN THE ACADEMIC CONTEXT	153
3.4.1 A Cautionary Note	153
3.4.2 Listening Comprehension Activities	156
3.4.2.1 The Nature of Listening Comprehension	156
3.4.2.2 The Listening Comprehension Tasks Students Perform in the Academic Context	160
3.4.2.2.1 Informing transactions	160
3.4.2.2.2 Directing transactions	164
3.4.2.2.3 Eliciting transactions	168
3.4.2.2.4 Formality	172
3.4.2.3 Problems in Listening Comprehension	176
3.4.2.3.1 General problems in listening comprehension	176
3.4.2.3.2 Specific problems in listening comprehension	180
3.4.2.4 Note Taking Activities	185
3.4.2.4.1 Methods of acquiring a written record	185

	<u>Page</u>
3.4.2.5 Problems in Note Taking	188
3.4.2.6 Constituent Enabling Skills	192
3.4.3 Reading Comprehension Activities	194
3.4.3.1 The Nature of Reading Comprehension	194
3.4.3.2 The Reading Tasks Students Perform in the Academic Context	198
3.4.3.3 Reading Difficulties Encountered by Students	207
3.4.3.4 Importance of Note Making	213
3.4.3.5 The Frequency with which Students Make Notes from Text-books	218
3.4.3.6 The Extent of Difficulty Encountered in Note-Making	220
3.4.3.7 The Specific Problems Encountered in Note-Making	220
3.4.3.8 Constituent Enabling Skills	221
3.4.4 Writing Activities in the Academic Context	226
3.4.4.1 The Nature of Written Production	226
3.4.4.2 The Writing Tasks Students Perform in the Academic Context	233
3.4.4.3 Problems Experienced in Written Production	242
3.4.4.4 The Standards Expected of British and Overseas Students	259
3.4.4.4.1 Variations in the allowances staff claim they make in marking the written work of overseas students	261
3.4.4.4.2 Variations in the allowances staff claim they make in assessing the written work of overseas students, as between coursework and examinations	267
3.4.4.5 Constituent Enabling Skills	269
3.4.5 Speaking Activities in the Academic Context	272
3.4.5.1 Introduction	272
3.4.5.2 The Speaking Tasks Students Perform in the Academic Context	272
3.4.5.2.1 Informing transactions	273
3.4.5.2.2 Students working together using English as a means of communication	277
3.4.5.2.3 Students actively taking part in discussions involving the teacher and the class as a whole	280
3.4.5.2.4 Eliciting transactions	281
3.4.5.2.5 The frequency with which the teacher asks the class questions	283
3.4.5.3 Problems Experienced in Spoken English	285
3.4.5.4 Constituent Enabling Skills	303
3.5 SETTING	305
3.6 INTERACTION	307
3.7 INSTRUMENTALITY	309
3.8 CONCLUSION	311

CHAPTER FOUR

	<u>Page</u>
4. TEST DESIGN AND CONSTRUCTION	315
4.1 TYPE OF TEST BATTERY NEEDED	315
4.2 FROM SPECIFICATION TO REALISATION	318
4.2.1 The Power of the Data: A Cautionary Note	318
4.2.2 Handling the Data	319
4.2.3 General Descriptive Parameters	322
4.2.4 Dynamic Communicative Characteristics	324
4.2.5 Task Dimensions	325
4.2.5.1 Size	325
4.2.5.2 Complexity	326
4.2.5.3 Functional/Referential Range	326
4.2.6 A Further Problem in Text Selection	327
4.2.7 The Problem of Format	330
4.3 ASSESSING READING SKILLS IN AN E.A.P. CONTEXT	332
4.3.1 General Descriptive Parameters	332
4.3.2 Formats for Testing Reading Comprehension	333
4.3.2.1 Multiple Choice Techniques	333
4.3.2.2 Short Answer Questions	338
4.3.2.3 Cloze and Gap-Filling	340
4.3.3 Testing the Range of Skills	346
4.3.4 Variety of Format	346
4.3.5 Conclusions	347
4.4 ASSESSING LISTENING SKILLS IN AN E.A.P. CONTEXT	349
4.4.1 General Descriptive Parameters	349
4.4.2 A New Paradigm	351
4.4.3 Formats for Testing Listening Comprehension	353
4.4.3.1 Job Sample Tests: Monologue and Interactive Discourse	353
4.4.3.2 Dictation and Listening Recall	360

	<u>Page</u>
4.4.4 Conclusions	366
4.5 ASSESSING WRITING SKILLS IN AN E.A.P. CONTEXT	367
4.5.1 Why Test Writing Skills?	367
4.5.2 Formats for Testing Written Production in an E.A.P. Context	369
4.5.2.1 Objective Formats	371
4.5.2.2 Integrative Tasks	374
4.5.3 The Assessment of Written Production: A Problem of Reliability	378
4.5.3.1 Impression Versus Analytical Approaches	378
4.5.3.2 Establishing Appropriate Criteria for Assessing Written Production in T.E.A.P.	390
4.5.3.3 Further Considerations in Designing Writing Tasks for Inclusion in a Test Battery	395
4.5.3.3.1 Number of writing tasks	395
4.5.3.3.2 Question choice	396
4.5.3.3.3 Amount of time allowed for each writing task: The ramifications of time limits	397
4.5.3.3.4 Checking the viability of the writing tasks	399
4.6 CONCLUSIONS	401

CHAPTER FIVE

5. T.E.A.P. PRE-TEST: PROCEDURES, REACTIONS AND ANALYSIS	406
5.1 ADMINISTRATIVE PROCEDURES AND REACTIONS	406
5.1.1 Introduction	406
5.1.2 Sampling	410
5.1.2.1 Selection of Sample	410
5.1.2.2 Obtaining the Sample	410
5.1.2.3 Breakdown of Sample	410
5.1.3 Preparation of Materials and Administration	413
5.1.4 Associated Examining Board's Examination Procedures	414
5.1.4.1 Marking Schemes	414
5.1.4.2 The Moderation of Question Papers and Marking Schemes	415
5.1.4.3 The Standardisation of Marking	417

	<u>Page</u>
5.1.5 Validity Criteria - Administration	418
5.1.6 Reactions to T.E.A.P.	420
5.1.6.1 General Issues Arising	420
5.1.6.2 Tutor and Candidate Responses to Specific Test Tasks	433
5.1.6.2.1 TO	433
5.1.6.2.2 TA	437
5.1.6.2.3 TB	440
5.1.6.3 Conclusions	443
5.1.7 Handling the Data	445
5.2 DATA ANALYSIS AT THE ITEM LEVEL	446
5.2.1 Introduction	446
5.2.2 Reading	448
5.2.2.1 Conclusions	452
5.2.3 Listening	453
5.2.4 Linguistic Competence	455
5.2.5 Writing	456
5.2.5.1 Introduction	456
5.2.5.2 Summary Tasks: T011, T032, TA31 and TB31	459
5.2.5.3 Editing Tasks: TA32 and TB32	465
5.2.5.4 Inter- and Intra-Marker Reliability	467
5.2.5.5 Conclusions	473
5.3 DATA ANALYSIS AT THE TASK LEVEL (INTERNAL)	474
5.3.1 Introduction	474
5.3.2 Approaching the Test Data	477
5.3.3 Relationship Between Tasks	481
5.3.3.1 Correlational Data	481
5.3.3.1.1 Reading	484
5.3.3.1.2 Writing	485
5.3.3.1.3 Listening	485
5.3.3.1.4 Linguistic competence	486
5.3.4 Factor Analysis Data	486
5.3.4.1 Conclusions	494

	<u>Page</u>
5.4 DATA ANALYSIS AT THE TASK LEVEL (EXTERNAL)	495
5.4.1 Introduction	495
5.4.2 Tutors' Estimates (Global and Communicative) of Students' Language Proficiency	495
5.4.2.1 Background	495
5.4.2.2 Reading	499
5.4.2.3 Listening	499
5.4.2.4 Writing	499
5.4.3 Students' Self-Assessments of Language Proficiency	500
5.4.4 Students' and Tutors' Estimates of Language Proficiency Compared with Aggregated Task Scores	502
5.4.5 Relationship Within and Between Tutors' and Students' Assessments of Language Proficiency	504
5.5 COMPILING THE FINAL FORM OF T.E.A.P.	506
5.5.1 Introduction	506
5.5.2 Decisions on Which Tasks to Include	506
5.5.2.1 Reading	506
5.5.2.2 Listening	507
5.5.2.3 Writing	507
5.5.3 Data Analysis at the Construct Level	508
5.5.4 Reliability of the Constructs	512
5.5.5 The Concurrent and Predictive Validity of the Constructs	515
5.5.6 Decision on Which Constructs to Include	520
5.5.7 Conclusions on the Validity and Reliability of the Constructs	521
5.5.8 Analysis of Performance in the Two Versions of the Test	522
5.5.9 External Validation of Sessions and Combinations of Sessions	528

	<u>Page</u>
5.6 INFLUENCE OF BACKGROUND VARIABLES ON ENGLISH PROFICIENCY	531
5.7 SETTING GRADE BOUNDARIES	538
5.7.1 Introduction	538
5.7.2 Judgements about Groups of Candidates	539
5.7.3 Judgements of Test Content	542
5.7.4 Maintaining Grade Boundaries	543
5.7.5 Future Grade Awarding	544

CHAPTER SIX

6. CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES	546
--	-----

<u>BIBLIOGRAPHY</u>	555
---------------------	-----

V O L U M E T W O C O N T E N T S

CHAPTER ONE APPENDICES

	<u>Page</u>
1.1 Letters and Pro Forma sent to Institutions Requesting Details of Overseas Student Numbers	596
1.2 Totals of New Non-Native Speaker (N.N.S.) Entries In the Further and Higher Education Sectors 1978-1979	612
1.2.1 Totals of New Entries Undergraduate and Post-graduate According to University and Subject as at 31st December, 1978	613
1.2.2 Totals of New First Year Students by Subject and Polytechnic as at 31st December, 1978	626
1.2.3 Totals of New First Year Students in the Further Education Sector by Subject and College as at 31st December, 1978	628
1.3 English Language Requirements for Entry into Institutions in the Further and Higher Education Sectors	633

CHAPTER THREE APPENDICES

3.1 Questionnaire on Communicative Characteristics sent to A.R.E.L.S. Schools	667
3.2 The Observations Conducted at Various Institutions in Further and Higher Education	670
3.2.1 Specimen Copy of the Observation Checklist	671
3.2.2 Details of the Observations Conducted 1979-1980	689
3.2.3 Attitudinal Tones most Frequently Encountered by Students in the Various Study Modes	694
3.3 The Questionnaire	696
3.3.1 Specimen Copies of the Pilot and Final Versions of the Questionnaire	697
3.3.2 Details of the Number of Returns to the Pilot and Final Versions of the Questionnaire	788

	<u>Page</u>
3.4 Overall Percentages of the Returns to Questionnaires, Concerning the Difficulties Encountered by Students in Listening, Reading, Writing and Speaking, Ranked in Order of Magnitude	802
3.5 Breakdown of the Amount of Difficulty Overseas Students Claimed to have Experienced in each Activity and the Frequency of Occurrence of that Activity, across Levels and Discipline Areas	821

CHAPTER FOUR APPENDICES

4.1 Answer Booklets and Source Booklets Used in the Pre-Test Administration of T.E.A.P. September-November 1982	827
4.1.1 Answer Booklet and Source Booklet for Session I	828
4.1.2 Answer Booklet and Source Booklet for Session IIA	867
4.1.3 Answer Booklet and Source Booklet for Session IIB	902
4.2 Transcripts of Tapes Used for TO31, TA21 and TB21 and Brief Descriptions of the Speakers on the Tapes	931
4.3 Enabling Skills on Which Individual Items Focus	946
4.3.1 Reading Skills	947
4.3.2 Listening Skills	949
4.4 Listening Recall Tests Used in Preliminary Trials at the Universities of Reading and Edinburgh	951

CHAPTER FIVE APPENDICES

5.1 Details of Participating Centres	955
5.2 Background Details	957
5.2.1 Background Details Questionnaire	958
5.2.2 Summary Tables of the Responses to the Questions in the Background Details Questionnaire	961

	<u>Page</u>
5.3 Instructions to Invigilators	982
5.4 T.E.A.P. Pre-test Mark Schemes	990
5.4.1 Session I (T0)	991
5.4.2 Session IIA (TA)	1006
5.4.3 Session IIB (TB)	1023
5.4.4 Sheets Used by Examiners for Recording Marks for the Various Tasks	1039
5.5 External Validity Pro Forma	1045
5.5.1 Students' Self-Assessment	1046
5.5.2 Teachers' Communicative Assessment (T.C.)	1049
5.5.3 Teachers' Global Assessment (T.G.)	1052
5.5.4 Subject Tutors' (Predictive Validity) Assessment	1054
5.6 Reactions to T.E.A.P.	1056
5.6.1 Specimen Copy of Follow Up Questionnaires	1057
5.6.2 Summary of Returns to the Follow Up Questionnaires	1066
5.7 T.E.A.P. Statement of Aims and Objectives	1072
5.8 Analysis at the Item Level	1076
5.9 Second Markings and Original Marks for each Writing Criterion	1096
5.10 Means and Standard Deviations of N.N.S. Test Population Sub-Groups According to Level and Subject Area	1102
5.11 Inter-task Correlations	1115
5.11.1 All N.N.S. and Various Task Amalgamations	1116
5.11.2 Various N.N.S. taking Combinations of Session	1126

	<u>Page</u>
5.12 Factor Analysis: Printouts of Factor Analyses and Varimax Rotations	1136
5.12.1 Set One Sci/Eng and A.S.A.B.S. Groups of Candidates taking Various Combinations of Session	1137
5.12.2 Set Two All N.N.S. taking Various Combinations of Session	1156
5.12.3 Set Three All N.N.S. taking Single Sessions	1166
5.13 Contingent Numbers for External Validity Studies	1173
5.13.1 Contingent Numbers for Correlations between Task Scores and Students' and Teachers' Assessments of Reading, Listening and Writing Abilities	1174
5.13.2 Contingent Numbers for Correlations between Skill Composite Scores and Teachers' Assessments of Students' Reading, Listening and Writing Abilities	1180
5.13.3 Contingent Numbers for Correlations between Total Test Scores, Total Teachers' Assessments and Subject Tutors' Assessments (Language and Academic Progress) at End of Term Two	1183
5.14 Influence of Background Variables	1185

LIST OF TABLES

CHAPTER ONE

<u>Table</u>		<u>Page</u>
1A	Number of Overseas Students Enrolled in Universities, Polytechnics and Further Education Colleges	9
1B	Total Number of Overseas Students Enrolled on Courses in the Three Largest Discipline Areas	10
1C	Number of Overseas Students Accepted for Courses in Universities 1976-1981	13
1D	Number of Overseas Students, Full-Time and Sandwich, Accepted for Courses in the Other (Non-University) Higher and Further Education Sectors in Tertiary Education 1977-1978.	14
1E	Subject Totals of New Overseas Entrants to the University Sector at Undergraduate and Post-graduate Level in the Three most Populated Disciplines	16
1F	Number of Overseas Candidates Applying and Accepted for University Undergraduate Places in Areas 3, 5 and 6 in 1977-1978 and 1980-1981	18
1G	G.C.E. Advanced Level Subject Entries Made by Overseas Candidates from Grant-Aided Further Education Establishments	19
1H	Totals Obtained from October 1979 Survey on Numbers of New Overseas Students, for whom English is not the Mother Tongue in the Country of Origin, Enrolling for the First Time in the 1978-79 Session	21

CHAPTER THREE

3A	Framework of Categories for the Description of Communicative Test Events	114
3B	Categories Used in Observation Checklist	126
3C	Main Features of the Classification Used in the S.T.O.S.	128
3D	Specification of Communicative Needs	134
3E	The Frequency of Occurrence in Observations of Teacher Making Statements	162
3F1	Summary of the Returns to the Questionnaire Concerning the Frequency of Teacher Giving Instructions to Look at Various Sources of Information	165

<u>Table</u>	<u>Page</u>
3F2 The Frequency of Occurrence in Observations of Teacher Directing Pupils to Sources of Information	167
3G1 Summary of the Returns to the Questionnaire Concerning the Frequency With Which Staff Ask Students Questions	169
3G2 The Frequency of Occurrence in the Observations of Teacher Asking Questions (or Inviting Comments)	171
3H The Frequency of Occurrence in the Observations of the Use of Informal Language	175
3I Part A) Summary of Returns to the Student Questionnaire Concerning Students' Estimates of the Amount of Difficulty They Experience in the General Listening Comprehension Areas Part B) Summary of Returns to the Staff Questionnaire Concerning Staff's Estimates of the Proportion of Students Experiencing Difficulty in the General Listening Comprehension Areas	177
3J Summary of Returns to the Student Questionnaire Concerning the Amount of Difficulty They Had in Understanding Teachers and Other Students Summary of Returns to the Student Questionnaire Concerning the Amount of Difficulty They Had in Recognising Various Amounts of Discourse	181
3K1 Summary of Returns to the Questionnaire Concerning the Frequency With Which Duplicated Notes Were Provided Summary of Returns to the Questionnaire Concerning the Frequency of Occurrence of Other Methods of Taking Notes	184
3K2 The Frequency of Occurrence in Observations of Various Methods for Gaining a Written Record of Spoken Discourse	187
3L Summary of Returns to the Questionnaire Concerning QuA4/5 (Staff) the Proportion of Students Encountering Difficulty in Note-Taking and A7 (Student) the Amount of Difficulty Experienced	189
3M1 Summary of Returns to the Questionnaire Concerning the Frequency With Which Students Carried Out Various Reading Tasks During Their Degree Programme	199
3M2 The Frequency of Occurrence of Various Reading Tasks Established on the Basis of Information Gathered During the Observations	201
3N Summary of Returns Concerning the Amount of Difficulty Experienced by Students in Various Reading Tasks	209
3O Note-Making Activities	219

<u>Table</u>	<u>Page</u>
3P Johnson's Overview of the Rhetorical and Communicative Functions to be Found in Academic Writing	231
3Q1 The Frequency With Which Students are Expected to Produce Various Types of Written Work	234
3Q2 The Frequency of Occurrence of Various Writing Tasks Established on the Basis of Information Gathered During the Follow Up Interviews	235
3R Summary of Returns from the Student Questionnaire Indicating the Amount of Difficulty Experienced by British and Overseas Students in Various Aspects of Written Work	243
3S Summary of Returns from the Staff Questionnaire Indicating the Proportion of British and Overseas Students the Staff Considered to have Displayed Various Defects in Their Written Work	245
3T Summary of Returns From the Staff Questionnaire Concerning the Importance the Staff Attached to Various Criteria in Their Assessment of Students' Written Work	247
3U Summary of Questionnaire Returns Concerning Whether the Staff Expect the Same Standards of Written Work From Overseas as From British Students Summary of Returns on Whether These Allowances Vary as Between Coursework and Examinations	260
3V1 a) Summary of Returns to the Questionnaire Concerning the Frequency With Which Students Gave Oral Reports or Short Talks b) Summary of Returns to the Questionnaire Concerning the Frequency With Which Students Worked Together With Others Using English as a Means of Communication	274
3V2 a) The Frequency of Occurrence in Observations of Oral Reports or Short Talks in Which Students Make Statements	276
3V2 b) The Frequency of Occurrence in Observations of Pupils Seeking Information or Consulting	279
3W1 Summary of the Returns to the Questionnaire Concerning the Frequency With Which Students Actively Took Part in Discussions Involving the Teacher and the Class as a Whole Summary of Returns Concerning the Frequency With Which Students Asked Staff Questions	282
3W2 The Frequency of Occurrence in Observations of Students Referring to the Teacher	284

<u>Table</u>	<u>Page</u>
3X Summary of Returns to the Student Questionnaire Concerning the Amount of Difficulty Experienced by Students in Various Activities in Spoken English	287
3Y1 Summary of Returns to the Staff Questionnaire Concerning the Proportion of Students They Considered to be Experiencing Difficulty in Various Speaking Activities	288
3Y2 Numbers of Staff who Registered Answers in the 'Don't Know' Category or Left the Box Blank for Questions in This Section	289
3Z Summary of Returns Concerning the Frequency With Which Students Encountered Various Specific Problems With Spoken English	295
3AA The Academic Setting	306
3BB Interaction in the Academic Context	308
3CC Instrumentality in the Academic Context	310
CHAPTER FOUR	
4A Summary of Frequency and Difficulty Data Relating to Reading Comprehension Activities	334
4B Ordered List of Reading Comprehension Enabling Skills in an E.A.P. Context	335
4C Summary of the Difficulty and Frequency Data Relating to Listening Comprehension Activities	350
4D Ordered List of Listening Comprehension Enabling Skills in an E.A.P. Context	358
4E The Frequency With Which Students Have to Produce Written Texts of Varying Lengths, Broken Down According to Discipline and Academic Level	369
4F Ordered List of Writing Enabling Skills in an E.A.P. Context	370
4G Structural Areas for Grammar Test	373
4H Academic Writing Scale	388
4I The Percentage of Academic Staff Who Considered Each Criterion Important in the Assessment of Written Production Ranked in Order of Preference	391
4J T.E.A.P. Attribute Writing Scales	392

CHAPTER FIVE

<u>Table</u>	<u>Page</u>	
5A	Replies to Question 7 of the Follow-up Questionnaire	428
5B	Quantification of Short Responses to Question 8 of the Follow-up Questionnaire	431
5C	T.E.A.P. Attribute Writing Scales	457
5D	Contingent Means for Each of the Criteria Applied In the Assessment of Writing	458
5E	Pearson Product Moment Coefficients for the Seven Assessment Criteria Applied to the Writing Tasks	460- 461
5F	Cluster Analysis Dendographs	462
5G	The Percentage of Candidates Getting the Same Marks on the Different Criterion Descriptors in Each of the Writing Tasks	464
5H	Correlations Between Accuracy Composites (Assessment Criteria 5 + 6 + 7) in the Various Tasks	466
5I	Coefficients of Correlation Between the Total Task Scores Awarded by the Three Second Markers and the Original Markers	468
5J	Coefficients of Correlation Between the Grades Awarded on Each Criterion by the Three Second Markers and the First Marker	469
5K	Total Difference in Marks for All Candidates Awarded on Each Criterion by the Four Markers	471
5L	Total Mark Difference Between Markers on Each Criterion in Each of the Three Tasks	472
5M	Differences in Criterion Grades Awarded When Examiners Marked Scripts a Second Time	472
5N	Comparison of N.S. and N.N.S. Performance on Each of the Tasks in the Pre-Test Versions of T.E.A.P.	475
5O	The Results of t-Tests on 'A' Level and Undergraduate Students' Mean Scores by Task and Session	479
5P	The Results of t-Tests on Science and Engineering Students' Mean Scores by Task and Session	480
5Q	Inter Task Correlations	482
5R	Set 1 Factor Loadings Greater than 0.6 after Varimax Rotation A.S.A.B.S. and Sci./Eng. Students on Combinations of Session	489

<u>Table</u>	<u>Page</u>	
5S	Set 2 Factor Loadings Greater than 0.6 After Varimax Rotation. All N.N.S. Candidates on Combinations of Session	492
5T	Factor Loadings Greater than 0.6 after Varimax Rotation. All N.N.S. on Single Sessions	493
5U	Correlations Between Task Scores and Teachers' Communicative Assessments of Students' Reading, Listening and Writing Abilities	497
5V	Correlations Between Task Scores and Teachers' Global Estimates of Students' Reading, Listening and Writing Abilities	498
5W	Correlations Between Task Scores and Students' Self-Assessments of Reading, Listening and Writing Abilities	501
5X	Correlations Between Reading, Listening and Writing Task Scores, Amalgamated by Session, and Students' and Teachers' Assessments	503
5Y	Coefficients of Correlation Between Students' Self-Assessments and Teachers' Global and Communicative Assessments of Proficiency in Reading, Listening and Writing.	505
5Z	Comparison of N.S. and N.N.S. Performance on Skill Composites in Various Session Combinations	509
5AA	Inter-Correlations Between Construct Scores	511
5BB	Pooled Cronbach α 's for Skill Composites in the Various Combinations of Session	512
5CC	Product Moment Correlations Obtained Between First and Second Markings of Various Test Tasks	514
5DD	Correlations Between Skill Composite Scores and Language Teachers' Assessments of Students' Reading, Listening and Writing Abilities at the Time of Taking The Test.	516
5EE	Correlations Between Skill Composite Scores and Subject Tutors' Assessments of Students' Language Abilities at the End of the Second Term	517
5FF	Correlations Between Total Test Scores, Total Teachers' Assessments and Subject Tutors' Assessments (Language and Academic Progress)	519
5GG	Differences Between the Performances of Sci./Eng. and A.S.A.B.S. Students on TO/TA Combination	524

<u>Table</u>		<u>Page</u>
5HH	Differences Between the Performances of Sci./Eng. and A.S.A.B.S. Students on TO/TB Combination	525
5II	Differences in the Performances on TA and TB of Sci./Eng. and A.S.A.B.S. Students Who Took Both	527
5JJ	Concurrent Validity. Whole Session(s) Versus Other Test Results	530
5KK	Maximum Possible Scores on Test Tasks After Application of Intended Future Weightings	531
5LL	Background Variable Correlations With Skill Component Scores for Reading, Writing, Listening and Grammar	536
APPENDIX FIVE		
5MM	Number of Students (N.S. and N.N.S.) Taking Single Sessions and Combinations of Sessions at Each Centre	956
5NN	The Number of Students Completing Every Part of TO, TA, TB, TO + TA, TO + TB, or TA + TB According to Academic Level and Discipline Area	962
500	The Ages of N.N.S. Students Taking Single Sessions and Combinations of Sessions	963
5PP	The Sex of N.N.S. Students Taking Single Sessions and Combinations of Sessions	964
5QQ	The Country of Origin of N.N.S. Students Taking Single Sessions and Combinations of Sessions	965
5RR	The Number of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to the First Language in the Country of Origin	968
5SS	The Academic Level of Students Taking Single Sessions and Combinations of Sessions	971
5TT	The Number of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Their Specific Discipline Areas	972
5UU	The Number of N.N.S. Students Taking Single Sessions and Combinations of Sessions, According to Broad Discipline Areas	974
5VV	Period of Residence in Britain (In Months) for N.N.S. Students Taking Single Sessions and Combinations of Sessions	975

<u>Table</u>	<u>Page</u>	
5WW	Amount of Time Students, Taking Single Sessions and Combinations of Sessions, Spent (Outside of Class) With People Who Speak English	976
5XX	The Number of Years Students, Taking Single Sessions and Combinations of Sessions, Had Spent Studying English in Language Classes in Their Own Countries	977
5YY	The Number of Years Students, Taking Single Sessions and Combinations of Sessions, Had Spent Studying English in Language Classes in Britain	978
5ZZ	The Number of Students, Taking Single Sessions and Combinations of Sessions, Who Had Previously Been Taught a Subject(s) in English, e.g. Science, History, Economics, etc.	979
5AAA	The Number of Students, Taking Single Sessions and Combinations of Sessions, Who Had Studied The Subject(s) in English at Primary Level, at Secondary Level, As an Undergraduate, or as a Post-Graduate	980
5BBB	The Number of Students, Taking Single Sessions and Combinations of Sessions Who Had (1) Never, (2) Sometimes, (3) Often, Previously Needed to Read Articles, Books, etc. in Their Subject Area of Specialization in English	981
5CCC	Student Replies to Question 1 of Follow-Up Questionnaire	1067
5DDD	Student Replies to Question 2 of Follow-Up Questionnaire	1068
5EEE	Student Replies to Question 4 of Follow-Up Questionnaire	1069
5FFF	Student Replies to Question 5 of Follow-Up Questionnaire	1070
5GGG	Student Replies to Question 6 of Follow-Up Questionnaire	1071
5HHH	T012 Short Answer Questions on Reading Comprehension	1077
5III	TA11 Multiple Choice Questions on Reading Comprehension	1078
5JJJ	TA11 Reading Comprehension Item Response Frequencies	1079
5KKK	TA12A Gap-Filling: Identifying Omissions in Text	1080
5LLL	TA12B Gap-Filling: Completing the Gaps	1081

<u>Table</u>	<u>Page</u>	
5MMM	TA13 Short Answer Questions on Reading Comprehension	1082
5NNN	TB11 Multiple Choice Questions on Reading Comprehension	1083
5000	TB11 Reading Comprehension Item Response Frequencies	1084
5PPP	TB12A Gap-Filling: Identifying Omissions in Text	1085
5QQQ	TB12B Gap-Filling: Completing the Gaps	1086
5RRR	TB13 Short Answer Questions on Reading Comprehension	1087
5SSS	T021 Dictation	1088
5TTT	T031 Short Answer Questions on Listening Comprehension	1089
5UUU	TA21 Short Answer Questions on Listening Comprehension	1090
5VVV	Short Answer Questions on Listening Comprehension	1091
5WWW	T041 Multiple Choice Questions on Knowledge of Grammar	1092/3
5XXX	Knowledge of Grammar Item Response Frequencies	1094/5
5YYY1	Second Markings and Original Marks for Each Writing Criterion (T011)	1097
5YYY2	Second Markings and Original Marks for Each Writing Criterion (TA31)	1098
5YYY3	Second Markings and Original Marks for Each Writing Criterion (TA32)	1099
5YYY4	Second Markings and Original Marks for Each Writing Criterion (TB31)	1100
5YYY5	Second Markings and Original Marks for Each Writing Criterion (TB32)	1101
5ZZZ	Coefficients of Correlation Between the N.N.S. Group Scores on All Tasks in All Three Sessions	1118
5AAAA	Coefficients of Correlation Between the N.N.S. Group Scores on Tasks in All Three Sessions (4 Amalgamations)	1120
5BBBB	Coefficients of Correlation Between the N.N.S. Group Scores on Tasks in All Three Sessions (8 Amalgamations)	1122
5CCCC	Coefficients of Correlation Between the N.N.S. Group Scores on Tasks in All Three Sessions (10 Amalgamations)	1124

<u>Table</u>	<u>Page</u>	
5DDDD1	Reading Comprehension (TO/TA) Coefficients of Correlation	1127
5DDDD2	Reading Comprehension (TO/TB) Coefficients of Correlation	1128
5DDDD3	Reading Comprehension (TA/TB) Coefficients of Correlation	1129
5DDDD4	Writing (TO/TA) Coefficients of Correlation	1130
5DDDD5	Writing (TO/TB) Coefficients of Correlation	1131
5DDDD6	Writing (TA/TB) Coefficients of Correlation	1132
5DDDD7	Listening Comprehension (TO/TA) Coefficients of Correlation	1133
5DDDD8	Listening Comprehension (TO/TB) Coefficients of Correlation	1134
5DDDD9	Listening Comprehension (TA/TB) Coefficients of Correlation	1135
5EEEE	Contingent Numbers for Correlations Between Task Scores and Students' Self Assessments of Reading, Listening and Writing Abilities	1175
5FFFF	Contingent Numbers for Correlations Between Task Scores and Teachers' Communicative Assessments of Students' Reading, Listening and Writing Abilities	1176
5GGGG	Contingent Numbers for Correlations Between Task Scores and Teachers' Global Estimates of Students' Reading, Listening and Writing Abilities	1177
5HHHH	Contingent Numbers for Correlations Between Reading, Listening and Writing Task Scores, Amalgamated by Session, and Students' and Teachers' Assessments	1178
5IIII	Contingent Numbers for Coefficients of Correlation Between Students' Self-Assessments and Teachers' Global and Communicative Assessments of Proficiency in Reading, Listening and Writing	1179
5JJJJ	Contingent Numbers for Correlations Between Skill Composite Scores and Teachers' Assessments of Students' Reading, Listening and Writing Abilities	1181
5KKKK	Contingent Numbers for Correlations Between Total Test Scores, Total Teachers' Assessments and Subject Tutors' Assessments at End of Term Two.	1184

<u>Table</u>	<u>Page</u>	
5LLLL	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Various Age Groupings	1186
5MMMM	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Sex	1187
5NNNN	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Language Groupings	1188
50000	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Academic Level	1189
5PPPP	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to the Length of Time They Have Been In Britain	1190
5QQQQ	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to the Amount of Time Spent (Outside of Class) With People Who Speak English	1191
5RRRR	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to the Number of Years They Had Spent Studying English in Their Own Countries	1192
5SSSS	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to the Number of Years Spent Studying English in Britain	1193
5TTTT	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Whether They Had Previously Been Taught a Subject(s) Through the Medium of English	1194
5UUUU	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to Whether They Had Studied the Subject(s) in English at (a) Primary Level (b) Secondary Level, or (c) as an Undergraduate or Post-Graduate	1195
5VVVV	The Means and Standard Deviations of N.N.S. Students Taking Single Sessions and Combinations of Sessions According to How Frequently They Previously Needed to Read Books/Articles, etc., (in English) in Their Subject Area of Specialisation	1196

LIST OF FIGURES

CHAPTER ONE

<u>Figure</u>		<u>Page</u>
1	Total Number of Overseas Students in Universities as Against Other Sectors in Tertiary Education	11
2	Total Number of Commonwealth and Foreign Students	23

CHAPTER FIVE

1	Drawing Grade Boundaries on the Basis of Test Scores	541
---	--	-----

LIST OF ABBREVIATIONS USED IN THE TEXT

A.E.B.	Associated Examining Board
A.R.E.L.S.	Association of Recognised English Language Schools
A.S.A.B.S.	Arts, Social, Administrative and Business Studies
A.T.E.S.O.L.	Association of Teachers of English to Speakers of Other Languages
B.A.A.L.	British Association for Applied Linguistics
B.C.	British Council
C.I.L.T.	Centre for Information on Language Teaching and Research
C.N.P.	Communication Needs Processor
C.P.E.	Certificate of Proficiency in English (Cambridge Board)
C.S.E.	Certificate of Secondary Education
D.E.S.	Department of Education and Science
E.A.P.	English for Academic Purposes
E.F.L.	English as a Foreign Language
E.L.B.A.	English Language Battery
E.L.T.S.	English Language Testing Service
E.P.T.B.	English Proficiency Test Battery
E.S.L.P.E.	English as a Second Language Placement Examination
E.S.P.	English for Specific Purposes
E.S.P.M.E.N.A.	English for Special Purposes in the Middle East and North Africa
E.S.T.	English for Science and Technology
E.T.I.C.	English Teaching Information Centre
F.C.E.	First Certificate in English
F.L.T.	Foreign Language Teaching
G.C.E.	General Certificate of Education
G.L.P.	General Language Proficiency
H.M.S.O.	Her Majesty's Stationery Office
H.N.D.	Higher National Diploma
J.M.B.	Northern Universities Joint Matriculation Board

L.E.A.	Local Education Authority
M.A.L.S.	Midlands Association for Linguistic Studies
N.A.F.S.A.	National Association for Foreign Student Affairs
N.F.E.R.	National Foundation for Educational Research in England and Wales
N.N.S.	Non-Native Speaker
N.S.	Native Speaker
O.T.	Objective Test
P.C.A.	Principal Component Analysis
P.E.P.	Political and Economic Planning
R.E.L.C.	Regional English Language Centre (Singapore)
R.P.	Received Pronunciation
R.S.A.	Royal Society of Arts
S.A.T.	Scholastic Aptitude Test
Sci./Eng.	Science and Engineering
S.D.	Standard Deviation
S.E.D.	Scottish Education Department
S.E.L.M.O.U.S.	Special English Language Materials for Overseas University Students
S.T.O.S.	Science Teaching Observation Schedule
S.U.S.U.	Southampton University Students' Union
T.A.	Session IIA of T.E.A.P. for Non-Scientists/Non-Engineers
T.B.	Session IIB of T.E.A.P. for Scientists and Engineers
T.E.A.P.	Test in English for Academic Purposes (A.E.B.) To be known as T.E.E.P., Test in English for Educational Purposes, when the test becomes operational.
T.E.L.P.	Test in English Language Performance (Middlesex Polytechnic)
T.E.S.	Times Educational Supplement
T.E.S.O.L.	Teachers of English to Speakers of Other Languages
T.H.E.S.	Times Higher Education Supplement
T.O.	Session I of T.E.A.P.: General Academic
T.O.E.F.L.	Test of English as a Foreign Language

U.C.C.A.	Universities Central Council on Admissions
U.C.L.A.	University of California, Los Angeles
U.K.C.O.S.A.	United Kingdom Council for Overseas Student Affairs
U.S.R.	Universities' Statistical Record
U.W.I.S.T.	University of Wales Institute of Science and Technology

C H A P T E R O N E

THE BACKGROUND: THE PROBLEM TO BE STUDIED AND THE PURPOSE
OF THE RESEARCH

CHAPTER ONE CONTENTS

	<u>Page</u>
1. THE BACKGROUND: THE PROBLEM TO BE STUDIED AND THE PURPOSE OF THE RESEARCH	
1.1 RESEARCH PURPOSE: The Need for a Suitable Test	4
1.2 OVERSEAS STUDENT NUMBERS: Defining the Population for the Test	7
1.2.1 The Overseas Student	7
1.2.2 The Number of Overseas Students Enrolled on Courses in Tertiary Education	8
1.2.3 The Number of Overseas Students Accepted Annually for Courses in Tertiary Education	11
1.2.4 Survey of Individual Institutions in Tertiary Education	20
1.2.5 Future Trends	22
1.2.6 Conclusions	24
1.3 THE PROBLEM FOR STUDENTS AND INSTITUTIONS	25
1.3.1 The Nature and Level of Proficiency in English Required in an Academic Context	25
1.3.1.1 Present Language Entry Requirements of Institutions in the Tertiary Sector	27
1.3.1.2 How much is Enough? Selection and Cut-off Points	31
1.3.1.3 The Native Speaker Norm	32
1.3.2 The Need for Early Identification of a Shortfall in Language Proficiency	34
1.3.3 A Brief Survey of the Literature on the Problems Overseas Students Encounter in English Medium Study	37
1.3.3.1 Is it Just a Language Problem?	37
1.3.3.2 Establishing a Framework for Enquiry: a brief review of the literature concerning the problems encountered in the various study modes students have to operate in at the tertiary level	41
1.3.3.2.1 Lectures	42
1.3.3.2.2 Seminars	48
1.3.3.2.3 Practicals	50
1.3.3.2.4 Reading	50
1.3.3.2.5 Essays, report writing, dissertation	52
1.3.3.3 The Need for Remedial English - S.E.L.M.O.U.S.	53
1.3.3.4 Is Remedial Assistance Enough?	54

	<u>Page</u>
1.3.3.5 How Important is Language Proficiency for Academic Progress?	54
1.3.4 Conclusions	58

1. THE BACKGROUND: THE PROBLEM TO BE STUDIED AND THE PURPOSE OF THE RESEARCH

1.1 RESEARCH PURPOSE: The Need for a Suitable Test

During the period 1976-78 the Associated Examining Board, Aldershot, was approached by a number of its Centres and asked to make available a test which would provide receiving institutions in tertiary education with a comprehensive picture of the English language proficiency of students for whom English was not the mother tongue. These Centres expressed the need for a language screening device which would enable them to make decisions on a candidate's acceptability for an academic course and, in addition, indicate those areas where remedial language help was required. It was in response to this that the present study came to be undertaken. The writer was appointed as research assistant with responsibility for the research and development of the test. An advisory Working Party, consisting of specialists from the language testing field and representatives from various institutions in tertiary education, was established and this body has periodically monitored the project since its inception, in line with standard Board procedures.

There was evidence in the literature of a need for an appropriate language entrance test which would relate closely to the communicative needs of students in an academic context. In 1971 Wingard (p.55) had expressed the view that it was:

"... partly because of the lack of firm agreement as to the skills actually needed by students pursuing university courses in a second language, that English proficiency testing designed to establish levels of competence for this purpose has so far proved rather disappointing."

and this still appeared to be the case six years later, when Cowie et al. (1977a, p.8) noted:

"Before effective action can be taken on a national scale to overcome the many language problems of overseas students in Britain, there is first an urgent need for a reliable proficiency test in English, capable of identifying and accurately assessing the language needs of each student, in terms of his prospective course of study in Britain."

There was discontent with the existing English language examinations acceptable as evidence of language proficiency for entry purposes to institutions in the tertiary sector. A survey of the literature indicated dissatisfaction with both the nature and the amount of information that current English examinations were able to provide concerning the language proficiency of the candidates who sat them (cf. Chaplen 1970; Laing 1971; Wijasuriya 1971; Holes 1972; Morrison 1974; Moller 1977; Cheung 1978; Kelly 1978; Ryan 1979; Taylor 1979; Pickett 1980; Barnes et al. 1981; Torbe et al. 1981 and Moller 1982).

The English Language Testing Service (E.L.T.S.) of the British Council appears to have responded to demands from test users for more detailed information concerning the language proficiency of applicants for English-medium, tertiary level study in their research and development of a new battery, (v. Carroll, B.J. 1978a) intended to replace the English Proficiency Test Battery (E.P.T.B.) devised by Davies (1965). No evidence has been provided that E.P.T.B. was inadequate, either in terms of reliability or predictive and concurrent validity. The new English Language Testing Service (E.L.T.S.) battery instead lays claim to greater face and content validity. The design of the new E.L.T.S. battery is said to reflect more closely a change in emphasis in language teaching from an atomistic approach to a broader sociolinguistic one (cf. Carroll, B.J. 1978a; Clapham 1981 and Seaton 1981) and to take account of developments in English for Specific Purposes (E.S.P.) where the concern is no longer with catering, in a single test, for the needs of all users regardless of the purposes for which the language is required.

While there is clearly a connection between a student's proficiency in English and the degree to which he will benefit from and contribute to his course of study, we felt it might be difficult to predict academic problems, due to weakness in English, if there was not available a more accurate picture of the communicative demands made upon the student in his course of study than any so far established, and a profile of the student's language ability with respect to these. A behavioural analysis of the student's situation and of the language involved in the exercise of roles in that

situation would enable us to identify more closely the communicative skills required for study purposes (cf. Price 1977b and Ryan 1979).

After first identifying those institutions and subject areas that most overseas students elected to study in and then analysing the communicative demands that were made on students currently enrolled in the first year of courses in these areas, we would examine the feasibility of realising the resulting specification in a proficiency test appropriate for students coming to study in the United Kingdom at a variety of levels, in a variety of disciplines. We were interested in discovering whether one general proficiency test, with varying cut-off points, would be sufficient for assessing the proficiency of all students, regardless of level or discipline, or whether we would have to construct a variety of specific tests to accommodate these different levels and discipline areas.

Our aim was to establish a test which would tell us about a student's command of the language regardless of the means by which it had been achieved. The important point was to establish whether that command was sufficient for what the students needed to do. Our test was to be essentially a proficiency test and only diagnostic in the limited sense that we would offer a restricted profile of the student's ability in different media, which could inform decisions on a candidate's suitability for a particular course of study and enable appropriate remedial action to be taken where necessary.

1.2 OVERSEAS STUDENT NUMBERS: Defining the Population for the Test

1.2.1 The Overseas Student

We took as our starting point, in 1979, the identification of the main subject areas in tertiary education in which overseas students, whose first language was not English, had enrolled.

The definition of overseas student adopted by the Department of Education and Science (D.E.S.) and the Universities' Statistical Record (U.S.R.) is for fee purposes and, therefore, the published figures do not include those non-native speakers of English who qualified for the home rate of fees by fulfilling the necessary residence requirements; nor those who have qualified for Local Education Authority (L.E.A.) awards (v. Bristow 1977). This imbalance is countered by the fact that, whilst we are interested in those students from overseas for whom English is not the mother tongue, included in the national figures are a certain number of students from countries where English is either the first language (e.g. Canada or the United States of America) or is a medium of educational instruction (e.g. as in the case of Nigeria). Though the number of students from the English-speaking countries, to a certain extent, affects the total numbers of students referred to in Tables 1A to 1G below, it would seem reasonable to accept that these tables provide a fairly accurate indication of the distribution of overseas students, for whom English is not the first language, engaged in full-time study as between educational sectors and particular subject group areas. This is borne out by our survey of individual institutions, referred to in Table 1H below, where we were able to ask specifically for the number of overseas students whose first language was not English and so filter out most native speakers of English.

1.2.2 The Number of Overseas Students Enrolled on Courses in Tertiary Education

To establish the numbers of overseas students enrolled on courses in this country the Universities' Statistical Record (U.S.R.), the British Council and the Department of Education and Science (D.E.S.) were approached for statistics relating to the numbers of overseas students in each particular subject group category in the different sectors of tertiary education. Details of these statistics are to be found in Tables 1A, 1B and 1C and Figures 1 and 2 below. They were the latest available data as of 1st May, 1983.

In Table 1A the total numbers of overseas students attending courses at universities, polytechnics and further education colleges, in each of the D.E.S.'s nine broad subject area categories, are set out. The figures refer to all overseas students in each particular sector, irrespective of their year of study.

Two categories stand out as being of major importance in numerical terms:

- (i) engineering and technology (area 3)
- (ii) social, administrative and business studies (area 6).

Together they account for about 60% of all overseas students in all sectors of tertiary education.

TABLE 1A

NUMBER OF OVERSEAS STUDENTS ENROLLED ON COURSES IN
UNIVERSITIES, POLYTECHNICS AND FURTHER EDUCATION COLLEGES

	Education	Medicine/Dentistry and Health	Engineering and Technology	Agriculture, Forestry and Veterinary Science	Science	Social, Administrative and Business Studies	Architecture and other professional and vocational subjects	Language, Literature and Area Studies	Arts, other than Languages	Totals
	1	2	3	4	5	6	7	8	9	10
UNIVERSITIES										
(Undergraduate)										
1976-1977	151	1004	6621	113	2499	2986	216	963	1154	15 707
1977-1978	241	1120	7399	112	2920	3248	221	1000	1075	17 336
1978-1979	286	1158	7334	99	3393	3391	249	1086	1229	18 225
1979-1980	289	1160	6820	90	3489	3411	288	1022	1042	17 611
1980-1981	342	1098	6029	68	3018	3238	335	621	655	15 404
1981-1982	412	1130	5184	78	2569	3023	359	468	509	13 755
(Post-graduate)										
1976-1977	1246	1316	4069	620	4211	3766	565	1257	912	17 962
1977-1978	1354	1427	3971	606	4103	4021	569	1184	923	18 158
1978-1979	1421	1534	4165	624	4069	4072	643	1165	919	18 612
1979-1980	1357	1595	3962	617	3773	3763	587	1074	838	17 566
1980-1981	1263	1531	3491	584	3230	3698	512	919	643	15 871
1981-1982	1372	1461	2918	596	2840	3467	536	816	604	14 610
POLYTECHNICS										
†1976-1977	298	281	5903	-	2083	3977	860	310	562	14 274
†1977-1978	243	250	6353	-	2169	4203	954	167	456	14 796
†1978-1979	234	182	6106	-	2085	3934	937	209	528	14 215
†1979-1980	289	195	5654	-	2078	3336	830	168	489	13 039
†1980-1981	195	177	5172	1	1914	3301	962	136	382	12 240
‡F.E. COLLEGES										
†1976-1977	131	257	7359	144	644	6415	1636	1299	1641	19 526
†1977-1978	1092	251	7134	174	790	5298	1529	201	1507	17 976
†1978-1979	1184	261	7725	154	899	4907	1560	280	1585	18 555
†1979-1980	1209	232	7462	168	954	4667	1732	218	1525	18 167
†1980-1981	1071	278	6797	149	794	3919	1406	281	1191	15 886

SOURCES: Unless otherwise indicated, figures were supplied by the Universities' Statistical Record (U.S.R.).

† The British Council (1978a, 1979, 1980a, 1981a and 1982)
From 1977-78 onwards figures include certain higher education establishments as well as further education.

‡ Excluding students studying for G.C.E. or C.S.E. examinations.

Table 1B below shows the total numbers of students enrolled for courses in the three largest categories during the sessions 1976-77, 1977-78, 1978-79, 1979-80 and 1980-81. The figures in brackets indicate the percentage of the total number of students in all the subject categories for any one year.

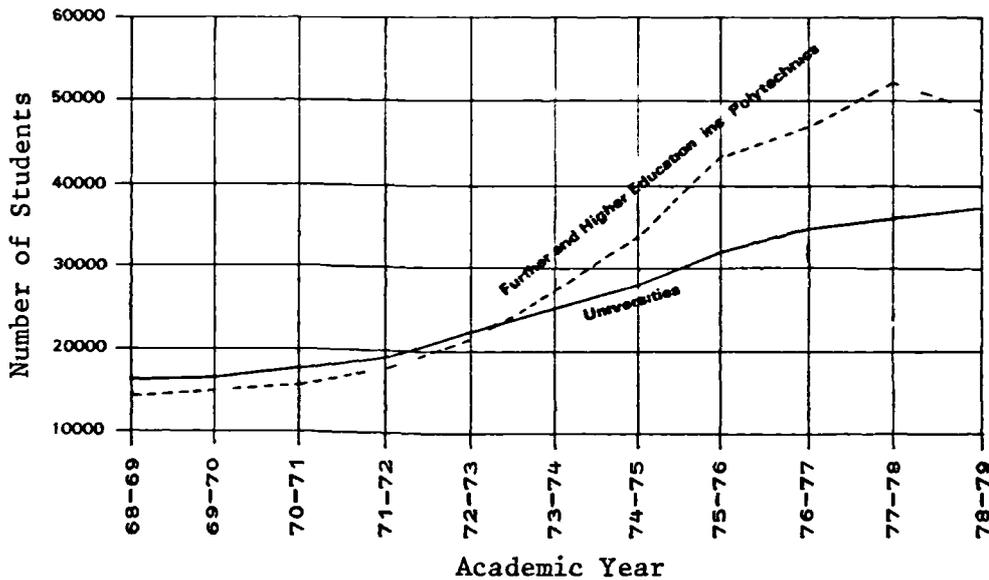
	Engineering & Technology <u>3</u>	Science <u>5</u>	Social, Administrative & Business Studies <u>6</u>
1976-77	23 952 (35%)	9 437 (14%)	17 144 (25%)
1977-78	24 402 (37%)	9 948 (15%)	16 430 (25%)
1978-79	25 330 (36%)	10 446 (15%)	16 304 (23%)
1979-80	25 058 (37%)	10 294 (16%)	15 177 (23%)
1980-81	21 489 (36%)	8 956 (15%)	14 156 (24%)

These three categories account for almost three-quarters of all overseas students enrolled at institutions in the tertiary education sectors, excluding students studying for G.C.E. or C.S.E. examinations.

An interesting trend in the education of overseas students in the United Kingdom is revealed in Figure 1 below which shows how the balance of overseas students in tertiary education has altered in recent years.

FIGURE 1

TOTAL NUMBER OF OVERSEAS STUDENTS IN UNIVERSITIES
AS AGAINST OTHER SECTORS IN TERTIARY EDUCATION



SOURCE: The British Council (1980a, p.3).

Since 1972 the other tertiary education sectors have attracted more overseas students than the universities. The figures indicate that in 1977-78 they catered for 14 331 more students than the universities though this increase had dropped to 12 499 by 1978-79 and, in the wake of increased fees for overseas students, the further education sector has been hit the hardest in terms of falling rolls.

1.2.3 The Number of Overseas Students Accepted Annually for Courses in Tertiary Education

The figures in Table 1C below relate to the annual number of new admissions to courses listed according to the nine D.E.S. broad subject area categories at both undergraduate and post-graduate level in the university sector, and those in Table 1D show the number of

new entrants in the other (i.e. non-university) higher and further education sectors. These figures are particularly relevant to our purpose, in that they tell us the broad subject areas in which overseas students are currently enrolling.

The figures in Tables 1C and 1D confirm the conclusions we drew from Table 1A about the distribution of overseas students as between subject categories. Area 3, engineering and technology and area 6, social, administrative and business studies are again the categories with the largest number of students.

It is noticeable in the university sector that more post-graduate than undergraduate overseas students enrol for courses in this country. Both Holes (1972) and Morrison (1974) drew attention to the far greater increase in the number of post-graduates as against undergraduates in the 1960s and early 1970s. Morrison (1974, p.2) records that, in the years 1964-71, there was a substantially greater increase in the numbers of post-graduates as against undergraduates, especially during the period 1968-71, over which: non-science post-graduates increased by 32%; science post-graduates increased by 33.4%; non-science undergraduates increased by 16.5% and science undergraduates increased by 8.1%. However, this markedly different rate of increase between post-graduates and undergraduates is not sustained for the years 1976-82, as can be seen in Table 1C below.

TABLE 1C

NUMBER OF OVERSEAS STUDENTS ACCEPTED FOR
COURSES IN UNIVERSITIES 1976-81

UNIVERSITIES	Education	Medicine/Dentistry and Health	Engineering and Technology	Agriculture, Forestry and Veterinary Science	Science	Social, Administrative and Business Studies	Architecture and other professional and vocational subjects	Language, Literature and Area Studies	Arts, other than Languages	Totals
(Undergraduate)	1	2	3	4	5	6	7	8	9	10
1976-1977	148	352	2682	37	1220	1406	90	628	891	7 454
1977-1978	145	375	2801	38	1393	1475	86	648	808	7 769
1978-1979	194	383	2392	31	1603	1594	108	698	951	7 954
1979-1980	191	361	2049	34	1443	1420	129	621	761	7 009
1980-1981	242	294	1736	20	933	1268	130	332	437	5 392
1981-1982	272	337	1748	32	881	1287	113	255	341	5 266
(Post-graduate)										
1976-1977	1111	630	2330	342	1991	2494	409	790	458	10 555
1977-1978	1214	655	2327	321	1843	2745	401	641	493	10 640
1978-1979	1236	670	2394	341	1844	2767	456	679	453	10 840
1979-1980	1168	720	2367	353	1616	2587	385	629	420	10 245
1980-1981	1061	624	1753	321	1235	2538	323	523	316	8 694
1981-1982	1127	668	1528	350	1152	2541	372	473	310	8 521

SOURCE: Figures supplied by the Universities' Statistical Record.

TABLE 1D

NUMBER OF OVERSEAS STUDENTS, FULL-TIME AND SANDWICH, ACCEPTED
FOR COURSES IN THE OTHER (NON-UNIVERSITY) HIGHER AND
FURTHER EDUCATION SECTORS IN TERTIARY EDUCATION 1977-78

1977-78

TOTAL ALL COLLEGES (includes Polytechnics)

Advanced

Education (teacher training)	623
Education (other)	113
Total Education	736
Medical, Health and Welfare	129
Engineering and Technology	3618
Agriculture	17
Science	1261
Social, Admin., Business Studies	4230
Professional and Vocational Studies .	735
Languages, Literature, Area Studies .	147
Arts excluding Languages	35
Music, Drama, Art and Design	<u>310</u>
Total Advanced	11 218

Non-Advanced

Education (teacher training)	-
Education (other)	20
Total Education	20
Medical, Health and Welfare	69
Engineering and Technology	3016
Agriculture	121
Science	394
Social, Admin., Business Studies	1249
Professional and Vocational Studies .	620
Languages, Literature, Area Studies .	6
Arts excluding Languages	-
Music, Drama, Art and Design	<u>578</u>
Total Non-Advanced	6073

G.C.E./C.S.E. 8573

Unrecognised 3696

GRAND TOTAL 29 560

SOURCE: Figures supplied by the Department of Education
and Science.

It is noteworthy that nearly 30% of all overseas students listed in Table 1D were following G.C.E. or C.S.E. courses.

Table 1E below gives more specific details of the university sector new entrants at post-graduate and undergraduate level in the three most popular broad discipline categories: area 3, engineering and technology; area 5, science; area 6, social, administrative and business studies. The figures in this table also reveal that there was no markedly different rate of increase in new post-graduates as against undergraduates in the years 1976-82.

Together with the information presented in Table 1C above they show, with few exceptions, that there was a general decline in the numbers of overseas students enrolling for courses in universities from the 1979-80 session onwards.

TABLE 1E

SUBJECT TOTALS OF NEW OVERSEAS ENTRANTS TO THE UNIVERSITY SECTOR AT UNDERGRADUATE AND POST-GRADUATE LEVEL
IN THE THREE MOST POPULATED DISCIPLINES

AREA 3 ENGINEERING AND TECHNOLOGY	UNDERGRADUATES						POST-GRADUATES					
	76-77	77-78	78-79	79-80	80-81	81-82	76-77	77-78	78-79	79-80	80-81	81-82
Aeronautical Engineering	96	79	89	48	55	47	17	22	24	24	16	10
Chemical Engineering	183	197	136	112	85	80	238	227	236	220	123	105
Civil Engineering	549	591	632	534	511	545	354	354	381	403	363	349
Electrical/Electronic Engineering	611	676	580	495	417	438	524	494	507	510	364	309
Mechanical Engineering	660	591	429	365	304	286	315	271	268	260	193	187
Production Engineering	94	71	44	44	44	47	97	115	115	97	53	46
Mining	33	40	28	31	31	40	32	38	31	44	30	23
Metallurgy	26	56	34	24	20	19	96	89	97	114	69	80
Other General and Combined Engineering Subjs. .	202	270	204	199	122	106	82	86	106	125	73	57
Surveying	23	25	37	30	37	46	13	17	24	18	15	16
Other Tech. and Comb. of Eng. and Tech.	110	105	94	83	110	94	502	551	519	477	454	346
AREA 5 SCIENCE												
Biology	64	72	76	62	37	46	177	166	147	139	118	109
Botany	9	4	9	3	0	3	75	77	66	60	55	41
Zoology	20	14	12	14	5	3	50	57	70	53	47	39
Physiology and/or Anatomy	13	13	24	13	7	10	23	16	17	17	13	18
Biochemistry	60	80	87	59	33	42	96	81	71	66	50	54
Other Gen. and Comb. Biological Sciences	34	47	52	32	18	9	8	13	35	22	43	38
Mathematics	370	407	471	463	369	372	588	543	580	464	342	337
Mathematics/Physics	13	8	14	20	9	6	-	-	-	2	4	2
Physics	65	110	133	141	102	64	262	248	251	215	151	139
Chemistry	178	175	160	153	116	89	501	431	403	391	261	222
Geology	11	19	26	15	7	19	109	120	121	103	81	78
Other Environmental Sciences	7	14	15	14	8	11	63	41	33	25	36	29
Other General and Comb. Physical Sciences	27	39	23	40	17	20	12	10	5	5	4	2
Comb. of Biological and Physical Sciences	216	224	288	221	205	185	9	22	12	14	5	2
AREA 6 SOCIAL, ADMINISTRATIVE AND BUSINESS STUDIES												
Business Management Studies	113	102	152	134	109	119	513	588	570	504	539	676
Economics	190	189	217	161	120	149	619	636	640	627	561	522
Geography	19	13	14	17	6	4	85	98	73	87	72	52
Accountancy	43	58	66	88	98	90	71	73	70	50	95	96
Government and Public Administration	100	81	64	60	34	28	349	404	404	377	429	440
Law	152	151	179	176	219	258	352	392	427	410	394	348
Psychology	63	61	50	41	32	18	72	99	83	65	43	38
Sociology	78	83	80	78	47	28	313	331	341	270	164	163
Social Anthropology	16	15	9	10	4	4	65	69	90	74	56	50

SOURCE: Figures supplied by the Universities' Statistical Record.

For the 1977-78 and the 1980-81 session, figures were also obtained relating to the number of successful applicants for undergraduate courses covered by the three most important subject categories; these are set out in Table 1F below.

These figures emphasise the demand for places, especially in subject areas 3 and 6. In engineering and technology approximately one student in three is accepted, whilst in science the acceptance rate is roughly one in four. This indicates a potential target population for the proposed test much larger than appears if one consults the statistics of those accepted.

In an additional survey of students taking Advanced ('A') level examinations for the General Certificate of Education (G.C.E.), we discovered the heaviest concentration was in the areas of mathematics and science. The figures below in Table 1G refer only to those students sitting G.C.E. 'A' level examinations in grant-aided establishments in the further education sector. They do not include students in the first year of two year 'A' level programmes or students in other types of establishments, e.g. colleges in the private sector. If it were possible to add these in, then the number of overseas students actually involved in 'A' level science and mathematics courses would almost certainly exceed the total number of students in this subject classification at the university level.

TABLE 1F

NUMBER OF OVERSEAS CANDIDATES APPLYING AND ACCEPTED FOR UNIVERSITY
UNDERGRADUATE PLACES IN AREAS 3, 5 AND 6 IN 1977-78 AND 1980-81

	ENGINEERING AND TECHNOLOGY			SCIENCE			SOCIAL, ADMINISTRATIVE AND BUSINESS STUDIES		
	Applying	Accepted	Acceptance Rate	Applying	Accepted	Acceptance Rate	Applying	Accepted	Acceptance Rate
1977-78	8360	2801	34%	2294	1393	60%	6039	1475	24%
1980-81	7440	2372	32%	2637	1224	46%	5821	1321	23%

SOURCE: Figures supplied by the Universities Central Council on Admissions (U.C.C.A.).

TABLE 1G

G.C.E. ADVANCED LEVEL SUBJECT ENTRIES MADE BY OVERSEAS CANDIDATES FROM
GRANT-AIDED FURTHER EDUCATION ESTABLISHMENTS

	1976-77*	1977-78†	1978-79†	1979-80†	1980-81†
English	190	150	90	80	140
History	170	120	110	63	90
French	100	190	100	30	20
Other modern languages	210	180	80	120	110
Art and Music	320	250	210	220	190
Other arts subjects	150	70	90	20	40
Mathematics	7960	7880	7620	8500	6770
Physics	3380	3390	3360	3470	2930
Chemistry	1880	1530	1560	1540	1480
Biology, Botany and Zoology	1120	730	810	690	750
Other sciences or technical subjects	160	230	160	180	120
British Constitution	1100	840	900	880	580
Economics	1780	1240	1410	1440	1220
Geography	250	170	150	190	120
Other social science or vocational ..	1170	990	1010	980	860
All subjects	19 940	17 960	17 660	18 403	15 420

* England and Wales

† England only.

‡ In mathematics there might be multiple entries by one candidate, e.g. combined mathematics, pure mathematics and applied mathematics.

SOURCE: D.E.S. Statistics of Education, Volume 2 School Leavers C.S.E. and G.C.E. 1978, 1979, 1980, 1981 and 1982.

1.2.4 Survey of Individual Institutions in Tertiary Education

As well as contacting appropriate advisory and administrative bodies for statistics relating to the numbers of overseas students in tertiary education in the United Kingdom, we had initially approached individual universities, polytechnics and further education colleges for details of overseas students accepted for courses in the 1978-79 session. We had asked each institution to provide details on a pro forma (v. Appendix 1.1, p.596) indicating the number of overseas students for whom English was not the mother tongue, studying in the various subject areas. In this way we hoped to filter out any overseas students, such as Americans, who were native speakers of English.

The universities were asked for details of post-graduates only, as at that stage, for reasons of time, convenience and unawareness of other sources, we planned to seek details of undergraduate admissions from the Universities Central Council on Admissions (U.C.C.A.) and the covering letter, accordingly, requested the university's permission for us to do this. We subsequently discovered that statistics relating to both undergraduates and post-graduates were available from the U.S.R. and finally obtained the figures from them, subsequent to our request having been approved by their Policy Committee and permission granted by the individual universities for their release to us.

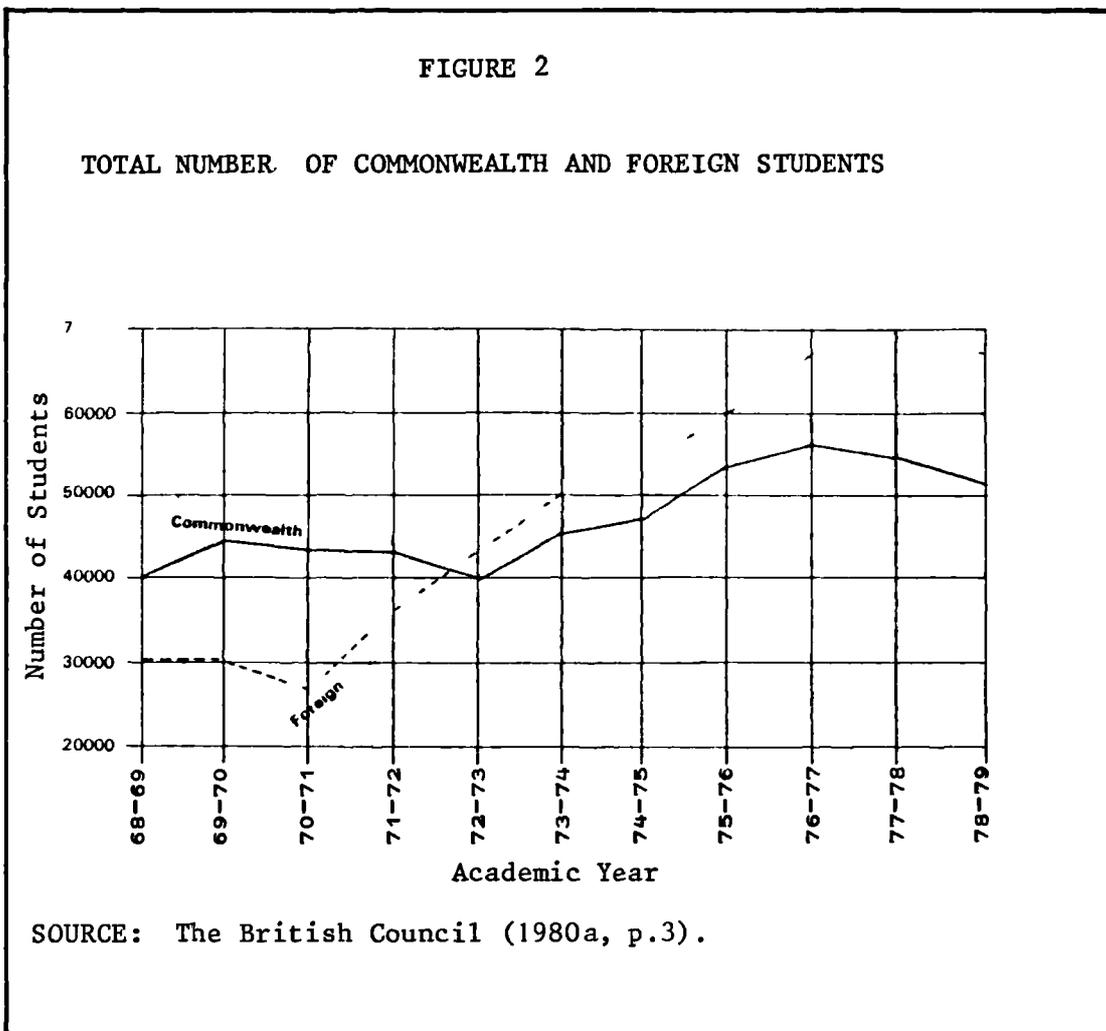
As a result of our own individual surveys sent out in October 1979 (v. Appendix 1.1, p.596) responses were received from 32 out of 64 of the universities, 101 out of 185 colleges and 14 out of 30 polytechnics to which we had sent requests. The totals for these sectors can be found in Table 1H below. The figures in this table lend further support to the conclusions concerning the most populated subject areas, made on the basis of the figures in Tables 1A, 1C and 1D above. In terms of numbers, areas 3, 5 and 6 are again seen to be the most heavily populated groups.

1.2.5 Future Trends

Although there is the possibility that the number of applicants might decline in view of the heavy increase in fees for overseas students beginning in the 1980-81 academic session, this does not necessarily entail that the number of entrants will drop. Institutions might accept candidates with weaker standards of English than in the past, particularly as non-E.E.C. overseas students are specifically excluded from university quotas and the Government is actively encouraging the recruitment of overseas students who are able to pay the increased fees. The problem of language competency will be exacerbated if the increased fees result in a decline in the number of Commonwealth students coming to this country and we see an increase in the balance of non-Commonwealth students, lacking the generally higher ability in English possessed by the former group. Albert (1976, p.7) had drawn attention to this changing composition amongst overseas students even before the current fees increase:

"... in 1972-73 the number coming from 'foreign' countries overtook the number coming from 'Commonwealth' countries for the first time and the proportions are now 53% foreign and 47% Commonwealth."

This trend can be seen clearly in the British Council's statistics for 1968-1979 summarised below.



Perren (1963, p.3) had noted that:

"... most students from non-Commonwealth countries, ... had never before been taught through the medium of English, although they might have spent a good deal of time in school, and after, learning English as a language."

Jordan (1977a, p.12) pointed to potential problems arising out of this situation:

"Overseas students likely to encounter the most serious language problems are those from non-Commonwealth countries. Consequently, an increase in their numbers points to an increase in the number of students with language difficulties. The number of overseas students at the University of Manchester reflects such an increase."

However, even within Commonwealth countries the language situation gives cause for alarm, for the position of English, as regards its role and status, is changing constantly in many of the countries

from which the majority of Commonwealth students are drawn. Edwards (1978, p.344) commented that:

"... (in Africa) it is gradually becoming an optional rather than compulsory subject and in Malaysia and the Philippines the transition from second to foreign language is almost complete."

The Southampton University Students' Union (S.U.S.U.) (1979) longitudinal study argued strongly that investigations into foreign students' command of English are even more urgent now, as the proportion of Commonwealth students with a second, as against that of students with a foreign, language background in English, decreases.

1.2.6 Conclusions

On the basis of the likelihood of the majority of overseas students studying in one of the three broad areas, engineering and technology, science, or social, administrative and business studies, it was decided that these were the areas in tertiary education on which we would focus our efforts to establish language needs, common across broad subject classifications. This needs analysis would be the first step towards devising a reliable and valid cross-disciplinary test of the communicative language ability of students coming to the United Kingdom to study in these subject areas, through the medium of English.

1.3 THE PROBLEM FOR STUDENTS AND INSTITUTIONS

1.3.1 The Nature and Level of Proficiency in English Required in an Academic Context

English proficiency, as far as receiving institutions are concerned, is often simply a question of whether a student possesses adequate English to be able to cope with his chosen course of study (v. Moller 1977). All too often administrators want a clear-cut yes/no decision. A number of questions arise when we talk of a student having adequate proficiency to cope with an academic course of study. What is adequate communicative ability in this academic context? What are the tasks a student has to cope with and what are the underlying enabling skills that a student has to use, receptively and productively, in his chosen course of study? Does he have what Davies (1965, p.11) called "control of English on all levels in appropriate situations"? What degree of it or lack of it is tolerated in different departments or by different staff within a department? Are differing levels of proficiency required in terms of different balances of language skills in different subject areas, or even for different courses within the same subject area? Does the difference extend as far as different components of the same course or even between different teaching situations in one component? What are the relative demands made on the overseas student at different levels: Advanced level G.C.E., undergraduate and post-graduate?

The question of what language proficiency is, will be a central concern of this thesis. Traditionally it has been defined in terms of performance in tests of linguistic competence, which assess the ability to produce grammatical sentences or utterances through a knowledge of linguistic rules. The more balanced of recent developments in approaches to language testing regard proficiency as a matter of communicative as well as linguistic competence, because the effective control of English in an appropriate situation requires command of use, as well as usage (contextual as well as linguistic or grammatical competence). For Davies (1977b, p.62) proficiency

considerations were a design influence involving an assessment of:

".. what the learners, whose proficiency is to be tested, need to do with the language, what varieties they must employ and in what situations they must use those varieties."

Heaton (1975a, p.164) echoed this:

"The proficiency test is concerned simply with measuring the student's control of the language in the light of what he will be expected to do with it in his future performance of a particular task ... The proficiency test is thus concerned with measuring not general attainment but specific skills in the light of the language demands made later on the student by his future course of study or job."

For Kelly (1978, p.218) the term proficiency test similarly denoted:

"... a test constructed to measure a candidate's ability to use the language of interest (the 'target' language) in certain specified communication situations ..."

and Moller (1981a), likewise, advocated a sociolinguistic-communicative approach where proficiency is seen as the ability actually to use the language in valid, sociolinguistic situations (cf. Morrow 1977, 1979; Carroll, B.J. 1978a, 1978b and 1980).

In the case of a foreign student being taught on an academic course through the medium of English, a circumscribed definition of proficiency is called for, as we are concerned here with proficiency in an academic context, rather than general English proficiency.

McEldowney (1976, p.5) observed:

"... neither the conversational or idiomatic English that is required for successful social intercourse, nor the type of English found in literature, is a central need for successful English medium study. What is considered to be central is a proficiency in the more expository, neutral, transactional type of English that is the medium of education in English-speaking countries."

She argued that, in order to be proficient in English-medium study, students need to control this academic 'expository' English in the following ways: they need to be able to understand the spoken mode for listening to lectures and discussions, to understand the written mode for reading text-books and other sources of information; to produce adequate written English in their set work and examinations

and to produce adequate spoken English when necessary, e.g. in discussions, asking questions and presentations.

The proficiency test we set out to develop was to be used for assessing the language ability of overseas students who applied to continue their studies in the target language, English. The communication situations which were of interest were those they would be expected to cope with in respect of their formal studies and candidates would be measured against the demands these made.

1.3.1.1 Present Language Entry Requirements of Institutions in the Tertiary Sector

In October 1979, a letter was sent to the academic registrars of all the universities, polytechnics and further education colleges in the United Kingdom requesting that they tell us the numbers of overseas students (for whom English was not the mother tongue) enrolling for the first time in 1978 in specific subject areas (v. Appendix 1.1, p.598). In addition to this request we asked:

"Could you also let us know whether you set any English language requirements which must be met before foreign students are allowed to join your courses?"

The replies to this question have been abstracted from the letters we received and are included as Appendix 1.3, pages 633-666.

The majority of universities will accept the Joint Matriculation Board (J.M.B.) test in English or the Cambridge Proficiency in English (C.P.E.) test, though some would still prefer Ordinary ('O') level English Language, and some accept or prefer other tests, such as the English Proficiency Test Battery (E.P.T.B.), the American Test of English as a Foreign Language (T.O.E.F.L.), Certificate in Secondary Education (C.S.E.), the Scholastic Aptitude Test (S.A.T.), or their own internally administered tests. The Cambridge First Certificate in English (F.C.E.) is not normally acceptable, though there are exceptions.

These findings accord with those of a similar survey made by James et al. (1977) who sought to ascertain from the replies

of the universities, the relative acceptability of the J.M.B. test as against the C.P.E. The authors of that report emphasised that, although the majority of universities state some proof of competence is required, at least for undergraduates, the reality could be quite different because the requirements were not applied uniformly (p.12):

"... while universities reserve the right, and justly so, to decide on individual cases, we have found the policy regarding an English qualification wildly fluctuating and often contradictory to the requirements as stated to us last year. Some of these contradictions include the acceptance of some overseas students without any qualification in English ..."

They also found that some universities did not recognise the J.M.B. test while others did not recognise the C.P.E. These inconsistencies are even more marked in the returns to our survey as we had asked what was acceptable to them and did not specifically enquire as James et al. (1977, p.12) did:

"... if an examination is called for, do you require the Joint Matriculation Board (J.M.B.) examination in English for foreign students or the Cambridge Proficiency examination in English, or do you regard them as equally acceptable?"

By leaving the question open ended, we received a surprising variety of replies concerning language entry requirements, as can be seen from Appendix 1.3, pages 633-666.

This picture is reinforced by the findings of Ryan (1979) who sent a questionnaire regarding English language requirements to 39 universities in May 1977. The replies he received substantiated the findings of James et al. (1977). Ryan (p.27) found that, in general:

"... most universities accepted '0' level English Language for overseas undergraduates and also accepted the Cambridge Proficiency Examination and the J.M.B. Admission Test in English (Overseas)."

The heterogeneity of entrance requirements is even more marked at the post-graduate level as Cowie et al. (1977a, p.7) remarked in their introduction to a Special English Language Materials for Overseas University Students (S.E.L.M.O.U.S.) report entitled English for Academic Purposes:

"The situation has quickly become serious since although it is usually necessary for students to provide evidence of their ability in English to follow courses at the undergraduate level, very few universities and polytechnics require formal English language qualifications for those admitted at the post-graduate level."

Ryan (1979, p.28), who had specifically requested information about overseas post-graduates, found:

"... three of the thirty one universities which replied stated that the same language requirements applied as to overseas undergraduates. Five universities said that the same requirements did not apply. Most universities left the question unanswered. One university summed up what is perhaps the general situation in the majority of universities with the honest remark - 'There are no entry requirements for overseas post-graduates (alas). Each department is a law unto itself'."

Some universities now require post-graduates to take an examination in English when they arrive, but this is not normally part of the selection process and the tests are normally used diagnostically for remedial purposes. Morrison (1974) cited the University of Newcastle as an example where all new post-graduates are required to take the English Language Battery (E.L.B.A.) test and a certificate that they have done so is required before they may register. It does seem that at some universities, very low scores on these tests in no sense preclude registration, though low scorers are 'required' to attend remedial language classes in the Michaelmas term. Various internal tests are compulsory at some other universities and strongly recommended at others (cf. Chaplen 1970 and Heaton et al. 1975). The availability of these tests is restricted though and many universities do not seem to administer their own tests, however individual departments may well attempt to do so.

What is apparent is a lack of any standardised policy on the part of the universities and the fact that all of them reserve the right to decide on an individual's particular case. We are certainly in agreement with James et al. (1977, p.12) in their conclusion that "some standardised policy should be evolved".

The replies from the further education sector indicate that there is even less of a consistent standard in language entry requirements,

policies varying within many of the institutions themselves, as well as between institutions. In many ways the situation is even more critical here as there is even less likelihood of any appropriate remedial language assistance being made available than there is in the universities. Larter (1962, p.119) commented:

"The extent to which proficiency in English is made a condition of entry to courses in Further Education here varies from the requirement of an external certificate to merely a nominal condition that the student 'shall be reasonably proficient in written and spoken English'. In fact the overseas student enrolling at a technical college is unlikely to have to offer any very firm proof of his language skill before he is permitted to enrol ... Although the importance of language skill is recognised, difficulties in assessing it and imposing definite conditions on entry, arise from the unreliability of overseas certificates and, as far as the technical colleges are concerned, the sheer administrative pressure of enrolment weeks."

Edwards (1978) argued that overseas qualifications were often misleading and Walker (1978) observed in his study of overseas students in the further education sector that there was a wide gulf between the possession of such paper qualifications and the ability actually to converse in the language. Larter (1962, p.124) questioned the validity of overseas qualifications as criteria of practical English skills:

"They have often tested the student's success in studying English as a formal subject and offer little guide to his potential skill in using it as a medium of study ..."

Edwards (1978, p.312) also referred to another problem:

"... many Training Schools showed a certain amount of mistrust concerning the 'paper qualifications' of overseas learners because of the possibility that they might not be genuine ... Some overseas candidates wrote to say that they had not got the necessary qualifications and could not afford the bribes for documents. Could the School help?"

Whilst the number of bogus certificates in circulation is in all probability very small, the difficulties of examining in the far flung corners of the earth are not. In a situation where untrained, unstandardised examiners provide assessments of ability on the basis of short interviews (as in the British Council's subjective assessment) there is obviously cause for concern. As Moller (1977, pp.

26-27) observed:

"... it is important to realise that there are limits to the information obtainable from any English language test. This is particularly true of tests administered overseas where control of the administration may not always be strict ... The chances of obtaining the opinion of the same assessor for more than one candidate in a given group of applicants for any one course in a British university are indeed slim. It, consequently, becomes necessary to put faith in the opinions and whims of a large number of assessors throughout the world ..."

He (p.25) pointed to a further limitation that tests carried out overseas may have, namely that there is usually a time lag of ten months between the taking of a proficiency test by a student overseas and his arrival in this country for matriculation:

"The question now arises as to the extent to which the assessment at the time of application is still valid at the time of matriculation ten months later."

and concluded (p.32) that any measure carried out overseas would:

"... be essentially the first step to further diagnostic assessment which could be most profitably carried out at the time of a student's arrival in Britain and prior to his or her pre-session English language orientation course. This could be followed by yet another measure as the student begins his academic course."

In Section 1.1 we referred to the fact that there was evidence in the literature that existing English examinations were, in some respects, qualitatively deficient. It seems that present attempts to establish students' language ability before entry to a course may also be considered quantitatively deficient in that entry requirements, where they exist, are not uniformly applied to all overseas applicants, particularly post-graduates.

1.3.1.2 How much is Enough? Selection and Cut-off Points

Davies (1965, p.13) outlined the problem in defining 'adequate' proficiency for English medium study:

"For a receiving institution it is the course of study rather than the student which raises the real proficiency problem. Do different courses need different levels of English proficiency?"

Davies went on to describe an English Proficiency Test Battery which could be used to predict whether a student had sufficient English to complete his academic studies successfully. In testing the hypothesis that different cut-offs are not needed he found that there was a single cut-off point for students at all academic levels, below which students might be deemed to be lacking the necessary English proficiency for successful completion of their studies. He added the warning (p.219):

"Where a cutting score is actually to be drawn depends entirely on the needs and wishes of a receiving institution."

Restrictions on the range of subjects within the academic levels meant he was unable to make separate analyses for academic subject variations in proficiency.

Sen (1970) found that, on a simple pass/fail basis, 35% of the students in her survey would have been misplaced by the Short Form version of the Davies test. Chaplen (1970) suggested that the problem with a single cut-off point is that it fails to take into account the wide variation in standards expected in departments within and across institutions. It may well be the case that different courses require a higher or lower cut-off point dependent on the language demands they make on students (v. Alderson et al. 1981).

We attempted to investigate whether different tests were needed for different discipline areas or whether one test was sufficient. It was our intention to provide a profile of a student's performance on the listening, reading and writing components of the test battery in the form of a set of behavioural grades. It would then be possible for receiving institutions to decide whether the levels reached by a candidate were adequate for the course he wished to pursue.

1.3.1.3 The Native Speaker Norm

In any discussion of what is adequate proficiency for a non-native speaker we need to have some idea of the proficiency of native speakers. Perren (1963, p.14) commented:

"... often we take it for granted that the British student has acquired the skills he needs with his G.C.E. But we know that the foreign student who has the same G.C.E. has not necessarily acquired the skills he needs."

One of the areas that will be considered below is that of 'the native speaker norm'. We attempted to establish how far, in practice, overseas students experienced language-based problems in their academic work, in excess of their British counterparts (v. Chapter 3 below). Very little work would seem to have been done on comparing the two groups' relative levels of language difficulty in the various academic tasks they have to perform and all too often the false assumption is made that the native speaker has no problems, with the result that the overseas student is expected to reach the standard of a mythical norm.

Austin Ward (1979) examined the competence of a sample of British craft students from a variety of courses in the further education sector. He comments (p.424) that only 17% of these demonstrated any 'satisfactory' competence in their written work:

"The analysis of the scripts for 'mechanical competence' revealed that the majority of students found great difficulty with spelling, punctuation and grammar."

He quotes (p.427) from The Bullock Report (D.E.S. 1975):

"Many allegations about lower standards today came from employers who maintain that the young people joining them cannot write grammatically, are poor spellers and generally express themselves badly."

One would expect a higher standard than this from the native speakers in the range of courses we are surveying, from G.C.E. 'A' level to post-graduate level, but evidence from the questionnaire returns described in Chapter 3 does indicate the extent and gravity of the language-based problems for some of the British students even at these levels.

1.3.2 The Need for Early Identification of a Shortfall in Language Proficiency

We need to be able to identify those overseas students who may possibly under-achieve because of a shortfall in their English ability. The problem is a serious one because, without some means of identifying very early on in which study modes difficulty may occur, it may well be late on in the first term before academic staff have sufficient evidence to make any decisions on a student's need to have remedial help with his English. Larter (1962), Chaplen (1970) and Morrison (1974) all drew attention to the fact that the departmental tutor is by no means certain to realise the extent of a student's problems immediately and that linguistic inadequacy can remain hidden for some time. If a student carefully controls his linguistic output by limiting himself to the structures he has confidence in utilising, his tutor may well be left with a false impression about his spoken English and also mistakenly infer a lack of any problems on the part of the student in comprehending spoken discourse. Of course, the student may not help matters by choosing to affect complete comprehension for various reasons, for example, through a desire not to lose face. It may also happen that it is not until the student has to produce written work at a late stage in the course that he himself becomes fully aware of the extent of his own problems. Morrison (1974, p.4) concluded:

"... departmental-tutor assessments of overseas post-graduates' command of English are inherently unreliable."

This is a disturbing state of affairs, especially for those on nine month post-graduate courses for, given the short length and heavy demands of these courses, it is likely that students will have very little time to spare for remedial English which is often seen as extra-curricular study. This pressure is likely to be even greater in the spring and summer terms and, as Jordan et al. (1973, p.46) concluded in their survey of 106 overseas post-graduate students at the Universities of Manchester and Newcastle:

"Once a student falls behind with his language improvement, rarely does he catch up, especially on a one year post-graduate course."

Given the proliferation of one year Masters and Diploma courses in most universities and the trend for non-Commonwealth and thus non-English medium educated post-graduates to attend them, Davies's (1965, p.117) warning that:

"... many of the more serious language problems are faced by the overseas post-graduate students, who are in Britain for one year or even less."

should not be ignored.

Chaplen (1970, p.1) noted that, in the case of undergraduates:

"The majority of non-native undergraduates gain their university entrance qualifications in Britain and will frequently have spent three or four years studying through the medium of English before going up to university."

a view supported by Morrison (1974). Thus, not many are likely to suffer severe problems and they probably have three years to overcome the difficulties.

Undergraduate matriculation often requires the applicant to submit evidence of his adequacy in English, whereas for a post-graduate there is not always this formal requirement. The receiving department's criterion for acceptability in the latter case is often just proof of academic qualification in a relevant area of study. Often the student feels that his acceptance onto a course tacitly implies, in the eyes of the receiving institution, the possession of adequate English to cope with the demands of the course. In a sense he is correct and acceptance would seem to impose an obligation on the receiving institution to ensure that the student's progress is not impaired by inadequate linguistic ability.

The transition from home country to the United Kingdom may bring unexpected problems for some overseas students. Chaplen (1970) and Morrison (1974) pointed out that post-graduates were likely to be experiencing English as a medium of instruction for the first time, in contrast, perhaps to an earlier exposure to it merely as a subject of study. Jordan et al. (1973) suggested that, by the standards of his home country, a student's English often seemed good or, at least, adequate; only after arrival in Britain does the full force of his language inadequacy strike him.

During the period of study in the United Kingdom, the amount of contact overseas students have with native speakers may be limited, with a resultant effect on language acquisition. Cultural barriers, British reserve, pressures of work, shortness of stay, language difficulties, age differential between post-graduates and undergraduates; all of these can reduce the possibilities of integrating into the community and of developing language ability. Daniel (1975) and Walker (1978) argued that, in fact, there is no reason to assume any integrative motivation on the part of the overseas student, for he may choose to reject our culture or, at least, consciously resist integration into it. Where there are large nationality groups it is much easier to stay within a particular group for reasons such as common language, cultural identity and social survival. The post-graduates are often enrolled on one year taught courses in classes which contain a sizeable majority of overseas students. These are the people with whom they will normally communicate, in and out of class. It is, perhaps, not surprising that Edwards (1978, p.342) found in her survey of overseas student nurses in Britain:

"At present there seems to be little real opportunity for many overseas learners to gain very much practice in the use of English other than the formulaic patterns of professional interchange ... Consequently, it is hardly surprising that it often takes so long before improvement is noticeable."

Given the limited opportunities for language improvement, especially for those on one year post-graduate courses, an early identification of those students with a shortfall in their English proficiency would seem to be essential. If they are to be accepted onto programmes of study, both they and the academic staff involved need to be appraised of any language problems and provision made for any remedial language work as early as possible in the course, or even before it begins.

1.3.3 A Brief Survey of the Literature on the Problems Overseas Students Encounter in English Medium Study

1.3.3.1 Is it Just a Language Problem?

Studies of overseas students in Britain date back to 1907 with the report of the Lee Warner Committee on Indian students and the point emerges clearly from subsequent studies that language is only one of a myriad of problems overseas students face on entry to this country. This point is emphasised in Singh (1963), Burns (1965), Davies (1965, 1977a), P.E.P. (1965), Dunlop (1966), Morris (1967), Holes (1972), Daniel (1975) and Walker (1978). An extensive and thorough analysis of approaches to identifying these problems in the literature can be found in Hawkey (1982).

The extent and nature of the language problem will be discussed below, but first the socio-cultural, cognitive and affective factors which often compound language difficulties will be briefly considered.

Dunlop (1966, pp.7-8) noted:

"It soon became clear that young people from one European country seeking education or training in another, met with fewer difficulties than those coming from another continent, and particularly than those whose cultural background had not fitted them for conditions of life and study in a European country."

Whether this is due to problems of "cultural incommensurability" (v. Edwards 1978) or social pressures due to colour prejudice or xenophobia is difficult to assess. There is obviously a good deal of sensitivity amongst overseas students about these attitudes, as can be seen in some of their essays in the collection edited by Tajfel et al. (1965). The National Foundation for Educational Research (N.F.E.R.) enquiry (Burns 1965) into the adjustment and attitude of overseas students holding Commonwealth bursaries in England and Wales discovered that a majority felt some sense of handicap because of a difference in race or colour. The 1965 P.E.P. report, based on interviews and questionnaires conducted amongst students from former British overseas territories, found that the

attitude of the British towards overseas students came in for some strong criticism. Reed et al. (1978) mentioned that the overseas visiting student often felt he was being treated as an intruder and confused with immigrants.

Both Holes (1972) and Edwards (1978) have commented on the problems associated with cultural incommensurability. Edwards instanced the 'shyness' of many South East Asians and Holes suggested this might be due to a combination of factors: lack of confidence in their English and fear of the English students' ridicule of clumsy or ill-formed sentences (especially in the case of the weakest students); a malaise induced by the unfamiliarity of a situation where their judgements might be questioned (particularly true of older students) and not least, an ignorance not so much of the language, but of the cultural admissibility of interrupting a lecturer or tutor to ask a question and how to ask the question itself. Holes (1972, p.28) quotes from Fishman (1969):

"... native members of such (speech) networks and communities slowly and unconsciously acquire sociolinguistic communicative competence with respect to appropriate language use. They are not necessarily aware of the norms that guide their sociolinguistic behaviour. Newcomers to such networks or communities ... must discover these norms more rapidly, more painfully and therefore more consciously."

This is evidenced in the way most foreign students ask questions, by approaching members of staff after the class. This is a strategy which is perhaps indicative of a gap in their communicative competence, namely ignorance of the sociolinguistic conventions associated with asking questions in public.

Related to cultural problems are differences in previous educational background and Edwards (1978) drew attention to the difficulties this may cause in terms of non-participation in group discussion and problems arising out of self teaching. Singh (1963) noted instances of excessive deference to staff and it would seem many overseas students have been taught to venerate their teachers, speaking only when spoken to or not answering for fear of losing face. Edwards cited previous exposure to didactic teaching methodology and Singh mentioned previous training by rote methods as factors influencing

the students' preference for being told what to do rather than organise their own study. On the basis of observations in the classroom, Edwards noted that (p.321):

"... the tendency towards rote-memorisation and regurgitation was apparent in some overseas learners' answers, often with little real cognisance of the question asked."

Many of these problems were also referred to by students themselves in the collection of overseas student essays entitled Disappointed Guests (Tajfel et al. 1965).

On arrival in this country a number of other cognitive, affective or social factors will affect an individual's language learning and/or acquisition. A survey of the literature suggests that, in particular, the learner variables of age (v. Hawkey 1982), aptitude (cf. Davies 1965; Gardner et al. 1972; Carroll, J.B. 1979 and Hawkey 1982), attitude and motivation (cf. Gardner et al. 1969, 1972; Burstall 1975 and James 1980a), personality (v. Altman 1980), cognitive style (v. McDonough 1981) and sex (v. Davies 1965) may either separately or together affect language learning and acquisition.

Davies (1965, p.12) noted:

"English proficiency itself may well be affected by a number of variables such as sex, age, home language and country, as well as by intelligence. From a theoretical point of view such influences could be highly important; but from the practical point of view of a receiving institution in this country what matters is not the biographical history of an overseas student but whether his English is adequate for the course he wants to take."

Nevertheless, he does admit (p.12) that such information could have practical value in giving advice as to:

"... the likelihood of so many students of a certain age, sex, language background, reaching the necessary level of proficiency; and also in making long term plans for English language teaching programmes in different areas where proficiency may be very high or very low."

Cowie et al. (1977a, p.7) observed:

"The most serious language problems are met by those students who arrive in Britain each year with scarcely any

ability to communicate at even the most elementary level. Language handicaps of this nature clearly give rise not only to serious academic problems, but also to considerable welfare problems."

Even those most proficient in English may encounter academic, social or welfare difficulties because of inadequate previous exposure to spoken English.

Rogers (1977) pointed out that many students are also unfamiliar with 'social English', the language of everyday interaction with its attendant conventions of formality and informality, dependent on socio-cultural setting.

It is well documented that inadequate English can lead to social difficulties for some students (cf. Perren 1963; Morrow 1977; Morrow et al. 1977; Morrison 1974 and S.U.S.U. 1979). Johnson et al. (1976, p.3) argued that overseas students have a real need for social language which they define as:

"... language which is used either simply to establish or confirm interpersonal relationships or else to perform other functions which depend for their successful realisation on the establishing or maintaining of such relationships."

in order to function as a 'whole person' within a foreign environment and culture. They noted that:

"... it seems likely that in many cases it is only when social adjustment has begun to take place that the student's psychological set is such that he can gain the maximum benefit from the course of instruction he is following."

Holes (1972, p.45) drew attention to the impossible situation language deficiency can place the overseas student in:

"Failure to integrate socially was a damaging consequence of inability to communicate effectively in English. This tended to isolate some students who found themselves in a vicious circle: their English did not improve much because they got no practice in speaking it and vice versa."

and (p.55):

"... those in need of constant language practice are usually disappointed in their search for it. They often

become disillusioned as a result and withdraw defensively into the safety of their national or regional group, where, of course, they speak no English."

Jordan et al. (1973, p.40) concluded:

"... one of their biggest problems was to find the opportunity to practise speaking English with native English speakers ... Consequently they made slow progress with their spoken English."

Jordan (1977a) also stated that 56% of overseas students in his survey found it difficult to meet British people to converse with and this led to restrictions on opportunities for practising spoken language.

Whilst we accept that there will be an inevitable overlap between language ability and the socio-cultural, cognitive and affective factors referred to above, the focus for this study will not be multidimensional (v. Hawkey 1982) but will be confined to investigating the difficulty students find in operating in what Candlin et al. (1978) term 'study modes'. These authors related 'study skills' and 'linguistic skills' through a framework where these 'study modes' are superordinates, with 'macro-skills' such as listening comprehension and note-taking as hyponyms. Their model does not split English according to the traditional four skills, what Corder (1973) described as 'epiphenomena' and thus the same skill may appear in more than one mode.

1.3.3.2 Establishing a Framework for Enquiry: a brief review of the literature concerning the problems encountered in the various study modes students have to operate in at the tertiary level

Once we examined the literature concerning the academic language needs of overseas students it soon became evident that the language proficiency of some overseas students could not be described as adequate when they started courses in this country. The inadequacies detailed below might well be merely the tip of the iceberg as, for the most part, they emerge from the work done in those institutions that are fortunate enough to have language servicing facilities which can provide remedial English language tuition.

A common factor among many of the overseas students entering educational institutions in the United Kingdom is that they have come not to learn English, but to learn something else which is taught through the medium of English (cf. Perren 1963; Davies 1965 and Jordan et al. 1973). It is likely that a number of these students will have some difficulty in operating in English during the course of their study, especially in those cases where no allowance is made as far as their academic work is concerned for their being overseas students, where they follow the same course, receive the same tuition and take the same examinations as native speakers. Candlin et al. (1978) argued that if few or no concessions of a linguistic or study skill nature are made during their courses, then in all probability they will need a high level of competence in a wide range of study modes.

We needed to establish a framework for specifying those features of target language use that might be important to a participant in his/her academic life. The primary informing source for this has to be real life events and descriptions of the relevant features of these were established through empirical research. It was useful though, in the first instance, to turn to secondary informing sources to see what information they provided concerning the study modes in which students had to operate and those in which they experienced difficulties. This provided an indication of the size and nature of the problem and informed us subsequently in our approach to data collection.

1.3.3.2.1 Lectures

Jordan (1977a) described how post-graduate students at Manchester and Newcastle heard most spoken English in a passive listening role, i.e. in situations where they were not called upon to respond at all, for example, listening to lectures, where no check is made on comprehension. In his study, 70% listed understanding spoken English as their biggest difficulty on arrival in the United Kingdom.

Edwards (1978) found a similar pattern in her study of overseas nurses coming to study in the United Kingdom and concluded that the overseas learners' language problems lay mainly in the field of

spoken English. She describes how in both understanding and speaking, over 78% admitted to having difficulties of some kind with this.

Davies (1965) took a group of non-native speakers of English studying at institutions abroad through the medium of English and compared them to an equivalent group in the United Kingdom, to see if any difference due to a short period of residence in the country was revealed. He discovered that in non-contextual listening tests the overseas group were inferior and concluded that this (p.235):

"... could only be accounted for by their lack of exposure to everyday British English in Britain."

Sen (1970) also found that 25% of the students she surveyed admitted to having listening comprehension difficulties. It would seem likely, therefore, as Morrison (1974) argued, that lack of exposure to natural spoken British English accounts for the initial difficulty that many overseas students experience on arrival.

The students' command of reading skills or the possession of a paper qualification in English language may have led to false expectations about their ability to follow the spoken word. The English they were exposed to in their own country, e.g. in language classes, may have been significantly different from British varieties of natural spoken English. In the academic environment in the United Kingdom the speaker is primarily concerned with the transmission of informational content and not with difficulties arising out of unfamiliarity with the language itself (cf. Larter 1962; James 1972, and Jordan et al. 1973).

Morrison (1974) drew attention to the fact that there is often some improvement in this area after a period of exposure to everyday British English, a factor noted also by Davies (1965), Chaplen (1970) and Jordan et al. (1973), though Chaplen did add the rider that the improvement is linked to the degree of attendance at remedial English classes where there is a concentration on the improvement of oral/aural skills.

Most research into the spoken receptive mode has focused on the lecture study mode, though at some levels it is obviously as relevant to the seminar and practical modes as well.

Wijasuriya (1971) found overseas students had difficulties in lectures due to a complex of factors. At the phonological level, both he and Morrison (1974) referred to the difficulties caused by the speed at which lectures are delivered. They also detail evidence of the difficulty occasioned by a variety of native speaker accents as do Larter (1962), Sen (1970), Jordan et al. (1973), U.K.C.O.S.A. (1974), Edwards (1978) and Walker (1978). Candlin et al. (1976) also referred to difficulties with phonology (elision, reduction, intonation and regional accent) and James (1977) noted the difficulties many overseas students have in decoding a lecturer's utterances because of unfamiliarity with English stress-timed rhythm and sometimes arbitrary lexical stress.

At the morphological level Greenall (1980) noted the difficulties students have with inflection and both he and Tinkler (1973) are concerned with the difficulties caused to students by passives.

At the level of lexical meaning, the difficulties met by students are as relevant to the reading and writing study modes as they are to lectures, seminars and practicals. Chaplen (1970), Wijasuriya (1971), Morrison (1974) and Candlin et al. (1976) all cited lexis as a source of difficulty, especially where there is no direct translatability back into the first language and Hutchinson et al. (1981) reported difficulty when there was a lack of precision in the vocabulary used instead of standard technical terms.

Ryan (1979, p.6), in a study which focused on language-associated problems amongst science and technology students at U.W.I.S.T., drew attention to the fact that each subject area has its own language register:

"... i.e. particular lexis and syntax that are features of that subject and to some extent distinguish it from other subjects, from the point of view of language."

and noted that this occasioned overseas students a good deal of difficulty.

Larter (1962) also referred to this as a particular problem and source of misunderstanding with many difficulties arising from the different socio-cultural associations words may have for students from different backgrounds, even though familiar with the lexical meaning of certain elements of vocabulary. Problems with technical register were also remarked on by Walker (1978) who cited examples of technical words which have a special meaning in engineering such as 'apron stud', but also have a different non-technical meaning.

As Widdowson (1978) pointed out though, register involves much more than technical vocabulary; including as it does, the rhetoric, the illocutionary acts of scientific English as well as the linguistic conventions of technical writing. A mastery of register is essential in written coursework; it also affects reading and listening comprehension in the chosen area of study.

At the level of cohesion, Wijasuriya (1971) and Widdowson (1978) found inter-sentence connectives a source of difficulty as was the referential system, especially logical connectives, according to Morrison (1974). Candlin et al. (1976) discovered in their survey of engineering lectures that the referential system (anaphora, etc., transition markers, logical connectors) was the greatest single cause of difficulty. James (1977) noted that comprehension was often impaired by an inability to grasp discourse markers and antithetic or concessive sentence connectives.

Wijasuriya (1971) examined 46 hours of taped lectures (psychology, clinical sciences, statistics, economics, administrative studies, electronics, biochemistry) for the purpose of identifying, classifying and quantifying the occurrence of discourse markers and inter-sentence connectives, i.e. cohesive elements. He argued that recognition of these features aids the understanding of the development of an argument; they signal change of direction in an argument and also enable the listener to predict the line of an argument. He analysed the problems overseas students are likely to have had with these inter-sentence connectives and discourse markers.

He found (p.151) that failure to decode discourse markers was a source of problems:

"They do not quite realise it when the direction of the argument changes and so on and they also find it hard to evaluate the relative importance of various points in the discourse."

He pointed out that logical connectives are unlikely to cause much difficulty per se as far as the foreign student following a lecture is concerned. The problem is that the connectives may contribute to the redundancy of language and, if their signalling potential is not recognised by the student, then this redundancy value is lost. Holes (1972), in his investigation of the difficulties of the lecture mode for students, also found that the greatest difficulties for those with low ability in English arose out of the fact that in lectures the language has a lower level of redundancy for them than it has for native speakers. Wijasuriya (1971, p.115) emphasised that:

"... the greatest problem the student has to face is the fact that many different types of cohesive unit can occur in the same sentence."

and cites the example: "... and so therefore this is what they did." The problems that connected discourse present the student with cannot be considered in isolation, they are cumulative and only so many factors can be retained in the short term memory at one time. Dinning (1977) pointed out that the problem is intensified by the fact that the memory span is shorter in the second language (L2) than in the first language (L1).

At the level of coherence, Holes (1972) cited register switching as a serious cause of difficulty in the lecture mode. He (p.33) drew attention to the fact that:

"... the distinctions between formal and informal language may be blurred in a lecture situation where there may be drastic change from formal to informal language when giving explanations ... it could be disastrous for the foreign student who may have been taught English in a combination of literary and formal English."

Rogers (1977) and Jordan (1977a) remarked on the difficulties of students not knowing the meaning or appropriate use of informal

colloquial English. Edwards (1978, p.315) commented in a similar vein:

"Since indigene learners also used slang and/or highly allusive language, which would exclude any outsider, it is likely an overseas learner misses a considerable amount of what is said in the classroom."

Sen (1970), Holes (1972), U.K.C.O.S.A. (1974) and Jordan (1977a) all referred to the difficulty of comprehending lectures in general and taking notes. Johns et al. (1977a) were concerned more specifically with the difficulty overseas students had in grasping the basic argument and relating it to the framework of the subject under discussion.

The study skill of note-taking was seen by James (1977) as a problem involving four main activities for the student: understanding the message; identifying the main points; deciding when to write them down and writing them down quickly, so that comprehension of the ongoing lecture is not interrupted, and clearly so that they will be understood at a later date.

In the study skills area Matthews (1978) referred to the problem of eliminating secondary and tangential remarks and Candlin et al. (1976) to problems in accurate transcription of information from the blackboard. Price (1980) cited the problem some students had with reorganising essential information from notes and graphic modes of presentation, and Wallace (1980) problems with reconstituting notes.

The difficulties in comprehension are often compounded by the teaching staff themselves, as Edwards (1978, p.316) observed:

"The instances of inappropriate methods of presentation of practical procedures, inadequate explanation of theory and lack of advance organisation of material - which were disturbingly frequent - observed in the classroom would almost certainly compound the initial problems of overseas learners and may well account for the difficulties in understanding, experienced by some indigenes."

1.3.3.2.2 Seminars

The perceived areas of difficulty in the literature referred to above in connection with the lecture mode nearly all relate to the seminar study mode as well. They were included in the foregoing section, on the grounds that this was where the acknowledged researchers originally identified them. There is however, evidence of additional problems occurring in the seminar mode.

Morrison (1974) set out to identify the study modes in the academic context which caused the greatest difficulties for overseas post-graduates in science oriented disciplines at the University of Newcastle-upon-Tyne and to identify, in such modes, those features which were a persistent cause of difficulty. The modes comprised not only seminars and tutorials where questions and discussion are the norm, but also informal lectures where these features are likewise to be found. He discovered that the greatest relative incidence of difficulty occurred in seminars and tutorials. These were closely followed by informal lectures, then formal lectures and lastly, individual discussion (v. Black 1971).

Rogers (1977, p.37) discovered in the courses he ran for post-graduate students in science and technology, that despite much previous work in the language, students:

"... were unable to participate in academic discussion: even less were they able to take part in social activities outside the company of other foreign students in the same plight as themselves."

Mackenzie (1977, p.41) also found that for the Latin-American students he taught:

"... lack of oral fluency often proves an insuperable obstacle to effective participation in seminars and tutorials."

It seems to be mainly in the production of coherent discourse that seminars put an additional burden on the overseas student, though Johns et al. (1977a) referred to problems in understanding the varying realisations of functions and Hawkey (1982) to problems associated with topic switching. Edwards (1978, p.316) instanced

where students:

"... had difficulty in remaining linguistically coherent when attempting to answer at length or when joining in discussions."

She comments (p.321) that if one adds to this difficulty the problems with accent, more general difficulties in understanding the teaching staff and:

"... the indigene learners' near total ellipsis when answering questions and the speed at which they do so."

one has some idea of the factors which might inhibit participation. Edwards concludes that:

"... what learner-initiated participation there was in the classroom activities was dominated by indigenes."

At the coherence level Holes (1972) and Jordan (1977a) noted difficulties in communicating functionally, especially in asking questions for appropriate purposes. Similarly the U.K.C.O.S.A. (1974) survey found:

"... a reluctance to ask questions due to a fear of using English. This contributes to the formation of nationality groups where students hardly ever speak in English."

Rogers (1977) and Jordan (1977a) both found that overseas students did not take an active part in discussions and gave restricted answers when questioned, due to problems with fluency and self-expression. Rogers indicated that some also had problems because they were too formal and polite; they had difficulty with 'openers', 'closers' and 'topic change'; they had difficulty with humour of all kinds or did not understand the various conventions of non-verbal behaviour.

Johns et al. (1977a) looked at the problems faced by students in seminars and found that the sheer variety of teaching situations, labelled seminars, hampered attempts to devise suitable teaching strategies, for remedial work. They noticed, in particular, that students seemed to have problems with turn taking, metacomment, mitigation and repair.

On the matter of study attitudes and habits, Brew (1980) noted the difficulty some overseas students had in handling the questioning approach when they come from an educational system where the authority of the teacher and of established knowledge goes unchallenged. Watt (1980, p.42) added:

"There is generally overmuch apparent respect accorded to the teacher; this inhibits argument and the development of a critical approach."

1.3.3.2.3 Practicals

A large number of the problems referred to in the sections above on lectures and seminars are equally relevant to practicals, on the phonological, morphological, lexical meaning, cohesion and coherence, study skills, attitudes and habits, levels of description.

With particular (though not exclusive) reference to practicals, Hutchinson et al. (1979) mentioned further problems overseas students have at the coherence level in dealing with 'glossing techniques', cultural references, assumptions of shared knowledge and in imposing a coherent structure on information received and produced.

In terms of study attitudes and habits related to practicals, Fitch (1980) drew attention to how ill-prepared overseas students are for experimental work and the recording of results in a log book and Dudley-Evans (1978) to problems some have in appreciating the significance of experimental results, no doubt due to the absence of such learning experiences in their own countries.

1.3.3.2.4 Reading

The study mode of reading is likely to remain important through most courses. Holes (1972) commented that although students considered reading text-books the most important area of their studies, many did not consider it difficult. Straker-Cook (1977, p.45) noted that:

"... the reading of specialist literature in English is the one skill that most students seem to have maintained prior to their arrival in Britain."

Against this viewpoint Jordan (1977a) recorded a general inability on the part of the overseas student to read quickly or understand the complexities of academic prose. He found that the average student had only one speed (i.e. slow) for silent reading - about 150-160 words a minute. Edwards (1978) noted similar difficulty with reading comprehension and U.K.C.O.S.A. (1974) mentioned "difficulty in reading effectively" as a common problem, as was "difficulty in understanding examination questions". Holes (1972) referred to slow reading as a universal complaint as well as a widespread lack of guidance in what to read. He got the impression that the cause of low reading speeds was (p.66):

"... in some measure due to a reverential attitude to books in general: each one read must be summarised with a great expenditure of time and effort. There was no 'skimming' of books for idea and information. This was the first time for some students that they had had to read books in any quantity, with specific aims in mind."

At the level of lexical meaning, Johns (1978) noted the problems caused by density of unknown vocabulary and metaphorical usage. Holes (1972, p.66) echoed this when he pointed out that it was not the actual level of syntactic difficulty in the text which caused problems, but rather:

"Terminological difficulties in new subjects were the biggest obstacle in reading, as is shown by the large numbers of students (among them the most fluent) who read their text-books with a dictionary at hand."

Ryan (1979) and Templeton (1973) also recorded difficulties encountered in mastering the technical register of a subject area.

As regards cohesion, Sim (1974) drew attention to the problems brought about by sentence connectors, Wallace (1980) to semantic markers and Johns (1978) to non-linear information structure.

On the coherence level Johns et al. (1977a) referred to the difficulties in handling modality, inference, probability, obligation and hypothesis. Imhoof et al. (1975), Widdowson (1978) and Johnson (1979) pointed to those in dealing with the illocutionary functions used to create different kinds of discourse, and Johns (1978) to problems in predicting the writer's intentions.

There is a lot of evidence that problems are also serious at the study skills level. Brew (1980) instanced the problem of sorting out the main points from supporting details and Wallace (1980) to abstracting the organisational pattern and main ideas from a text. Johns (1978) and Morrow (1980a) referred to problems in surveying for gist and scanning for information and Wallace (1980) to those in surveying and referencing.

1.3.3.2.5 Essays, report writing, dissertation

Jordan et al. (1973) and Jordan (1977a) argued that as the course develops, writing skills eventually become more important and will generally, according to Jordan (1977a, p.18):

"... supersede understanding as a cause of major difficulty."

as written work has to be submitted.

Larter (1962) felt that particular difficulties in this area may only surface even later in the course when writing has to be done under the pressure of time constraints as in examinations; by then it may well be too late for remedial action. Jordan (1977a) likewise noted an acute problem in writing quickly.

There are close links between the problems overseas students face in the reading mode and their own attempts at written production (v. Johnson 1981). At the phrase structure level, Greenall (1980) pointed to additional problems in the use of the article, spelling, passivisation, relativisation and complex nominalisation, and Johns et al. (1977b) to problems in the use of determiners.

At the coherence level, Kaplan (1966), Johnson (1977a) and Houghton (1980) noted difficulties students from different cultures had in achieving a coherent discourse structure in English, what Bruner (1975) terms 'analytic competence', (the ability to structure thought in linguistic terms). Jordan (1977a) and Rogers (1977) referred to difficulties in writing concisely and Greenall (1980) to evidence of a possible tendency on the part of the overseas student to aim at minimal content.

The study skill of organisation (v. Wallace 1980) and researching of a piece of writing (v. Johns 1978) would seem to present great difficulty. Wallace (1980) referred to difficulties in organising essays, Dudley-Evans (1978) to organising reports and Price (1980) to organising a dissertation.

1.3.3.3 The Need for Remedial English - S.E.L.M.O.U.S.

In view of the many difficulties in English for academic purposes faced by overseas students, a number of lecturers formed the S.E.L.M.O.U.S. group in June 1972. Their aim was to share experiences in dealing with overseas students' English language difficulties on the basis of investigation of language needs and the production of relevant teaching materials for use in any necessary remedial work. This group is very much aware of the extent of the language problems involved as one of its members, Jordan (1977a, p.13) stated:

"Once a student has been accepted and arrives at the university, no matter how poor his command of English, he is rarely asked to delay his academic studies in order to attend a full-time course of English. Yet S.E.L.M.O.U.S. members estimate that about 30% of those students that they teach are in need of full-time English tuition ranging from 3-12 months."

S.E.L.M.O.U.S. has been prominent in highlighting the particular language needs of overseas post-graduate students and Price (1977a) described how its members have been instrumental in setting up pre-session courses to try to improve the language performance of students before their academic studies get under way. Given that present language entry requirements are not uniformly applied (see Section 1.3.1.1 above) and vary as between institutions and that present attempts to assess student language ability are considered not wholly satisfactory, there is a serious possibility that remedial teaching, either pre- or in-session, will be necessary because of a shortfall in some overseas students' language abilities.

The very existence of the S.E.L.M.O.U.S. group adds weight to the contention that language problems amongst overseas students are a cause for concern and there is a felt need for better testing

procedures which will point to areas where remedial work is needed. Language disabilities do not necessarily entail failure for the overseas student, but it is likely that the more proficient they are in the language, the more they will benefit from their chosen courses of study.

1.3.3.4 Is Remedial Assistance Enough?

Morrison (1974) drew attention to the possibility that the crucial factor might be the individual's language ability when starting a course and that average improvement during a year-long course is insignificant when compared to differences between individuals on entry. He refers to work done at Gothenburg which suggested that some students were better on starting than others were after a year's study.

Evidence is available that universities have in the past accepted students with critically low scores on the Davies Test (E.P.T.B.) and English Language Battery (E.L.B.A.). In these cases language proficiency on entry would certainly seem to be an important factor governing a student's relative performance.

1.3.3.5 How Important is Language Proficiency for Academic Progress?

A composite picture of the overseas student's situation is presented by Sen (1970) who carried out by questionnaire, English test and interview, a study of 2367 overseas students and 553 nurses from 130 countries, studying in Britain from 1964-66. Amongst other things, she investigated their academic and language difficulties and the problems they encountered in adjusting to English society. She subsequently followed up this work with a study of how successful they were on their courses. She was not concerned solely with the language difficulties they encountered, as her aim was more to describe the interaction of the general criteria: academic ability, English language proficiency, adequate financial resources and adeptness in adjusting to new social situations. Like Larter (1962), Davies (1977a), Mackenzie (1977), she considered academic study and

welfare problems inseparable for most students. She also raised the important question of the extent to which language is critical to a student's academic success.

Davies (1977a) noted that Sen (1970) found English language to be non-significant as a predictor of academic success. Sen (1970, p.163) claimed on the basis of evidence from her follow up study:

"... that the extent of the use of and familiarity with the English language has little relevance to their final performance. This seems to contradict the experience of teachers in this country and in the United States who have placed great emphasis on the English language proficiency of overseas students."

She arrived at this conclusion from the results of student performance in five sub-tests of the Short Form version of the Davies Test. She found that these tests did not discriminate between the passes and failures for the qualification courses considered (cf. Davies 1977a; Moller 1977 and 1982) and so she argued (p.154):

"... for practical purposes the tests do not, alone, provide a useful guide to final performance in these courses. Of course, the diagnostic - as against the predictive - value of the tests is a different matter."

However, she does qualify these remarks (p.163):

"... the criterion of final performance used in the present survey i.e. success or failure, is crude. Certain modifications of the test in the light of the present results and its validations on the basis of actual examination marks with larger samples for each type of course might provide a useful instrument for the selection of overseas students."

Hawkey (1982) expressed concern over Sen's use of an objective, norm-referenced proficiency test, designed to act as a broad guide for placement purposes, without consideration of whether a more criterion-referenced performance battery would be more suitable where diagnosis and prediction are involved.

Earlier in her study Sen drew attention to the serious nature of some of the problems. Over a third of the 2367 students found writing essays difficult and a large number had difficulties in following lectures and tutorials. If one examines her relative

figures for different national groupings, however, they point to what may be an even greater cause for concern in the future, namely the worse plight of the foreign as compared to Commonwealth students. She notes that the Middle Eastern students, who scored less than the other groups in the Davies test (p.58):

"... on the whole seem to find most difficulty with their studies. Only few expressed no difficulty with lectures, tutorials and in contacting their teachers, and a high proportion found writing essays and reference reading 'very difficult'."

A majority of the students in her survey were from the Commonwealth and in terms of their previous use of English as a medium of instruction, the earlier age at which they had begun to learn English, and their use of English at home, Sen found they had had considerably more exposure to the language than their foreign counterparts. This is a real cause for concern, for, as was shown by Figure 2, page 23 above, the balance between foreign and Commonwealth students coming to study in this country changed in the early 1970s and now there are considerably more foreign than Commonwealth students.

The English ability of the foreign students will, in all probability, be lower than that of the Commonwealth students because of a much more limited exposure to the language and it might be reasonable to infer that these students are likely to encounter language difficulties proportionate to this in their academic courses.

Davies (1977a, p.36) also raised the question of how serious a problem language is, on the basis of his findings in a survey completed for the Scottish Education Department (S.E.D.) on the English proficiency of foreign students in Scotland, in the non-university sector:

"... it would be a mistake to exaggerate the place of language among foreign students' problems. To do so can be a means of evading the examination not only of other problems of learning and teaching but also of welfare problems. What all foreign students require is a minimum general English, a language core. Thereafter other factors (e.g. academic ability) are important in contributing to students' performance, their welfare and their success."

Davies found in his survey that students ranked English language as being the most serious problem, followed by academic status, social contacts and accommodation, which seems at least to contradict the findings of Stevenson (1974) referred to in the same paper, who carried out a survey of the welfare problems of overseas students at Edinburgh University in which language was not placed high in the rank ordering of difficulties encountered.

The view that language is an important factor in academic progress is however, indicated by the N.A.F.S.A. (1961) survey and subsequent studies. Hebron's (1967) survey (quoted in Walker 1978) of overseas students at Rutherford College, found language to be important and Ryan (1979) quotes evidence from Campbell's (1974) study of undergraduate and post-graduate students following courses in technical subjects at the Loughborough, City and Birmingham Universities. Campbell concluded that difficulties students had in communication were not solely attributable to an insufficient proficiency in English, but more to a complex mixture of linguistic, academic, socio-cultural and practical problems. However, a majority of the students in Campbell's study did indicate that understanding lecturers, writing and speaking were areas where they had particular difficulty. Walker (1978) found that in the interviews he carried out with academic staff at the technical college level, the study skills area was often cited as presenting a major problem for the overseas student. Hawkey (1982) refers to the British Council's (1981) Survey of the Factors affecting the Performance of O.D.A.-Sponsored Study Fellows, the main finding of which was that the 'unsatisfactory' performance of 30% of the sample was due mainly to inadequate language proficiency and motivation and a failure to adapt successfully to the new academic and social environment. As Davies (1977a) pointed out, the importance of language is open to debate. Opinion in the literature is mixed though the majority of studies reviewed do indicate that serious problems exist for a number of overseas students.

Sen (1970, pp.161-162) mentioned an important caveat in interpreting studies of the difficulties experienced by overseas students:

"... the relative proportions of students who are willing to admit difficulties - whether academic or personal - may not always give an accurate picture of the extent to which such difficulties are encountered. It is our impression, supported we believe by the general tenor of the figures reported, that the picture these students and nurses present of themselves is unduly favourable and optimistic. This suggests that where problems appear or criticisms are voiced, they deserve more, rather than less, attention than the cold tabulations would imply."

This was borne out by Chaplen's (1970) work on this subject and by Jordan (1977a, p.14) who found:

"In 1972-73 and again in 1974-75 students' self assessment ratings were examined and compared to the students' scores in the Chaplen test. Overwhelmingly the results showed that students at the lower end of the scale in the tests grossly over-estimated their language ability."

Walker (1978) similarly pointed out that student awareness of language inadequacy is often lacking and that there was a general tendency for students to over-estimate their own language ability. This is obviously an important factor to be borne in mind when looking at surveys carried out by questionnaire and interview which report that language is not seen by students to be a problem.

1.3.4 Conclusions

Although many of the problems overseas students face in this country can be attributed to social and other factors, there is no doubt from our search of the literature that a certain number also arise through the relationship of the student's ability in language and the requirements of his study. Thus, as Davies (1965, p.12) pointed out:

"... while English proficiency may not be the most important factor for all overseas students, it is one important factor for them all, one that it may be possible to isolate for testing purposes, perhaps, indeed one - the only one! - that can be radically improved when revealed."

Adequate proficiency in English may not guarantee a student success in an academic course and lack of it may not entail failure. It is, as we have noted above, only one of the factors involved. It does,

however, mean that a student will be able to compete on a more equal footing with native speaker counterparts and have the opportunity to derive the maximum benefit from a course of instruction in this country.

Our survey indicates that it is the problems caused by the 'higher order' language skills at the coherence and cohesion levels, together with study skills and study attitudes and habits, which receive most attention in the English for Academic Purposes (E.A.P.) literature. Far less attention is paid to 'lower order' language skills at the phonological, morphological, lexical and phrase structure level and there seems to be an implicit assumption that an ability in the latter is subsumed by an ability in the former. This emphasis would seem to be in accord with the communicative approach that is currently in vogue in language teaching and to a lesser extent in language testing.

It is also noticeable that there is very little evidence in the literature of any attempts to survey the problems encountered by overseas students across disciplines and levels. Most of the studies we have quoted from are concerned only with small sub-sets of the population in which we are interested and most comments relate to a particular academic course in which the author is involved. There is a need to collect data on a wider basis to indicate the nature of study-related activities and resultant problems, which will inform the construction and evaluation of the test tasks in our proposed Test in English for Academic Purposes (T.E.A.P.). In this way we will try to arrive at valid and reliable measures of the state of a student's academic language proficiency, prior to the commencement of a course of study or as soon as possible after it. The value of this needs-based communicative approach to test task construction is discussed in Chapter 2 below and its feasibility will be a recurrent theme of this thesis.

In Chapter 3 below, the findings of a series of systematic observations of the functional English medium tasks faced by students on science, engineering, social, administrative and business studies courses at various levels (in tertiary education)

are reported. In addition the returns to 560 staff and 1470 student questionnaires, designed to provide a wider basis for generalisations concerning the nature of the tasks, and insight into particular problems experienced in their performance, are analysed. In this way it is hoped to specify more clearly than before those language-related skills required for study purposes across a variety of disciplines.

CHAPTER TWO

PRINCIPLES AND APPROACHES IN LANGUAGE TESTING

CHAPTER TWO CONTENTS

	<u>Page</u>
2. PRINCIPLES AND APPROACHES IN LANGUAGE TESTING	
2.1 BASIC CONSIDERATIONS IN TEST DESIGN	63
2.1.1 The Concept of Validity	63
2.1.1.1 Construct Validity	64
2.1.1.2 Content Validity	70
2.1.1.3 Face Validity	72
2.1.1.4 Criterion-Related Validity	73
2.1.2 The Concept of Reliability	75
2.1.3 Test Efficiency	76
2.1.4 Validity and Reliability - An Inevitable Tension?	77
2.2 A CRITICAL REVIEW OF APPROACHES TO LANGUAGE TEST DESIGN	80
2.2.1 Introduction	80
2.2.2 The Psychometric-Structuralist Era	81
2.2.3 The Psycholinguistic-Sociolinguistic Era	83
2.2.4 The Communicative Paradigm	88
2.2.5 The Problem of Extrapolation	92

2. PRINCIPLES AND APPROACHES IN LANGUAGE TESTING

2.1 BASIC CONSIDERATIONS IN TEST DESIGN

The concepts of validity, reliability and efficiency, affect all aspects of test design, whatever the linguistic paradigm influencing our approach. In Section 2.1 below we discuss the nature of the concepts and then in Section 2.2 we review the major approaches to language testing in the light of these.

In Section 2.1 we examine the status of the various types of validity and how the concept of validity relates to that of reliability. Given the restrictions on the time and resources available to us, we were constrained in this work to focus on the a priori validation of test tasks, paying particular attention to content validity in our attempts to improve the description of the domain of English for Academic Purposes from which we wished to sample. We were aware that linguistics had not yet given us a proven, theoretical basis on which to build definitions of what we were testing, but nevertheless we felt a need to address ourselves to the question of establishing construct validity for our tests. The growing interest in testing communicative ability is accordingly examined in Section 2.2 below. We recognised the importance of criterion-related validity and, as far as time and resources permitted, we sought also to establish this for our test at the performance stage, though within the limitations of the present project, this by necessity, received less attention than the a priori validation of the test.

2.1.1 The Concept of Validity

The validity of a test is usually defined as the extent to which it measures what it is supposed to measure (cf. Lado 1961; Pilliner 1968; Heaton 1975a; Ingram 1977; Kelly 1978 and Moller 1982). Anastasi (1982, p.131) described it as a question of what the test

measures and how well it does so. She pointed out that:

"... all procedures for determining test validity are concerned with the relationships between performance on the test and other independently observable facts about the behaviour characteristics under consideration."

The methods employed for establishing validity can be looked at under the categories of construct, content, face and criterion-related validity.

2.1.1.1 Construct Validity

Anastasi (1982, p.153) outlined the relationship between the various types of validity:

"... content, criterion-related and construct validation do not correspond to distinct or logically co-ordinate categories. On the contrary, construct validity is a comprehensive concept which includes the other types."

Kelly (1978) and Hawkey (1982) similarly regarded construct validity as a superordinate concept embracing all other forms of validity. Kelly (1978, p.2) saw it as "the extent to which the test measures what it is supposed to measure." This echoed Cronbach's (1971, p.463) comment that:

"Every time an educator asks 'But what does the instrument really measure?' he is calling for information on construct validity."

More specifically, Anastasi (1982, p.144) defined it as:

"... the extent to which the test may be said to measure a theoretical construct or trait... Each construct is developed to explain and organise observed response consistencies. It derives from established inter-relationships among behavioral measures ... Focusing on a broader, more enduring, and more abstract kind of behavioral description ... construct validation requires the gradual accumulation of information from a variety of sources. Any data throwing light on the nature of the trait under consideration and the conditions affecting its development and manifestations are grist for this validity mill."

She then described specific techniques that can contribute to construct validation, e.g. correlations with other tests, factor analysis, internal consistency measures, convergent and

discriminant validation, and argued that the theoretical construct, trait or behaviour domain measured by any test can be defined in terms of the operations performed in establishing the validity of the test.

Hawkey (1982), in a review of the literature on validity, agreed with Davies (1965) and Stevenson (1974) that very little attention had been accorded to construct validity in the language testing literature. He argued (p.124) that in what Spolsky (1976) described as the 'psychometric-structuralist era' of language testing, a prevalent view was that:

"New language is to be learnt as a stimulus-response habit formation process where discrete elements of the target language, identified as different from those in L1 and, thus, more difficult, are 'drilled in' until L1 habits no longer 'interfere'. The phonological, morphological, syntactic and lexical components of language are isolable, as are the four skills of listening, speaking, reading and writing. Here was a model nicely susceptible to discrete-item testing and thus to the kind of statistical analysis required by the psychometrists."

Because of confidence in this prevailing theoretical paradigm and the fact that it lent itself easily to testing, there was according to Hawkey, little need for much prior deliberation about the match between theory and test. Kelly (1978, p.2) developed a related argument:

"... the empiricism and operationalism of scientists in general in the first half of this century (cf. Lyons, 1977:120ff) and of those working on psychological and educational measurement in particular made the idea of working with such a concept as 'construct validity' unattractive ... As a result of such attitudes, other notions of validity more consistent with the principles of operationalism were substituted, in particular the notions of concurrent and predictive validity."

Construct validity is viewed from a narrow perspective in much of the current American literature (cf. Palmer et al. 1981a; Bachman et al. 1981a). It is seen principally as a matter of the a posteriori statistical validation of whether a test has measured a construct which has a reality independent of other constructs. The concern is much more with the a posteriori relationship between a test and the

psychological abilities, traits, constructs, it has measured than with what it is that should be elicited in the first place.

Others (cf. Davies 1977b; Kelly 1978; Morrow 1979; Alderson et al. 1981; Hawkey 1982) have shown a greater concern for test validation at the a priori stage. For Cronbach (1971, p.443):

"Construction of a test itself starts from a theory about behaviour or mental organisation derived from prior research that suggests the ground plan for the test."

Davies (1977b, p.63) argued in a similar vein:

"... it is, after all, the theory on which all else rests; it is from there that the construct is set up and it is on the construct that validity, of the content and predictive kinds, is based"

and Kelly (1978, p.8) supported this view, commenting that:

"... the systematic development of tests requires some theory, even an informal, inexplicit one, to guide the initial selection of item content and the division of the domain of interest into appropriate sub-areas."

Morrow (1979) also put the case strongly for trying to ensure that we consider more rigorously whether what we are testing is what we think we are testing, and whether what we think we are testing is what we ought to be testing. He argued that whatever the approach testers adopt, they need to be far more explicit about what it is that they are testing. It would seem to follow that the more fully we are able to describe the construct we are attempting to measure at the a priori stage the more meaningful are the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test.

Because we lack an adequate theory of language in use, attempts to determine construct validity at an a priori as against an a posteriori stage seem to involve us, to a great extent, in matters which relate more evidently to content validity. We need to talk of the communicative construct in pre-theoretical descriptive terms, and as a result, we become involved in questions of content relevance and content coverage. Thus for Kelly (1978, p.8) content validity seemed

"an almost completely overlapping concept" with construct validity, and for Moller (1982, p.68):

"... the distinction between construct and content validity in language testing is not always very marked, particularly for tests of general language proficiency."

The project we report in this thesis is essentially concerned with the research, specification and development of an experimental English for Academic Purposes (E.A.P.) test battery. A greater degree of explicitness at the a priori stage of test construction was felt to be necessary, if we were subsequently to make meaningful statements about a candidate's performance, which would be of use to those providing remedial support within receiving institutions. We would, however, regard this a priori validation as essentially a first, though crucial, step in the total validation process of an experimental test.

Davies (1983, p.1) argued forcefully that external validation based on data is always to be preferred:

"The external criterion, however hard to find and however difficult to operationalise and quantify remains the best evidence of a test's validity. All other evidence, including reliability and the internal validities is essentially circular"

and he quotes Anastasi on the need for independently gathered external data:

"Internal analysis of the test, through item-test correlations, factorial analysis of test items, etc. is never an adequate substitute for external validation."

However, since our primary concern was to collect appropriate information on a candidate's performance for the purposes of profile reporting, we were more obliged than those whose major interest was in predictive validity also to establish content/construct validity for our test by identifying, prior to test construction, appropriate communicative tasks which it should include.

Having made rigorous attempts at an a priori stage to make the test as valid as possible, we would then seek to establish the validity of T.E.A.P. against external criteria. If we by-passed the first stage

with its emphasis on content validity then the type of test we had available for external validation procedures would not, in all likelihood, have suited the purpose for which our test was intended.

A pencil and paper test of phoneme discrimination, for instance, might be found to correlate highly with an external criterion, e.g., another established test concurrently administered or a measure taken at a later date, such as final academic grades. It would however, be of less value to those providing remedial English language support, who, rather than a single score, require information about the particular study modes in which a student has difficulty operating. One would not be able to allocate students effectively to remedial language classes on the basis of performance in this type of test.

It is also salutary to point out that most G.C.E. examinations and existing language proficiency examinations, e.g. the C.P.E. and the J.M.B. Test in English (Overseas), because of their public, operational nature, are not overly interested in concurrent or predictive validity whereas, as Davies (1982) points out, these are matters of major concern for most standardised, closed E.F.L. tests. Correlating the results of one year's examination with other examinations or against some future criterion is perhaps viewed as a fruitless exercise when a new set of examinations is already in preparation for the following year and the results already issued for current candidates. Only experimental tests such as T.E.A.P. or closed tests such as E.L.T.S. or E.L.B.A. feel obliged to concern themselves with a posteriori validation procedures. Open examinations which are held annually tend to rely more heavily on construct, content and face validity. The major concern of this thesis is to explore ways of establishing content, construct and face validity for an E.A.P. proficiency test at the test construction stage. Attention is paid to a posteriori validation techniques such as internal consistency measures, factor analysis and, where available, inter-test correlations, though time and resources available have constrained how much of this we are able to do.

The a priori validation of an E.A.P. proficiency test would seem to demand that we test integrated macro-skills rather than micro-elements

in isolation. Given that our aim is to test the communicative competence of overseas students in an E.A.P. setting, it is doubtful whether tests of linguistic competence alone are appropriate because the constructs for such tests are necessarily based on discrete linguistic levels, not on integrative work samples. Since the essence of communication is an ability to combine discrete linguistic elements in a particular context, it seems essential that this ability should be assessed by tests of integrated skills, rather than by tests of discrete linguistic levels in isolation.

The content of an E.A.P. proficiency test based on work samples from the target situation, is qualitatively different from the content of a test of linguistic competence based upon discrete linguistic items. In the case of the E.A.P. proficiency test which aims at assessing communicative competence, the main justification for item selection is a careful sampling of the communicative tasks required of students in English medium study. In the case of a test of linguistic competence, a test may be considered valid if its content is based on an adequate sample of 'typical' discrete linguistic elements.

Davies (1965, p.49), commenting on the criteria adopted for selecting items for his English Proficiency Test Battery, argued that a proficiency test battery should have both a work sample justification and a linguistic justification:

"The linguistic demands of a Proficiency Battery are that it should be based on some language (or language testing) theory, that it should sample each of the major linguistic categories, that it should look at these features both in isolation and combination; it should consider the importance of these features in the communication needs of the candidates who will take the test... it should attempt to test control over the language in action rather than the language itself or knowledge about the language."

It is interesting to speculate how far language testers can escape from the influence of the prevailing language paradigm. Davies (1965, p.52), working within the psychometric-structuralist paradigm, argued that the proficiency tester:

"... starts off from the theory that language can (or should) be analysed into linguistic parts, into language levels..."

but despite this, he attempted to investigate the feasibility of including job sample tests in his battery: what he termed the 'performance approach'.

The proficiency tester today is more influenced by what Moller (1981b) has described as the sociolinguistic-communicative paradigm. According to Canale et al. (1980, p.34) communicative testing:

"... must be devoted not only to what the learner knows about the second language and about how to use it (competence) but also to what extent the learner is able to actually demonstrate this knowledge in a meaningful communicative situation."

Thus a good deal more attention has now to be paid to content and face validity than was the case under the previous orthodoxy. We feel however, that it is a matter for empirical investigation rather than conjecture, whether in an E.A.P. proficiency test one can by-pass assessing linguistic competence through discrete items based on linguistic categories, in favour of assessing communicative competence using integrative testing techniques based on a work sample justification. The experimental version of T.E.A.P. primarily had such a work sample justification but in addition, we felt it necessary to include components which sampled major linguistic categories; for, as Moller (1981b, p.44) argued:

"It is clear that communicative testing does test certain aspects of proficiency. But it is important to be aware that testing language proficiency does not amount just to communicative testing. Communicative language performance is clearly an element in, or a dimension of, language proficiency. But language competence is also an important dimension of language proficiency and cannot be ignored. It will also have to be tested in one or more of the many ways that have been researched during the past 30 years. Ignoring this dimension is as serious an omission as ignoring the re-awakening of traditional language testing in a communicative setting."

2.1.1.2 Content Validity

Inevitably we have commented on content validity in addressing ourselves to the concept of construct validity. We have pointed out that a primary purpose of our test is to provide receiving institutions with a

profile of the student's E.A.P. proficiency, indicating in broad terms the particular study modes where deficiencies lie. Content validity is considered important for this purpose as it is principally concerned with the extent to which the selection of test tasks is representative of the larger universe of tasks of which the test is assumed to be a sample (v. Bachman et al. 1981a).

Similarly, Anastasi, (1982, p.131) defined content validity as involving:

"... essentially the systematic examination of the test content to determine whether it covers a representative sample of the behavior domain to be measured."

She (p. 132) made the following important points:

1. "the behavior domain to be tested must be systematically analysed to make certain that all major aspects are covered by the test items, and in the correct proportions";
2. "the domain under consideration should be fully described in advance, rather than being defined after the test has been prepared";
3. "content validity depends on the relevance of the individual's test responses to the behavior area under consideration, rather than on the apparent relevance of item content."

The directness of fit and adequacy of the test sample is thus dependent on the quality of the description of the target language behaviour being tested.

J.B. Carroll (1961) pointed to the importance of, and the difficulties involved in, defining the area of language from which the sample is to be taken and the resultant problems this has for sampling. Moller (1982, p.37) also referred to the problems involved:

"In the case of a proficiency test, however, the test constructors themselves decide the 'syllabus' and the universe of discourse to be sampled. The sampling becomes less satisfactory because of the extent and indeterminate nature of that universe."

To the extent that the content is made explicit the concern also

becomes one of face validity (v. Porter 1983) - perhaps the most contentious validity we might invoke.

2.1.1.3 Face Validity

Anastasi (1982, p.136) pointed out that face validity:

"... is not validity in the technical sense; it refers, not to what the test actually measures, but to what it appears superficially to measure. Face validity pertains to whether the test 'looks valid' to the examinees who take it, the administrative personnel who decide on its use, and other technically untrained observers. Fundamentally, the question of face validity concerns rapport and public relations."

Lado (1961), Davies (1965), Ingram (1977), Palmer (1981) and Bachman et al. (1981a) discounted the value of face validity. Bachman et al. argued, (p. 55):

"Since there is no generally accepted procedure for determining whether or not a test demonstrates this characteristic, and since 'it is not an acceptable basis for interpretative inferences from test scores', we feel it has no place in the discussion of test validity."

If the test does not have face validity though, it may not be acceptable to the students taking it, or the teachers and receiving institutions who may make use of it. If the students do not accept it as valid, their adverse reaction to it may mean that they do not perform in a way which truly reflects their ability. Anastasi (1982, p. 136) took a similar line:

"Certainly if test content appears irrelevant, inappropriate, silly or childish, the result will be poor cooperation, regardless of the actual validity of the test. Especially in adult testing, it is not sufficient for a test to be objectively valid. It also needs face validity to function effectively in practical situations."

Though she added the caution (p.136):

"To be sure, face validity should never be regarded as a substitute for objectively determined validity... The validity of the test in its final form should always be directly checked."

In all forms of validity referred to so far, knowing what the test is measuring is crucial. There is a further type of validity which we might term criterion-related validity where knowing exactly what a test measures is not so crucial.

2.1.1.4 Criterion-Related Validity

This is a predominantly quantitative and a posteriori concept concerned with the extent to which test scores correlate with a suitable external criterion of performance; what Ingram (1977, p.18) termed 'pragmatic validity'. Criterion-related validity divides into two types (v. Davies, 1977b), concurrent validity, where the test scores are correlated with another measure of performance, usually an older, established test, taken at the same time (cf. Kelly 1978; Davies 1983) and predictive validity, where test scores are correlated with some future criterion of performance (v. Bachman et al. 1981a.)

Davies (1965, pp. 149-50) commented that predictive validity is:

"... established by the prognostic success of a test, its later confirmation of the expectations implied in the test's results. To estimate predictive validity follow-up studies are necessary."

and went on to say that this and concurrent validity (v. Davies 1983), i.e., external validation based on data, are always preferable to the 'armchair speculation of content validity.'

Though this concept of validity is more in keeping with the demands of an empiricist-operationalist approach (v. Kelly 1978): the problem remains that a test can be valid in this way without our necessarily knowing what the test is measuring. Morrow (1979, p.147) drew attention to the essential circularity of employing these types of validity in support of a test:

"Starting from a certain set of assumptions about the nature of language and language learning will lead to language tests which are perfectly valid in terms of these assumptions but whose value must inevitably be called into question if the basic assumptions themselves are challenged."

Moller (1982) also referred to the problems in establishing sufficiently valid criterion measures against which to correlate. As Jakobovits

(1970, p.75) had pointed out:

"the question of what it is to know a language is not well understood and, consequently, the language proficiency tests now available and universally used are inadequate because they attempt to measure something that has not been well defined."

There is a danger in a validation study of this type that one might be forced to place one's faith in a criterion measure which may in itself not be a valid measure of the construct in question. One cannot claim that a test has criterion-related validity because it correlates highly with another test, if the other test itself does not measure the criterion in question.

Incidentally, validity and reliability estimates based on correlational data must be treated with caution. A high correlation may indicate the measurement of two different attributes which are themselves quite highly correlated amongst the population of testees. On the other hand, a low correlation may indicate that two quite different attributes are being measured or may merely reflect a high level of error variance in one or both of the tests.

Morrow argued (1979, p.147):

"Validity exists only in terms of specified criteria and, if the criteria turn out to be the wrong ones, then validity claimed in terms of them turns out to be spurious"

and Hawkey (1982, p.153) echoed this:

"Although this procedure can reveal signs of apparent unreliability in a pilot test, such evidence with its inherent risk of circularity, would require very close scrutiny indeed before it led to any major change in evaluation approaches. At this developmental stage in communicative testing, other tests available as criteria for concurrent validation are likely to be less integrative/communicative in construct and format, and thus not valid as references for direct comparison."

Though we appreciated the need for caution in our interpretation of these criterion-related validity measures, we still considered them to be useful concepts. For example, one might be very wary of tests that produced results seriously at variance with those of other tests measuring the same trait, especially if the latter had been found to

have true construct validity.

Similarly, in the case of predictive validity, it may be that in certain circumstances the predictive power of the test is all that is of interest. If all one wants is to make certain predictions about future performance on the basis of the test results, this might entail a radically different test from that where the interest is in providing information to allow effective remedial action to be taken. If predictions made on the basis of the test are reasonably accurate then the nature of the test items and their content might not be important.

One of the main aims of our Test in English for Academic Purposes (T.E.A.P.) is to provide receiving institutions with a profile indicating those language areas in which overseas students coming to take English medium academic courses, would seem to be at a linguistic disadvantage as compared to their native speaker counterparts. Thus content, construct and face validity were considered important for our purpose, though we were also aware of the need to establish concurrent and predictive validity.

2.1.2 The Concept of Reliability

Another criterion against which any language test has to be judged is its reliability (v. Anastasi 1982). This concept is particularly important when considering language tests within the communicative paradigm. For as Davies (1965, p.14) pointed out:

"... reliability is the first essential for any test; but for certain kinds of language test may be very difficult to achieve."

Guilford (1965) identified a number of aspects of test reliability but, in the work being reported here, the major concern was with only three of these. We were concerned with the reliability between different markers when marking a test of written expression, and investigated how the kind of written production called for in the T.E.A.P. could be most reliably marked using trained markers. It was also necessary to ensure that relevant sub-tests were internally consistent in the sense that all items in a sub-test were judged to be

measuring the same attribute. The third aspect was parallel forms reliability, the requirements of which will have to be borne in mind when future alternative forms of the T.E.A.P. have to be devised.

Because of the constraints on us and because of the many problems involved (cf. Perren 1968; Carroll, J.B. 1973; Beardsmore 1974; Clark 1975; Fisher et al. 1979; Underhill 1981) it was not possible to investigate the development of a reliable test of oral production in this project. This is not to say an oral test cannot be devised which is both valid and reliable, though the time and resources required for administration and in the prior training of examiners are quite considerable. In our analysis of the language tasks facing students in the academic context and the difficulties met in coping with them, we have investigated spoken medium tasks, as it is hoped that this may at least encourage future research and development into the provision of a valid and reliable oral component (v. James forthcoming). In this work though, we shall be confining ourselves to examining how reliable we can make the reading comprehension, listening comprehension, and writing sub-tests of our proposed Test in English for Academic Purposes.

2.1.3 Test Efficiency

A valid test is of little use if it does not prove to be a practicable one. This involves questions of economy, ease of administration, scoring and interpretation of results. The longer it takes to construct, administer and score, and the more skilled personnel and equipment that are involved, the higher the costs are likely to be.

The duration of the test may affect its successful operation in other ways, e.g. a fatigue effect on the candidates, administrative factors such as staff to invigilate, and the availability of rooms in which to sit the examination; all have to be taken into consideration. It is thus highly desirable to make the test as short as possible, consistent with the need to meet the validity and reliability criteria referred to above. If the aim is to provide as full a profile of the student's abilities as is possible then there is obviously a danger of conflict

here, for although hard pressed administrators seem in our experience to want a single overall grade, remedial language teachers would prefer as full a profile as possible. (cf. Moller 1977; Alderson et al. 1981).

2.1.4 Validity and Reliability - An Inevitable Tension?

Within the limitations of this work the concern was primarily with validation at the test construction stage, and only to a lesser extent with a posteriori validation at the performance stage. The resources to do thorough, professional, concurrent and predictive validity studies, such as conducted by Moller (1982) and those at present being conducted on the British Council's ELTS battery by the Institute of Applied Language Studies, at the University of Edinburgh, were not available to us, though it is envisaged that such studies will be done on T.E.A.P. at the Universities of Reading, Lancaster and Southampton over the next few years. Our concerns were of necessity with content, construct and face validity though the predictive and concurrent validity of our tests was also examined as far as circumstances allowed. We sought to examine how far the 'communicative' and 'E.S.P.' paradigms were applicable at an a priori stage to the research and development of a valid language proficiency test for students attempting to follow academic courses through the medium of English.

This might have proven to be a sterile endeavour unless we could also ensure the reliability of our tests. While one can have test reliability without test validity, a test can only be valid if it is also reliable. (v. Kelly 1978)

There is sometimes said to be a reliability-validity tension, (cf. Guilford 1965; and Davies 1978). This tension exists in the sense that it is sometimes essential to sacrifice a degree of reliability in order to enhance validity. If, however, validity is lost to increase reliability we finish up with a test which is a reliable measure of something other than what we wish to measure. The two concepts are, in certain circumstances, mutually exclusive but if a choice has to be made, validity "after all, is the more important". (v. Guilford 1965, p.481).

Validity is important also because it is related to the way in which test performance levels are defined. Houston (1983) describes the difference between norm and criterion referenced methods of doing so and discusses some of the difficulties of specifying appropriate performance criteria when the latter method is chosen. Popham (1978, p.2) provided the following functional definitions of these approaches:

"... a criterion-referenced test is designed to produce a clear description of what an examinee's performance on the test actually means. Rather than interpreting an examinee's test performance in relationship to the performance of others as is the case with many traditional tests, a good criterion-referenced test yields a better picture of just what it is that the examinee can or cannot do."

Davies (1978, p.158) made the connection with language testing and expressed certain reservations about criterion referenced tests:

"... there are difficulties in using criterion referenced tests for language: there is no finite inventory of learning points or items; there are very many behavioural objectives; there are variable (or no) external criteria of success, fluency, intelligibility, etc; there is no obvious way of establishing adequate knowledge, of saying how much of a language is enough."

thus putting in a language testing context some of the difficulties referred to later by Houston (1983). Clearly, criterion-referencing of performance levels is possible only to the extent that the test has a high degree of content validity.

Rea (1978) argued that simply because tests which assess language as communication cannot automatically claim high standards of reliability in the same way that discrete item tests are able to, this should not be accepted as a justification for continued reliance on highly reliable measures having very suspect validity. Rather, we should first be attempting to obtain more reliable measures of communicative abilities. This seems a less extreme and more sensible position than that adopted by Morrow (1979, p. 151), who argued polemically:

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

Rea's viewpoint was shared by Read (1981a, pp. X-XI), who reported that a recurring theme at the April 1980 R.E.L.C. Seminar on 'Evaluation and

Measurement of Language Competence and Performance' was that:

"... subjective judgements are indispensable if we are to develop testing procedures that validly reflect our current understanding of the nature of language proficiency and our contemporary goals in language teaching."

Read went on to emphasise that:

"... this does not mean a return to the old pre-scientific approach. It is generally accepted that a substantial, verifiable level of reliability must also be attained, if test results are to have any meaning."

Moller adopted a similar approach (1981a, p.67):

"While it is understood that a valid test must be reliable, it would seem that in such a highly complex and personal behaviour as using a language other than one's mother tongue, validity could be claimed for measures that might have a lower than normally acceptable level of reliability."

He argued that, although reliability is something we should always try to achieve in our tests "it may not always be the prime consideration" and offers a possible compromise position, (p.67):

"In constructing test batteries that contain different types of task, for example, certain of the sub-tests may be required to exhibit a high degree of reliability. Other sub-tests, particularly tests of communicative use, may quite properly exhibit lower reliability without adversely affecting the overall validity of the battery."

Hawkey (1982, p.149) commented in a similar vein:

"... the reliability of a test cannot be ignored without a harmful effect on the validity of the instrument. But it is likely that, if the construct validity of communicative tests is to be ensured, the reliability question is going to have to be accepted as subordinate, though worked at fairly hard by item analysis and correlational operations."

In our Test in English for Academic Purposes we attempted to develop test formats and evaluation criteria that would provide the best overall balance between reliability, validity and efficiency in the assessment of communicative skills.

2.2 A CRITICAL REVIEW OF APPROACHES TO LANGUAGE TEST DESIGN

2.2.1 Introduction

To inform us in making decisions on the best formats for our test battery we examined critically the alternative approaches to language testing and their limitations in terms of our stated criteria of validity, reliability and efficiency.

Davies (1978) argued that by the mid '70's, approaches to testing seemed to fall along a continuum which stretched from 'discrete' item tests at one end, to integrative tests such as cloze at the other. He took the view that in testing, as in teaching, there was a tension between the analytical on the one hand and the integrative on the other and considered that (p.149):

"... the most satisfactory view of language testing and the most useful kinds of language tests, are a combination of these two views, the analytical and the integrative."

He went on to say that it was probable in any case, that no test could be wholly analytical or integrative. For Davies (p.149):

"The two poles of analysis and integration are similar to (and may be closely related to) the concepts of reliability and validity. Test reliability is increased by adding to the stock of discrete items in a test; the smaller the bits and the more of these there are, the higher the potential reliability. Validity, however, is increased by making the test truer to life, in this case more like language in use."

Oller (1979) on the other hand, felt that testing should focus on the integrative end of the continuum. He made a strong case for following the swing of the testing pendulum away from what Spolsky (1976) had described as the "psychometric-structuralist" era, the so called 'discrete point' approach to testing, to what he termed "the psycholinguistic-sociolinguistic era"; the age of the integrative test. In our examination of these approaches, for the sake of description we treat them as if they were 'distinct', or 'pure' types. We recognise that, in practice, most tests contain an element of each, either in the test format or the assessment procedures adopted but, whilst we agree with Davies (1978) that the distinction between the two is neither real nor absolute, we nevertheless feel that they can

be usefully examined in terms of the particular focus they represent.

2.2.2 The Psychometric-Structuralist Era

The clear advantages of testing 'discrete' linguistic points are that they yield data which are easily quantifiable, as well as allowing a wide coverage of items. Tests which focus on 'discrete' linguistic items are efficient and have the usual reliability of marking associated with objectively scored tests, but both the 'discrete-point' approach, and the various formats employed in it, suffer from the defects of the construct they seek to measure. The problem with this approach to the measurement of proficiency is that it depends on proficiency being neatly quantifiable in this fashion. Oller (1979, p.212) outlined the deficiencies in terms of the construct validity of a hypothetically pure form of this approach:

"Discrete-point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) separately with little or no attention to the way those elements interact in a larger context of communication. What makes it ineffective as a basis for teaching or testing languages is that crucial properties of language are lost when its elements are separated. The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the parts separately, the whole is greater than the sum of its parts... organisational constraints themselves become crucial properties of the system which simply cannot be found in the parts separately."

Oller is on fairly safe ground here as most people would probably agree that testing a candidate's linguistic competence is a necessary, but not sufficient, component of a test battery. In real-life for example, we actually require people taking a driving test to demonstrate that they can perform the task and we do not depend solely on a pencil and paper test that informs us about the extent of their knowledge concerning the principles of driving. Similarly, those who have to make assessments about a piece of music will make them on the piece as a whole, not on selected parts of it. Chaplen (1970, p.XXVIII) criticised isolated skills tests from this point of view, arguing that:

"It seems unlikely that measurements of the component skills most commonly isolated can provide either singly or in aggregate, a satisfactory measurement of the gestalt."

This is a view shared by Savignon (1972) who found that grammatical competence was not by itself a good predictor of communicative skills.

Kelly (1978) argued that if the goal of applied linguistics is seen as the applied analysis of meaning, e.g. the recognition of the context specific meaning of an utterance as distinct from its system giving meaning, then we as applied linguists should be more interested in the development and measurement of ability to take part in specified communicative performance, the production of and comprehension of coherent discourse, rather than in linguistic competence. This echoed Spolsky's (1968) earlier point that perhaps instead of attempting to establish a person's knowledge of a language in terms of a percentage mastery of grammar and lexis, we would be better employed in testing that person's ability to perform in a specified socio-linguistic setting. Rea (1978, p.51) has expressed a similar view:

"... although we would agree that language is a complex behaviour and that we would generally accept a definition of overall language proficiency as the ability to function in a natural language situation, we still insist or let others impose on us, testing measures which assess language as an abstract array of discrete items, to be manipulated only in a mechanistic way. Such tests yield artificial, sterile and irrelevant types of items which have no relationship to the use of language in real life situations."

Morrow (1979) argued that if we are to assess proficiency, i.e. potential success in the use of the language in some general sense, it would be more valuable to test for a knowledge of and an ability to apply the rules and processes, by which these discrete elements are synthesized into an infinite number of grammatical sentences and then selected as being appropriate for a particular context, rather than simply to test a knowledge of the elements alone. For Morrow (1979, p.145):

"... knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language."

This characterisation of proficiency as the ability to do something rather than just a knowledge about something can also be found in Davies (1965), Harris (1969) and Clark (1975).

2.2.3 The Psycholinguistic-Sociolinguistic Era

At the risk of generalisation we might say that, in response to a feeling that 'discrete-point' tests were insufficient indicators of language proficiency, the testing pendulum swung in favour of global tests in the 1970s, into what Spolsky (1976) termed the psycholinguistic-sociolinguistic era, an approach to measurement that was in many ways contrary to the allegedly atomistic assumptions of the 'discrete-point' tests (v. Davies 1978). It was claimed by Oller (1979) that global integrative tests such as cloze and dictation went beyond the measurement of a limited part of language competence achieved by 'discrete-point' tests with their bias towards testing the receptive skills; that such tests could measure the ability to integrate disparate language skills in ways which more closely approximated to the actual process of language use. For Oller (1979, p.37):

"The concept of an integrative test was born in contrast with the definition of a discrete point test. If discrete items take language skill apart, integrative tests put it back together. Whereas discrete items attempt to test knowledge of language one bit at a time, integrative tests attempt to assess a learner's capacity to use many bits all at the same time, and possibly while exercising several presumed components of a grammatical system, and perhaps more than one of the traditionally recognized skills or aspects of skills."

Read (1981a, p.X) succinctly described the psycholinguistic/sociolinguistic era:

"From a psycholinguistic perspective, language came to be seen as less of a well-defined taxonomic structure and more of a dynamic, creative, functional system. It was recognized that natural language contains a considerable amount of redundancy, so that it is difficult to show that any single linguistic unit is indispensable for communication... The sociolinguistic contribution centres on the concept of communicative competence, which represents a broadening of Chomsky's notion of competence to cover not only knowledge of rules for forming grammatical sentences but also rules for using those sentences appropriately with different contexts... Thus the psycholinguistic and sociolinguistic perspectives have enlarged the basis on which the validity of a test is to be judged. New criteria have become introduced that cannot be measured by the standard 'objective' methods.'

Oller maintained that provided linguistic tests such as cloze require "performance" under real life constraints, e.g. time, they are at least a guide to aptitude and potential for communication, even if they do not test communication itself. They are also practicable to administer, economical to set and mark, and have respectable reliability figures associated with them.

Work by Alderson (1978a) however, has raised serious questions about the validity of these integrative measures as testing devices. He demonstrated that there is no such animal as "the cloze test" and, even in using the same passage, results are affected by altering the point where the deletions are started from, or by using a different nth rate deletion. The evidence is similarly contradictory about the differing scoring methods to be adopted in marking a cloze procedure and it has even been suggested that a cloze test is a much less effective measure for assessing "general proficiency", in that it correlates less well with other established general proficiency measures, when used on monolingual as against multilingual groups (v. Klein-Braley 1981). Perhaps more crucial than any of these reservations is the question of what performance on a cloze test really tells us about a candidate's language ability.

A major cause of concern is the assumption made by Oller (1976, 1979, 1980) that General Language Proficiency (GLP), the grammar of expectancy his integrative tests are tapping into, is a single principal factor underlying all language skills. His concept of 'overall proficiency' has inevitably merged into a hypothesis of an underlying unitary competence. This is a view implicit in his concept of the internalised expectancy grammar and, though it is one which is seductive for the purpose of those having to take administrative decisions, as Davies (1981b) points out, it conflicts with substantial evidence in favour of at least two competencies, namely reception and production (v. Vollmer 1981a). The differences between knowing how to analyse input and knowing how to construct output would seem to more than outweigh the correspondences between the two processes. Pedagogical experience would also suggest that the different performance tasks an individual is faced with, result in a variety of different proficiencies being

exhibited in the completion of these tasks.

Davies (1981b) emphasised that although Oller claims that his integrative tests represent total language proficiency better than any other single test or combination of tests, this is not in itself, an argument in favour of the unitary competence hypothesis, as measures such as cloze and dictation are so integrative, that they contain most or all language abilities anyway.

High correlations between cloze and other measures may only reflect that they are measuring different skills which are highly correlated among individuals, however this does not mean that there will be no individuals whose performances in the various skills differ considerably.

A group of testees may have scores in two tests which correlate very highly, in the sense that both tests put the individuals in more or less the same rank order, but since correlational measures take little or no account of mean scores, the group's scores may be centred on very different means in the two tests, indicating quite different levels of performance overall. In other words, correlational data do not provide evidence about standards.

The empirical evidence that has been marshalled in favour of the "unitary competence hypothesis" is open to some doubt and there is a growing body of evidence favouring a divisibility hypothesis (cf. Vollmer 1979; Bachman et al. 1981a; Vollmer 1981a and b; Hughes 1981a).

Principal component analysis is often used to substantiate the "unitary competence hypothesis" but this method is essentially designed to simplify data, and would be expected to produce one factor from a battery of seemingly different language tests. More crucially, this general language proficiency factor does not necessarily explain all the variance in the results, and the percentage of variance explained differs from study to study (v. Vollmer 1981a). Because of the existence of factors, other than the principal component, which explain

reasonable proportions of the remaining variance, it is often possible by pursuing further factor analysis, for example Varimax rotation of the factor structure, to obtain a number of independent factors each of which makes a sizeable contribution to the total variance.

There is also evidence in the literature that the format of a task can unduly affect the performance of some candidates (v. Murphy 1978a, 1980). It is our contention that this makes it necessary to include a variety of test formats for assessing each construct rather than rely on a single overall measure, such as cloze. In this way, we might give candidates a better chance of demonstrating potentially differing abilities (v. Vollmer 1979, 1981a).

Though the tests Oller is advocating are global in the sense that they require testees to exhibit simultaneous control over different aspects of the language system, they are nevertheless indirect. Although the tests might integrate disparate language skills in ways which more closely approximate to actual language use, one would argue that their claim to the mantle of communicative validity remains suspect, as only direct tests which simulate as closely as possible relevant authentic communication tasks, can claim to mirror actual communicative interaction (cf. Kelly 1978; Morrow 1979). As Moller (1982, p.25) pointed out, they do not:

"... require subjects to perform tasks considered to be relevant in the light of their known future use of the language."

Advocates of communicative language testing would argue that Oller's view pays insufficient regard to the importance of the productive and receptive processing of discourse, arising out of the actual use of language in a social context with all the attendant performance constraints, e.g. the interaction based nature of discourse, unpredictability and behavioural outcomes (cf. Morrow 1979 and Moller 1981b).

Both Rea (1978) and Morrow (1979) have emphasised that though indirect measures of language abilities claim extremely high standards of reliability and concurrent validity as established by statistical techniques, their claim to other types of validity remains suspect.

Morrow (1979) cited as evidence for this the fact that neither cloze nor dictation offers the opportunity for spontaneous production by the candidate and the language norms which are followed are those of the examiner (or original author of the text), not of the student himself. Neither testing procedure offers the possibility for oral or non-controlled written production, and since the oral and written skills are generally held to be highly important, some means of assessing them reliably in communicative situations should be found. Although integrative measures appear to correlate highly with other similar measures of general language proficiency, there is some empirical evidence that cloze correlates only moderately with tests of written production (v. Weir et al. 1978) and with spoken production (v. Vollmer 1981a). Given that the tests concerned are reliable, this would suggest the possibility that proficiency in these areas cannot be adequately predicted by a test of overall proficiency.

Morrow also claimed both cloze and dictation are fundamentally suspect since they are tests of underlying ability (competence) rather than actual performance. In other words, they depend basically on a knowledge of the language system rather than the ability to operate this system in authentic settings. Carroll, B.J. (1980, p.9) reached the same conclusion:

"... this (cloze test) is still essentially usage based. The task does not represent genuine interactive communication and is, therefore, only an indirect index of potential efficiency in coping with day to day communicative tasks."

Even if it were decided that indirect tests such as cloze were valid in some sort of derived fashion, it still remains true that performing on a cloze test is not the same sort of activity as reading. The pedagogical consequences of including this type of test measure in a battery might be harmful if it results in candidates being taught specifically to handle indirect assessment tasks, in preference to teaching them to cope with more realistic tasks.

Kelly (1978, p.241) made the further point that some candidates may manage to succeed in the indirect task by training of a certain kind

and thus invalidate the test:

"... indirect tests are subject to attacks on their validity in those cases where it is possible to by-pass the ability in question and develop proficiency in the assessment task alone."

He also noted (op. cit. pp. 245-246) that:

"Analysis of a student's responses to an indirect test will not provide any relevant information as to the reasons for the student's difficulties in the authentic task, of which one assumes, the indirect test is a valid and reliable measure. By their very nature, indirect tests can provide evidence for level of achievement, but cannot diagnose specific areas of difficulty in relation to the authentic task."

Integrative tests such as cloze only tell us about a candidate's competence. They do not tell us anything directly about a student's performance ability, and their main value in their unmodified form, appears to lie in ascribing competence levels, rather than relating candidates' performance to any external criteria. They are perhaps only of limited use when the interest is in what it is that the individual student can or cannot do in terms of the various language tasks he is to be faced with in real life situations.

The deficiencies in the type of information the 'discrete point' approaches of the psychometric-structuralist era, and the more integrative approaches of the psycholinguistic-sociolinguistic era could provide, meant that we needed to investigate the 'communicative paradigm' to see whether this approach might be more suitable for our purposes.

2.2.4 The Communicative Paradigm

For communicative testers to lay claim to construct validity for their approach, explicitness is required both at the test design stage where one is concerned with the required result (v. Hawkey 1982) and at the evaluation stage where one is estimating the acquired result. It is not necessarily the case that communicative tests will look radically different from existing tests; but there may be strong pragmatic reasons for trying to demonstrate any differences in either the test content or the marking schemes to be applied. Whatever the

differences it is only through a greater degree of explicitness in terms of the specification for test task construction and the criteria to be adopted in assessment, that a progressive dialectic can ensue.

Canale et al. (1980) provided us with a useful starting point for a clarification of the terminology necessary for forming a more definite picture of the construct, communicative testing. These authors took communicative competence to include grammatical competence (knowledge of the rules of grammar), sociolinguistic competence (knowledge of the rules of use and rules of discourse) and strategic competence (knowledge of verbal and non-verbal communication strategies). Whereas for Hymes (1972), communicative competence included the ability to use the language, as well as having the knowledge which underlies actual performance, Morrow (1979) felt a distinction needed to be made between communicative competence and communicative performance, the distinguishing feature of the latter being the fact that performance is the realisation of Canale et al.'s (1980, p.6) three competences and their interaction:

"... in the actual production and comprehension of utterances (under general psychological constraints that are unique to performance)"

Morrow (1979) and Canale et al. (1980) argued that communicative language testing as well as being concerned with what the learner knows about the form of the language and about how to use it appropriately in contexts of use (COMPETENCE), must also deal with the extent to which the learner is actually able to demonstrate this knowledge in a meaningful communicative situation (PERFORMANCE) i.e. what he can do with the language, or as Rea (1978, p.4) put it:

"... his ability to communicate with ease and effect in specified sociolinguistic settings."

It is held that the performance tasks candidates are faced with in communicative tests, should be representative of the type of task they might encounter in their own real-life situation and should correspond to normal language use where an integration of communicative skills is required with little time to reflect on, or monitor language input and output.

An idea of recent thinking may be gained by looking at the work of testers generally supportive of this broader sociolinguistic model of communication, where there is a marked shift in emphasis from the linguistic to the communicative dimension. The emphasis is no longer on linguistic accuracy, but on the ability to function effectively through language in particular contexts of situation. Cooper's (1968) view that existing test frameworks, because they concentrated on linguistic competence, might fail to assess a person's communicative ability, was taken up by Morrow (1979, p.149) who argued that traditional tests did not give:

"... any convincing proof of the candidate's ability to actually use the language, to translate the competence (or lack of it) which he is demonstrating into actual performance 'in ordinary situations' i.e. actually using the language to read, write, speak or listen in ways and contexts which correspond to real life."

Carroll, B.J. (1980, p.1) adopted a similar line:

"... the prime need of most learners is not for a theoretical or analytical knowledge of the target language, but for an ability to understand and be understood in that language within the context and constraints of particular language-using circumstances."

For him (op. cit., p.7):

"... the ultimate criterion of language mastery is therefore the learner's effectiveness in communication for the settings he finds himself in."

These statements reflect an emphasis in language teaching and, more recently, testing that has been placed on use and the concern that has been shown with communicative functions rather than with the formal language patterns of usage (cf. Campbell et al. 1970; Hymes 1972; Widdowson 1978).

Despite the value of this "use/usage" distinction it seems we still have a problem with terminology in communicative approaches to testing. References are frequently made in the literature to testing communicative performance, e.g. B.J. Carroll's book (1980) is entitled Testing Communicative Performance, and Morrow (1979) similarly ignored the illogicality of testing performance, when he used the sub-heading

"Performance Tests". Though we can talk of testing performance if the reference is to an individual's performance in one isolated situation, as soon as we wish to generalise about ability to handle other situations, competence would seem to be involved.

Strictly speaking, a performance test is one which samples behaviour in a single setting with no intention of generalising beyond that setting - any other test is bound to concern itself with competence. The very act of generalising beyond the setting actually tested, implies some statements about abilities to use the language and/or knowledge of it. It would be more accurate in discussing communicative language testing not to claim to be doing anything more than evaluating samples of performance, in certain specific contexts of use, created under particular test constraints for what they can tell us about a candidate's underlying communicative competence.

For Kelly (1978, p.350):

"To take part in a communicative event is to produce and/or comprehend discourse in the context of situation and under the performance conditions that obtain. It is the purpose of a proficiency test to assess whether or not candidates are indeed capable of participating in typical communication events from the specified communication situation(s)."

As a working definition we might accept that communicative performance relates to the transmission and reception of particular meanings in particular contexts, and what can be tested is the quality and effectiveness of the performance observed in these circumstances (v. Moller 1981b).

Though we accept that linguistic competence must be an essential part of communicative competence, the way in which they relate to each other, or either relates to communicative performance has in no sense been clearly established by empirical research. A good deal of work needs to be done in comparing results obtained from linguistically based tests with those which sample communicative performance, before one can make any positive statements about the former being a sufficient indication of likely ability in the latter or in real-life situations. No realistic comparisons are possible until reliable and effective, as

well as valid, methods for assessing proficiency in performing relevant communicative tasks are investigated. In this work we attempted to establish and construct such test tasks for inclusion in a trial battery, and compared them operationally with tasks which related more to the earlier 'discrete point' and integrative eras.

Before addressing ourselves more closely to the questions of 'what to test?' and 'how to test?' we need to consider briefly the problem of the generalisability of test results. This is an unavoidable issue, whatever approach to testing we adopt, and it is particularly germane to testing in the communicative paradigm, given the unlikelihood of our ever developing an adequate grammar of language in use from which to sample.

2.2.5 The Problem of Extrapolation

Other than serious marker reliability problems, associated with the assessment of performance, which we will deal with in Chapter 4 below, the main issue affecting our adoption of a 'communicative' approach to language testing was that of generalisability. Any test can be seen as a sampling instrument that provides evidence on which to base inferences that extend beyond the available data. For our purposes the evidence provided by test performances had to be evidence relevant to the whole domain of interest, that is, the test had to be valid; it had also to be capable of allowing stable predictions to be made about a candidate's performance in any part of the domain, in other words, the test had to be reliable.

The communicative type of E.A.P. test that seemed most suited to our purposes implied the specification of performance tasks closely related to the learner's practical activities, that is, to the communicative contexts of situations he would find himself in; therefore we were faced with the problem of the generalisability of the tasks we selected.

For Kelly (1978, p.225) the possibility of devising a construct valid proficiency test, i.e. one that measured ability to communicate in the target language, was dependent on the prior existence of "... appropriate objectives for the test to measure."

Advocates of performance based tests (cf. Morrow 1977, 1979; Carroll, B.J. 1978a, 1980) seem to be arguing that it is only necessary to select certain representative communication tasks, as we do not use the same language for all possible communication purposes. In the case of proficiency tests, these tasks are seen as inherent in the nature of the communication situation for which candidates are being assessed. Caution, however, demanded that we waited until empirical evidence was available before making such confident statements concerning the identification of these tasks, as only after examining the feasibility of establishing suitable objectives through empirical research based on real people coping with real situations, would we have any grounds for making claims that we had selected a representative sample of operational tasks to assess performance ability. Much of the work described in Chapter 3 was undertaken with this purpose in mind.

Even if it were possible to establish suitable objectives, viz., successfully to identify relevant communicative tasks and underlying constituent enabling skills for our target population, we would still face reliability and validity problems. If, as Rea (1978) and Morrow (1979) suggest, we seek to construct simulated communication tasks which closely resemble those a candidate would face in real life and which make realistic demands on him in terms of language performance behaviours, it might be difficult to do so reliably or validly. Communication is not coterminous with language and much communication is non-linguistic. Often the conditions for actual real-life communication are not replicable in a test situation, which is, by necessity, artificial and idealised and, to use Davies's (1978) phrase, Morrow is perhaps fruitlessly pursuing the chimera of authenticity. Further, even if our sample of communicative tasks possessed content and face validity might they not still lack generalisability in terms of the other communicative tasks we are not able to include? Are assessments of performance on these tasks made under particular linguistic and social constraints and thus not relatable to competence as 'characteristic abilities'? In other words, if we select, if we sample from a domain, how can we be sure that ours is an adequate sample?

Kelly (1978, p.226) observed that any kind of test is an exercise in sampling and from this sample an attempt is made to infer students' capabilities in relation to their performance in general.

"That is, of all that a student is expected to know and/or do as a result of his course of study (in an achievement test) or that the position requires (in the case of a proficiency test), a test measures students only on a selected sample. The reliability of a test in this conception is the extent to which the score on the test is a stable indication of candidates' ability in relation to the wider universe of knowledge, performances, etc., that are of interest."

He pointed out (p.230) that even if there is available a clear set of communication tasks:

"... the number of different communication problems a candidate will have to solve in the real world conditions is as great as the permutations and combinations produced by the values of the variables in the sorts of messages, contexts of situation and performance conditions that may be encountered."

Thus on the basis of performance on a particular item, one ought to be circumspect to say the least, in drawing conclusions about a candidate's ability to handle similar communication tasks.

Morrow (1977, p.53) was also aware of the problems of extrapolation. He succinctly set out the problem:

"The very essence of a communicative approach is to establish particular situations with particular features of context, etc., in order to test the candidate's ability to use language appropriate in terms of a particular specification. While it is hoped that the procedures discussed will indeed be revealing in those terms, they cannot strictly speaking reveal anything of the candidate's ability to produce language which is appropriate to a situation different in even one respect from that established."

Alderson (Alderson et al. 1981, p.59) also accepted that to follow the communicative paradigm one needed to define what it was that students had to do with language in a specific situation or series of situations, but recognised that by specifying performance in this manner:

"... one might end up describing an impossible variety of situations, which one cannot encompass for testing purposes."

In order to make stable predictions of student performance in relation to the indefinitely large universe of tasks, it would seem necessary to sample candidates' performances on as large a number of tasks as is possible, which conflicts immediately with the demands of test efficiency. The larger the sample of tasks and the more realistic the test items, the longer the communicative test will have to be.

However, as Alderson (ibid.) noted:

"... it may be that the issue of extrapolation is not (yet) of crucial importance: even if we cannot generalise from performance in one situation to performance in a variety of situations, if we can say something about performance in one situation, then we have made progress, and if we can say something important about performance in the target situation so much the better. Ultimately the student will have to perform, despite the statistical evidence of the relationship between predictor and predicted, or the theoretised relationship between competence and performance."

Morrow (1977) observed that in the case of conventional language tests aimed at measuring mastery of the language code, extrapolation would seem to pose few problems. The grammatical and phonological systems of a language are finite and manageable and the lexical resources can be delimited. The infinite number of sentences in a language is made up of a finite number of elements and tests of the mastery of these elements are extremely powerful from a predictive point of view. Thus, we might tend to agree with Davies (1978, p.225):

"What remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning... Grammar is far more powerful in terms of generalisability than any other language feature."

However, Kelly (1978) put forward an interesting argument against this viewpoint. It is not known, for example, how crucial a complete mastery of English verb morphology is to the overall objective of being able to communicate in English, or how serious a disability it is not to know the second conditional. According to Kelly (1978, p.17) we do

not possess a:

"... reliable knowledge of the relative functional importance of the various structures in a language."

Given this failing, it would seem ill advised to make any claims about what students should be able to do in a language on the basis of scores on discrete-point tests of syntax or lexis. The construct 'ability to communicate in language' involves more than a mere manipulation of certain syntactic patterns with a certain lexical content. In consequence, it would appear that we still need to attempt to devise measuring instruments which can assess performance ability.

As a way out of the extrapolation quandary, Kelly (1978, p.239) suggested a two-stage approach to the task of devising a test that represents a possible compromise between the conflicting demands of the criteria of validity, reliability and efficiency.

"The first stage involves the development of a direct test that is maximally valid and reliable, and hence inefficient. The second stage calls for the development of efficient, hence indirect, tests of high validity. The validity of the indirect tests is to be determined by reference to the first battery of direct tasks. Clearly, where valid and reliable but inefficient tests already exist for the construct in question, then the research strategy calls for the development of efficient, indirect tests whose results correlate highly with those of the existing test."

Thus, retreat from direct evaluation of performance may be acceptable, provided relationships or even correlations between data from competence testing and predicted behaviour have been established.

As far as large-scale proficiency testing is concerned, another viable solution might be to focus attention on language use in individual and specified situations while retaining, for purposes of extrapolation, tests of the candidate's ability to handle those aspects of language which are generalisable to all language use situations, namely the grammatical and phonological systems.

Morrow (1979, p.152) saw a third way out of the extrapolation quandary. His argument is that a model (as yet unrealised) for the performance of

global communicative tasks may show, for any task, the enabling skills which have to be mobilised to complete it:

"The status of these enabling skills *vis-à-vis* competence: performance is interesting. They may be identified by an analysis of performance in operational terms, and thus they are clearly, ultimately performance-based. But at the same time, their application extends far beyond any one particular instance of performance, and in this creativity they reflect an aspect of what is generally understood by competence. In this way they offer a possible approach to the problem of extrapolation."

He asserted that (p.153):

"An analysis of the global tasks, in terms of which the candidate is to be assessed, will usually yield a fairly consistent set of enabling skills"

and argues that assessment of ability in using these skills would therefore yield data which are relevant across a broad spectrum of global tasks, and are not limited to a single instance of performance.

For Morrow (1979, p.153), a working solution to the problem would be the development of tests which measure both overall performance in relation to a specified task, and the strategies and skills which have been used in achieving it:

"Written and spoken production can be assessed in terms of both these criteria. In task-based tests of listening and reading comprehension, however, it may be rather more difficult to see just how the global task has been completed. ... it is rather difficult to assess why a particular answer has been given and to deduce the skills and strategies employed. In such cases, questions focusing on specific enabling skills do seem to be called for in order to provide the basis for convincing extrapolation."

He is aware though, that there exists in tests of enabling skills a fundamental weakness in the relationship between the whole and the parts, as a candidate may prove quite capable of handling individual enabling skills yet still not be able to communicate effectively.

Another problem is that it is by no means easy to identify these enabling skills nor are there any guidelines for assessing their relative importance for the successful completion of a particular communicative task, let alone their relative weighting across a

spectrum of tasks. Morrow would appear to assume that we are not only able to establish these enabling skills, but also able to describe the relationship that exists between the part and the whole in a fairly accurate manner (in this case, how 'separate' enabling skills contribute to the communicative task). He would seem to assume that there is a prescribed formula; possession and ability to use enabling skills $X + Y + Z = \text{successful completion of communicative task (1)}$ whereas it would seem likely that the added presence of a further skill or the absence of a named skill might still result in successful completion of the task in hand.

Given that our aim is to predict a candidate's success in coping with real-life language activities we describe in Chapter 3 below the communicative tasks students meet whilst operating in an academic context and the enabling skills which appear to contribute to successful performance. On the basis of this analysis we constructed an experimental test battery to enable us to compare test tasks that related as far as possible to actual situations, i.e., direct tests, with various indirect measures that sampled performance on various linguistic tasks which had no counterpart in real-life terms. We are aware that it is not possible, under test conditions, to reproduce an exact copy of real-life performance but we believed that it might be possible to identify salient features as a yardstick for judging how near we had come in our efforts at direct testing or, as Kelly (p.234), described it, eliciting:

"... samples of authentic communicative performances from testees under maximally 'life-like' conditions."

The extrapolation problem faced by those adopting a more 'communicative' approach to language test design seems to relate to the wider issue of the status of laws in the behavioural sciences. In the physical sciences laws are extrapolations of replicable phenomena. Scientists in these domains can directly confront what it is that they wish to investigate, formulate hypotheses and repeat experiments as many times as they wish to check the falsifiability of their hypotheses. Because of problems associated with the infinite variability of language in use and the problems involved in population sampling, the scientific

paradigm is a difficult one to follow in educational measurement.

Hawkey (1982) described the classical scientific paradigm as a hypothetico-deductive methodology formulating quantifiable, narrow, parsimonious hypotheses, tested through the observation of the behaviour of a random sample of the target population, followed by a statistical analysis of the results according to pre-ordained procedures. This approach was not available to our particular research design; the construction of an English for Academic Purposes Test, which would show up the language deficiencies of overseas students operating in a variety of study modes, in a variety of disciplines at a variety of levels. In our efforts to establish a work sample justification for our test we had to take account of a large number of variables, some of which were not predictable, all interacting in socio-cultural contexts. Thus we were faced with a task sampling problem, a validity problem.

Unlike the scientists in the paradigm described by Hawkey we also faced serious problems in terms of population sampling. We had to be content with taking a sample from the population in which we were interested. This had powerful practical implications because our target population was transient, widely dispersed and varied in terms of accessibility. Most of the sampling, therefore, was by necessity opportunistic. This was a population sampling problem, a reliability problem.

In the task specification stage of test construction described in Chapter 3 below we had to content ourselves with what Hawkey (1982, p.16) described as an 'illuminative evaluation' paradigm, where the focus was on the description of complex phenomena, the resolution of significant features, and the comprehension of relationships. Ours was an empirical, fact finding study, rather than an attempt to provide a full explanatory model of communicative performance in an academic context. Our aim was to provide a descriptive framework in terms of which we might describe and analyse communication tasks of relevance to a broad spectrum of students in an E.A.P. context, prior to test construction. No claims are implied about how the language user

operates when involved in these communication tasks or how he learns to perform such tasks (v. Kelly 1978). The aim was simply to provide a specification, coarse but robust, of the general communicative tasks facing our target students in their academic context.

The research design was intended to result in a pilot test which, in addition to possessing the general properties of a good measurement instrument, would provide valid and reliable information on which to make inferences about candidates' ability to take part in communicative events of interest. While we were interested in the statistical procedures applicable to the results of the trial tests, we were primarily concerned with providing an a priori specification, upon which to base operational versions of T.E.A.P.

Our approach to data collection, for the establishing of our descriptive framework, is outlined in Chapter 3 below. It was integrative, quantitative research, being employed alongside qualitative approaches where the particular focus warranted it. The use of observation, interview, questionnaire and test data would, however, seem to employ the kind of mutually corroborative and refining variety of data collection approaches, that might be advocated by adherents at various positions along the research paradigm continuum (v. Hawkey 1982).

C H A P T E R T H R E E

IN PURSUIT OF A NEW PARADIGM

CHAPTER THREE CONTENTS

	<u>Page</u>
3. IN PURSUIT OF A NEW PARADIGM	
3.1 TOWARDS A FRAMEWORK FOR DESCRIPTION OF COMMUNICATIVE TASKS	105
3.1.1 The English for Specific Purposes (E.S.P.) Approach	105
3.1.2 Establishing the General Descriptive Parameters of Communication	114
3.1.3 Establishing the Dynamic Communicative Characteristics	116
3.1.4 Establishing the Test Task Dimensions	120
3.2 METHODS OF DATA COLLECTION: Filling Out the Framework	122
3.2.1 The Observations	122
3.2.1.1 Introduction	122
3.2.1.2 The Observation Method of Data Collection	125
3.2.1.2.1 Events and activities	127
3.2.1.2.2 Setting	130
3.2.1.2.3 Interactions	131
3.2.1.2.4 Instrumentality	131
3.2.1.2.5 Dialect	132
3.2.1.2.6 Target level dimensions	132
3.2.1.2.7 Target level tolerance conditions	136
3.2.1.2.8 Communicative key	137
3.2.1.2.9 Enabling skills	138
3.2.1.2.10 Conclusions	140
3.2.2 The Follow Up Interviews	140
3.2.3 The Questionnaire Method of Data Collection	142
3.3 THE PURPOSE OF LECTURES, SEMINARS/TUTORIALS AND PRACTICAL CLASSES	145
3.3.1 The Lecture	145
3.3.2 The Seminar and the Tutorial	148
3.3.3 The Practical Class	151
3.4 ACTIVITIES IN THE ACADEMIC CONTEXT	153
3.4.1 A Cautionary Note	153
3.4.2 Listening Comprehension Activities	156
3.4.2.1 The Nature of Listening Comprehension	156
3.4.2.2 The Listening Comprehension Tasks Students Perform in the Academic Context	160

	<u>Page</u>
3.4.2.2.1 Informing transactions	160
3.4.2.2.2 Directing transactions	164
3.4.2.2.3 Eliciting transactions	168
3.4.2.2.4 Formality	172
3.4.2.3 Problems in Listening Comprehension	176
3.4.2.3.1 General problems in listening comprehension	176
3.4.2.3.2 Specific problems in listening comprehension	180
3.4.2.4 Note Taking Activities	185
3.4.2.4.1 Methods of acquiring a written record	185
3.4.2.5 Problems in Note Taking	188
3.4.2.6 Constituent Enabling Skills	192
3.4.3 Reading Comprehension Activities	194
3.4.3.1 The Nature of Reading Comprehension	194
3.4.3.2 The Reading Tasks Students Perform in the Academic Context	198
3.4.3.3 Reading Difficulties Encountered by Students	207
3.4.3.4 Importance of Note Making	213
3.4.3.5 The Frequency with which Students Make Notes from Text-books	218
3.4.3.6 The Extent of Difficulty Encountered in Note-Making	220
3.4.3.7 The Specific Problems Encountered in Note-Making	220
3.4.3.8 Constituent Enabling Skills	221
3.4.4 Writing Activities in the Academic Context	226
3.4.4.1 The Nature of Written Production	226
3.4.4.2 The Writing Tasks Students Perform in the Academic Context	233
3.4.4.3 Problems Experienced in Written Production	242
3.4.4.4 The Standards Expected of British and Overseas Students	259
3.4.4.4.1 Variations in the allowances staff claim they make in marking the written work of overseas students	261
3.4.4.4.2 Variations in the allowances staff claim they make in assessing the written work of overseas students, as between coursework and examinations	267
3.4.4.5 Constituent Enabling Skills	269
3.4.5 Speaking Activities in the Academic Context	272
3.4.5.1 Introduction	272
3.4.5.2 The Speaking Tasks Students Perform in the Academic Context	272
3.4.5.2.1 Informing transactions	273
3.4.5.2.2 Students working together using English as a means of communication	277

	<u>Page</u>
3.4.5.2.3 Students actively taking part in discussions involving the teacher and the class as a whole	280
3.4.5.2.4 Eliciting transactions	281
3.4.5.2.5 The frequency with which the teacher asks the class questions	283
3.4.5.3 Problems Experienced in Spoken English	285
3.4.5.4 Constituent Enabling Skills	303
3.5 SETTING	305
3.6 INTERACTION	307
3.7 INSTRUMENTALITY	309
3.8 CONCLUSION	311

3. IN PURSUIT OF A NEW PARADIGM

3.1 TOWARDS A FRAMEWORK FOR DESCRIPTION OF COMMUNICATIVE TASKS

3.1.1 The English for Specific Purposes (E.S.P.) Approach

To help us establish the purposes for which our target group of students needed English, we drew upon recent developments in the English for Specific Purposes (E.S.P.) field as an informing source. Recent approaches to E.S.P. (cf. Jones 1974; Widdowson 1975; Jupp et al. 1975; Candlin et al. 1976; Lee 1976; Mackay et al. 1978; Munby 1978 and Candlin et al. 1978) no longer viewed particular areas of English, e.g. English for science, as a register or group of registers defined in terms of their formal linguistic properties. They differed from earlier approaches (cf. Close 1965; Ewer et al. 1969 and Swales 1971) in that they did not consider these areas as formally differentiated varieties of English usage, i.e. ways of manifesting the syntactic and lexical resources of the language. They were more concerned with predicting the communicative demands to be made on the learner in performing a role or series of related roles, rather than with the linguistic structures of the language per se. It was no longer a matter of simply identifying the appropriate register, but more a case of investigating the characteristic interactions in which language users were engaged, where they had actively to process discourse in the spoken or written medium and participate in real communicative behaviour. The objectives in an E.S.P. approach to testing and teaching are determined by a behavioural analysis of target situations and of the language used in the exercise of roles within those situations.

In Stage I of the project (described in Chapter 1 above) we established the levels, the discipline areas and the institutions in which overseas students were enrolling in the further and higher education sectors in the United Kingdom. On the basis of the information gathered during that stage, we decided to focus our research on students following courses in the general subject areas of science, engineering and social, business and administrative studies. In

Stage II the communicative demands that were made on students following courses in these general discipline areas were investigated.

A search of the literature revealed that there was no available profile of the representative tasks facing students in the various study modes, in the three broad discipline areas under review. It was perhaps surprising, given the number of overseas students there were in these disciplines, how little was known about the parameters of the communicative activities they were involved in. B.J. Carroll (1978a) in An English Language Testing Service Specifications was of little help here, as what he described was essentially pre-theoretical, backed by seemingly little or no empirical evidence. Carroll (p.6) defended the lack of real data on the language needs of students as follows:

"Although it would be desirable to derive our data from comprehensive observational studies of the participants actually engaged on their courses, we decided that less time-consuming methods would be sufficient to assess the basic adequacy of our approach to test specification."

He went on to point out (p.17) that this speculative approach unfortunately meant that:

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

and further:

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

He nevertheless argued (p.5):

"... in such a way ... we will be able to identify common and specific areas of need upon which an appropriately diversified test design can be based."

and claimed the test's "ultimate validation" would be evidenced by the "effectiveness of the tests based on their results". If the tests were to prove invalid, how is one to know whether it was the specification or poor test construction that was at fault? Even if the tests appear to possess predictive and concurrent validity, how certain can one be, given the nature of the original specifications of student needs, that this demonstrates a correct match between the

tests and the language demands made on real students by their courses of study?

The specifications for the English Language Testing Service Battery (E.L.T.S.) described by B.J. Carroll (1978a) are the only significant attempt so far made in the United Kingdom to carry out full scale modular E.S.P. testing and thus they serve to illustrate the practical problems facing those adopting an E.S.P. approach to testing.

Carroll, (p.2) covertly attacked the 'Davies Test' (E.P.T.B.), which the new E.L.T.S. battery was to replace, when he argued for:

"... ways of devising a more up-to-date system which will be able to cope with a problem the size and diversity of which the earlier system had not been designed to meet."

He claimed (p.3) that the new E.L.T.S. approach:

"... will cater more completely for the many different types of programme (of courses of study) we are testing for."

He stated that, in the past, the progress of some students had been adversely affected by their language inadequacy and, in certain cases, this had led to failure in their courses. No evidence however, is produced in support of this statement nor is any attempt made to specify the nature of these inadequacies. Further, he does not provide any evidence that the 'Davies Test' had failed to identify students with problems. Nor does it necessarily follow that, if it was deficient, what was needed to cope with the problem was a battery of modular subject specific tests as its replacement. It may be that what was required, in fact, was a better, more valid, single proficiency test.

It seems reasonable to hypothesise that different academic departments might place different communicative demands on overseas students and that an 'A' level science course may make different demands on a student from a post-graduate course in European politics. Without empirical evidence, however, we cannot infer the need for different tests for these groups as it could be that what are required are

different levels or different combinations of proficiencies in coping with the various study modes. A post-graduate engineering student, for example, might need a higher total score in some of the sub-tests but not need to do anything essentially different in terms of study mode operations from a G.C.E. 'A' level science student. The differences in the study mode activities he would need to engage in, might be of degree rather than kind.

If we accept for the moment that existing proficiency tests are inadequate and that it is possible to talk meaningfully about proficiency for a specific purpose, we must face the question of how specific this is to be. Carroll (1978a, p.4) claimed that E.L.T.S. catered for:

"... a process of diversification of test instruments to meet the diversity of the test situations."

However, this tells us only in an arbitrary way where the diversification stops.

Previous analyses of students' language for course design purposes have suggested that each course should be different with respect to the aims, interests, needs, wishes, etc. of the students taking it (cf. Crofts 1976 and Allwright et al. 1977). Real communication takes place in unique situations and, so it is argued, one cannot generalise. Is it possible, therefore, to draw up a meaningful profile for, say, all civil engineering students? Does the specific purposes approach not, in the end, lead us to a particular student on a particular course at a particular point in time; a position Carroll seems to have started from in his student profiles.

Carroll's six specifications, even if they had been based on rigorous empirical observation, might still be considered inadequate in that there are not enough of them. What he specified for a civil engineering student is not necessarily going to fit a chemical engineering student or even a civil engineering student at another university or polytechnic. It may be a vicious circle impossible to escape from in E.S.P. testing, that the more specific one tries to be, the greater the need for even more specificity.

If one was to offer a multiplicity of specific modules then one would also meet serious problems at the construction stage, even if the tests were then 'closed' to maintain security and the same versions used again. If the intention was to produce new versions every year, the problems of construction would be considerable and the parallel nature of the tests that much more difficult to ensure.

There would be further practical problems in constructing multiple specific tests in that it is sometimes difficult to match students to tests, e.g. whether post-graduate students in urban and regional studies, whose course will include law, economics and technology, should take the social science, the general academic or the technology module of the new E.L.T.S. battery.

Even if one was able to confine the investigation to a particular course at a particular institution a fully comprehensive description might still not be feasible. To illustrate this point we can take an example from the specific end of the needs analysis spectrum where an extremely thorough but limited approach, based on discourse analysis, is exemplified by the work of Candlin et al. (1976). Over a two-year period they managed to analyse closely the discourse of a small number of engineering lectures in one department in an attempt to (p.15):

"... consider not only cohesion in the analysis of texts, the description of 'linguistic elements at a suprasentential level' but also 'the communicative use to which utterances are put in the performance of social actions', textual coherence."

The approach to discourse advocated by these authorities and others such as Sinclair et al. (1975) was extremely elaborate and time-consuming. It had the advantage of systematically describing in very fine detail a very limited selection of discourse, but the extent to which one could extrapolate from these data was open to question. We needed to investigate the demands that were made on students in all study modes and examine what was common across institutions and levels, in the various sub-divisions of engineering, science, and social, business and administrative studies. The meticulous, rigorous method of enquiry advocated by the discourse analysts was beyond our scope. An adequate treatment would have

necessitated a large number of additional research projects and so we were forced to gather our information at this precise level of analysis from secondary informing sources such as Wijasuriya (1971), Straker-Cook (1975), Candlin et al. (1976), Montgomery (1977) and Mead (1980). Because of the enormity of the task of devising a multiplicity of specifications and then constructing a separate test for each, we were constrained to focus on those representative abilities students needed to be able to function effectively in the various study modes in which they had to operate.

T.E.A.P. was to be an E.S.P. test insofar as it dealt with that branch of E.S.P. known as English for Academic Purposes (E.A.P.). This focus allowed us to maintain a precarious foothold on the slippery slopes of E.S.P. and saved us from the otherwise uncertain descent into the over-specific. We felt that this was a tenable, albeit compromise position, between the extremes of the specific and the general; between a test of a knowledge of the finite grammar of the language at one end of the continuum and a more communicative, task-oriented test, designed for individuals, on one specific course of a particular academic programme, at the other. It was our aim to identify the general communicative tasks that students had to engage in whilst operating in the various study modes in the target learning situations. The information established concerning what was common in these study modes across academic disciplines and levels, helped inform the construction of the test battery described in Chapter 4 below.

This battery was intended to be a valid and reliable E.A.P. proficiency test which would provide information on the student's ability to operate in an E.A.P. context. On the basis of test results we would present institutions with a profile of the student's abilities in the broad E.A.P. skill areas, which could then be matched against the various communicative demands made on the student by a particular course of study. As well as being useful for administrative purposes in making decisions on acceptance or rejection, it was also intended to be useful pedagogically in providing information for any necessary remedial action to be taken.

More traditional proficiency tests like E.P.T.B. or E.L.B.A. were not primarily intended to yield information of this kind. Thus if they are employed as placement tests for remedial English classes the result is usually a heterogeneous mixture in terms of E.A.P. disabilities and needs. The aim of our tests was to identify the areas of students' language weaknesses relative to their communicative needs. In this way students with a common problem might be allocated to a remedial group where there was a common goal in overcoming this problem. Obviously the information available from our proficiency test would necessarily relate to proficiency in particular activities in certain study modes, but the information that a student was not able to cope with a limited set of realistic E.A.P. test tasks should prove useful.

Our first task was to establish what these E.A.P. activities were. The value of the needs analysis approach received support in the literature (cf. Jones et al. 1975, 1976; Gorosch 1976; Mackay et al. 1978; Webb 1977; Hughes et al. 1977; Allwright et al. 1977 and Munby 1978). Criticisms had been made, though, concerning the impracticality of comprehensively establishing language needs in this fashion. Porter (1983, p.192), in replying to these doubts, commented:

"It is not always possible to specify these needs with any accuracy, but it should prove possible to isolate a number of generally necessary or useful linguistic activities and by the same token to rule out a number of activities on the grounds of limited applicability..."

It seemed reasonable to hypothesise that there were essentially different types of language proficiency and that it would be inappropriate to assess language proficiency in types of communication for which the person being assessed had no need. Porter (1983, p.195) stressed in support of tests based on analyses of needs:

"It is simply not acceptable to argue that for many language learners needs are not predictable if language in use varies with the communicative activity and its purpose. Selections of activities must be made for teaching purposes, and it is inconceivable that activities with no foreseeable relevance to the learner will be selected."

Ignoring for the moment questions concerning the value of the communicative paradigm for language testing, it seemed necessary,

before we could make any statements about the relative merits of direct as against indirect forms of testing language proficiency (v. Kelly 1978), to develop valid and reliable, direct, communicative measures which effectively sampled the domain we were interested in. We thus started from the assumption that it was both desirable and feasible to evaluate samples of performance, in certain specific contexts of use, created under particular test constraints, for what they could tell us about a candidate's underlying competence. Having made this leap of faith it seemed expedient to attempt to develop a framework of categories for description of the type described by Hawkey (1982), which would help us to identify the activities our target group was involved in and to construct realistic and representative test tasks corresponding to these. In the following section we describe the general sets of evaluational requirements for the design and construct validation of test tasks which informed us in our data collection procedures.

In the research and development of T.E.A.P. we were fortunate in that we were able to build on the earlier work of Kelly (1978), Munby (1978) and Hawkey (1982). We drew upon their research in the construction of a framework of categories for the description of communicative test events: general descriptive parameters, dynamic communicative characteristics and task dimensions of target language behaviour. By applying these categories at the a priori test task validation stage we hoped to avoid some of the problems which had arisen in some earlier efforts at communicative testing where no attempt had been made to produce explicit specifications of the candidates' projected language needs in the target situation before test task construction took place. Though we would be cautious in claims for the directness of fit possible between test realisation and specification, we would argue that this approach enabled us to come closer to matching test tasks with appropriate activities in the target behaviour than would be possible using non-empirical approaches.

In order to pursue the communicative paradigm we felt that tasks should, as far as possible, be included in the testing operation with due regard to their directness of fit with criteria which accurately and adequately describe the significant aspects of the target

activities and the conditions under which they are normally performed. The concern was thus with content validity at the a priori stage as it no longer seemed sufficient to rely solely on more quantitative, post hoc validation procedures to establish what it was that we had tested (v. Chapter 2 above). Unless a communicative testing system was initially matched against such a framework, it was difficult to see how we could ever get near to describing accurately the construct that we were attempting to measure. The more fully that we could describe the construct through our concern with content validity at the a priori stage, the more meaningful were the validation procedures that could subsequently be applied to the results of the test(s).

What follows (v. Table 3A below) is a provisional attempt at such a framework of descriptive categories. It owed a lot to Munby (1978) in phase I, the General Descriptive Parameters of Communication and a lot to Morrow (1977, 1979) and Kelly (1978) in phase II, Dynamic Communicative Characteristics. For the most part though, the framework was derived from the work of Hawkey (1982) particularly in phase III, Task Dimensions.

TABLE 3A		
FRAMEWORK OF CATEGORIES FOR THE DESCRIPTION OF COMMUNICATIVE TEST EVENTS		
PHASE I	PHASE II	PHASE III
General Descriptive Parameters of Communication	Dynamic Communicative Characteristics	Task Dimensions
Activities	Realistic context	Size of text
Setting	Relevant information gap	Grammatical complexity and range of cohesion devices required
Interaction	Intersubjectivity	Functional range
Instrumentality	Scope for development of activity by participants	Referential range
Dialect	Allowance for self- monitoring by participants	
Enabling skills	Processing of appropriately sized input	
	Normal time constraints operative	

3.1.2 Establishing the General Descriptive Parameters of Communication

The parameters established by Munby (1978), as part of his processing "model" for syllabus definition, are useful to testers as a checklist against which they can evaluate the appropriacy of the performance based test tasks being developed. If the intention is to simulate in the testing situation those events and component activities students are faced with in the real world, then it is necessary to have a systematic basis for describing these. If a set of general descriptive parameters applicable to events in the target situation are

established these can then be used to evaluate the degree of similarity between the test tasks and the activities students are involved in, or are likely to be involved in, while operating in their real world situations. Additionally the set of descriptors provides a basis for comparing existing alternative test formats in terms of the appropriacy of the test tasks they involve vis à vis the situation the target population is likely to find itself in. We list below those parameters we felt it important to collect information on.

Checklist of General Descriptive Parameters

- (a) Activities - the sub-tasks students have to cope with while participating in events, e.g. in a lecture situation a student might have to listen to the lecturer, take notes from the discourse, copy down dictation or notes from the board, read handouts, etc.
- (b) Setting - the physical and psychosocial contexts of the events, e.g. do students have to operate in the relative quiet of a lecture theatre and seminar room or in the noisier environment of a workshop?
- (c) Interaction - the role set and social relationships students are involved in, e.g. student-student, student-tutor.
- (d) Instrumentality - the medium, mode and channel of the activities within events, e.g. the activity of dictation would be characterised as spoken productive medium; monologue spoken to be written mode and face to face (unilateral) channel.
- (e) Dialect - the dialects and accents the students are exposed to, e.g. R.P. or regional varieties of accent.
- (f) Enabling skills - the underlying skills which appear to be necessary to enable students to operate in the

various activities, e.g. in search reading to get information specifically required for assignments one might identify component skills such as scanning for specifics or separating the essential from the non-essential in text.

As well as collecting data concerning the frequency of activities and the nature of attendant performance constraints that the target population had to cope with, we felt an additional focus for test design was provided by establishing the extent of the difficulty overseas students experienced in coping with these as compared with their native speaker counterparts. As there was limited time available for testing, this would enable us to concentrate on those tasks which exhibited a high frequency of occurrence and where there was the greatest shortfall between the desired performance level and test population behaviour.

3.1.3 Establishing the Dynamic Communicative Characteristics

In communicative approaches to language testing there would seem to be an emphasis not on linguistic accuracy, but on the ability to function effectively through language in particular settings and contexts (v. Chapter 2, Section 2 above). This involves the notion that linguistic activity in the tests should be of the kinds and under the conditions which approximate to real life (cf. Kelly 1978; Rea 1978; Morrow 1977, 1979 and Carroll, B.J. 1978a and 1980).

Davies (1978) argued that we need to do little more than ensure that we have a test of context as well as grammar, in the sense of making our test items more realistic. Rea (1978) took a stronger line and argued the case for constructing tests that involved simulated communicative tasks which directly resembled those which testees would encounter in real life and which made realistic demands on them in terms of language performance behaviours.

The issue would seem to be whether there are in fact dimensions of language use that are not part of existing tests and which from a

communicative perspective need to be incorporated, since it is important that testees be exposed to them (v. Moller 1981b).

Even if the communicative paradigm threw up these hitherto ignored features of language in use (cf. Morrow 1977, 1979; Carroll, B.J. 1978a and 1980) a fundamental problem would still seem to exist in reconciling the realities of communication with the theoretical and practical requirements of assessment. As Davies (1978) pointed out, the conditions for actual real life communication are not replicable in test situations which appear to be by necessity artificial and idealised.

We would agree with him about the illogicality of chasing the chimera of full authenticity, but would nevertheless argue that we should try and make our tests as realistic as possible in terms of the real life situation. For only if we try to make our test simulate as closely as possible the tasks students face in the academic context and the conditions under which these are normally performed, are we in a position to judge whether less direct measures of the same abilities can furnish us with similar evidence about student performance. It seems that we need to try to make our tests as direct as possible in the first instance in order to be able to compare the relative effectiveness of more traditional 'discrete point' and integrative tests which are attempting to measure the same construct. In communicative tests we should aim to provide the opportunity for what Widdowson (1978, p.80) termed "authentic" language use, i.e. putting the learner in positions where he is "required to deal with ... genuine instances of language use" in a way that corresponds to "his normal communicative activities".

If testers are committed to recreating as many of the conditions of real communication as is feasible in their tests, we agree with Hawkey (1982, p.164) that they need to be able to describe what happens "when the parameters of communicative events trigger each other off". While there are no definitive descriptions of such characteristics available, using Morrow (1977, 1979), Kelly (1978) and Hawkey (1982) and a small pilot survey of A.R.E.L.S. schools (v. Appendix 3.1, pp.667-669) as our major informing sources, we

have attempted to list below a preliminary checklist of characteristics so far observed.

Extreme caution must be exercised in using this set of characteristics as criteria for judging the worth of a proficiency test, but given the unlikelihood of our arriving at an adequate theory of communication or of language in use, in the near future, we are at present forced to pay heed to those parameters considered important by practitioners in the language teaching field. No claim is made for the comprehensiveness of this list nor that there is no overlap between categories. In addition, some of the characteristics are more appropriate to one medium rather than another, e.g. intersubjectivity relates more clearly to oral rather than written interaction. It does seem to follow, though, that the more our test tasks reflect the dynamic communicative characteristics appropriate to the target activities, then the more relevant the language behaviour that might result. Even if we are not able to incorporate all of these features into our tests for practical reasons, we still need some sort of yardstick whereby we can judge our own tests and compare them to other tests to see in what aspects they might be considered 'communicatively deficient'.

Checklist of Dynamic Communicative Characteristics

- (a) Realistic context - the test tasks should be seen to be appropriate to the candidates' situation, e.g. as regards type of activity and subject matter.
- (b) Relevant information gap - candidates should have to process new information as they might in their real life situations.
- (c) Intersubjectivity - the tasks should involve candidates both as language receivers and language producers. In addition the language produced by the candidates should be modified in accordance with what their expectations of the addressee

are perceived to be, what Morrow (1979) described as the interaction-based nature of language in use.

- (d) Scope for development of activity by candidates - the tasks should allow candidates the possibility of asserting their communicative independence, e.g. allowance should be made for the creative unpredictability of communication in a number of the tasks set and in the marking schemes adopted.
- (e) Allowance for self-monitoring by candidates - the tasks should allow candidates to use their discourse processing strategies to evaluate their own communicative effectiveness and make any necessary adjustments in the course of an event, e.g. candidates might take notes while listening to a short lecture and then be given the opportunity to monitor these notes before writing them up in a note completion task. Sections of the notes revised for this task can then be used as a basis for part of an extended writing task.
- (f) Processing of appropriately sized input - the size and scope of task activities should be such that candidates are processing the kind of input that they would normally be expected to, e.g. both the written and spoken texts candidates are exposed to, might be much longer than the texts candidates normally encounter in most existing tests.

- (g) Normal time constraints operative - the tasks should be accomplished under normal time constraints, e.g. a dictation or a lecture would only be heard once by the candidates and they might be expected to apply processing strategies just as in a normal academic environment.

3.1.4 Establishing the Test Task Dimensions

Finally, we derive from Hawkey (1982, p.166) that part of the framework for describing the dimensions of particular events. This section of the framework serves two purposes. Firstly, it provides a description in more objective linguistic, stylistic terms whereby the target test task can be related more closely to the dimensions of the equivalent target language activity and, secondly, it enables the testers to plot their performance evaluation criteria against the dimensions inherent in the task itself. We list below those task dimensions we employed in our enquiry into language needs and at the test realisation stage.

Checklist of Task Dimensions

- (a) Size of text - the length of the text, receptive and/or productive that is involved in the event.
- (b) Grammatical complexity and range of cohesion devices required - the degree of syntactic complexity and the range of cohesion devices likely to be required in the event (v. Widdowson 1978).
- (c) Functional range - the degree of variety of illocutionary acts involved in the event (v. Widdowson 1978).
- (d) Referential range - the breadth and depth of lexical knowledge required to handle activities in the event.

For the purposes of our enquiry we regarded phase I as being the most important. Our data collection procedures described below concentrated on establishing the general descriptive parameters of the situation our target population had to operate in, so that we might have a valid, empirical base for the design of our test formats, especially as regards performance tasks and constituent enabling skills. In establishing these parameters we also collected data on phase II and phase III, though for the reasons referred to above in connection with dynamic communicative characteristics and the practical problems discussed below in relation to task dimensions, these data played a less important role at the test realisation stage.

3.2 METHODS OF DATA COLLECTION: Filling Out the Framework

3.2.1 The Observations

3.2.1.1 Introduction

Having established the framework within which to conduct our research, what we needed was a systematic method of enquiry above the precise discourse analysis level which would enable us to observe and record the communicative activities in which students would be involved in the various study modes across a broad spectrum of disciplines, levels and institutions.

If we were to discover what it was that was common across disciplines, in terms of language related study skills that were expected of students, we felt it necessary to observe the activities they were involved in first hand, to establish empirical data on which a questionnaire method of data collection might be based. It would be of little use attempting to establish a more comprehensive description of communicative activities through a questionnaire survey if we omitted to request information concerning vital functional tasks, or if we asked the question in the wrong manner because of our unfamiliarity with the nature of the particular task. Observation of students in situ would help us to ask more intelligent and intelligible questions. Gradually, after a number of trial runs and an extensive search of the literature we developed a working document for use in the observations (v. Appendix 3.2.1, pp.671-688).

As there had been no previous systematic enquiry into the language tasks facing students in such a wide variety of disciplines, we had to devise an enquiry method from scratch. We needed to see how the set of contextual parameters described in Section 3.1 above could be converted into an observation checklist for the collection of data which would enable us to design appropriate performance tasks for a test battery.

We took as our starting point the needs analysis 'model' of Munby (1978) who claimed to have devised 'a dynamic processing model' for systematically constructing a profile of a learner's needs. Though reservations are expressed in the literature about the value of Munby's model for course design purposes, by those who desire a process - as against a goal oriented approach (cf. Brumfit 1980; Mead 1981 and Widdowson 1981), Hawkey (1982) argued that these reservations might not apply to the use of such models in the construction and validation of performance test tasks if one is more concerned with product than process. However, as Davies (1981a, p.332) pointed out, Munby's model is not necessarily "a blueprint which can be automatically applied", but would more aptly be described as (p.333) "a checklist of things to take into account in determining language communication needs".

Experience in applying Munby's model to actual events in the academic context in the first few observations showed that, as it stood, it did not readily lend itself to use as an observation schedule for the gathering of empirical data. In collecting and reporting data the priority of first specifying communicative events, activities and interactions was paramount. If a model is viewed as an analytical tool rather than a 'generating mechanism' a progression from larger to smaller and from sociological to linguistic features is required. In any analysis which is empirically based the procedures need to be inductively, rather than deductively organised (v. Carroll, B.J. 1978a). Events and activities therefore assume a primary importance in our framework whereas they come low down in seventh place in Munby's model and B.J. Carroll's (1978a) interpretation (v. Table 3D, p.134 below). By specifying these first the subsequent components logically derive from what has gone before. Wilkins (1980, p.5) illustrated this point when he referred to Munby's (1978, p.99) example of an air traffic controller and a pilot:

"... it is our knowledge of the language activities that constitute the communicative event that enables us to establish that spoken language, in dialogue form, using radio as channel, is involved, and yet the model suggests that we specify these before we know what the communicative event is."

Munby admitted the extent to which the model lacked linguistic and

sociological evidence and relied on subjectively obtained information. Mead (1981, p.76) took up this point and argued:

"... it is difficult to see how even minimal sets of realisation categories can be distinguished without data based evidence. No indication is given of which factors are significant or non-significant in determining the appropriacy of the language used in particular situations."

Munby's model fails to tell us how a description of the psychosocial features could generate E.S.P. specifications appropriate to learners' needs. It does not indicate what lexical and grammatical realisations are appropriate when these features are present. No matter how much psychosocial data are processed they alone will not provide explicit linguistic realisation rules. Munby's parameters of events, interactions, setting, instrumentality are, as Mead (1981) pointed out, only a framework within which to interpret instances of language and not an instrument for predicting them. It is descriptive not generative. When this interpretation is based on actual data rather than introspection, a more reliable information base may be available.

One option available to us was the possibility of tape recording a number of events. Given the wide ranging nature of our enquiry this would only have provided us with a limited number of transcripts of classes, which, in the manner of Wijasuriya (1971), Holes (1972), Morrison (1974), Straker-Cook (1975), Montgomery (1977) and Mead (1981), could then have been analysed minutely for the occurrence of various features of discourse, such as lexical and grammatical cohesion devices or discourse markers. At this level, though, the occurrence possibilities of any of these factors is infinite and even given unlimited resources and unlimited time, we were by no means certain to arrive at a comprehensive description of the discourse features of even one subject area, let alone all the subject areas we were interested in.

Our aim was to try and find a level at which we could compare activities in study modes across different subject areas. An alternative not open to us, though open to remedial agencies within a particular institution, would be to push the E.S.P. argument to its logical conclusion and write a test for a specific group, on a

specific course. For only through a thorough discourse analysis of a strictly delineated area might it be possible to ensure that the test precisely and 'authentically' reflected the language situation of the intended test population.

At the other end of the continuum one might argue that if it is necessary to abstract to make comparisons between subject areas, then one need only test those linguistic features which are the commonest link between seemingly disparate subject areas, e.g. grammatical items. Because of the reasons discussed in Chapter 2, the latter option, though inviting, did not suit our purpose. We thus decided to explore the possibility of trying to find a level of discourse at a broad functional level where we might make comparisons of the study modes students had to operate in, across a range of subject areas and levels.

Despite serious reservations about Munby's processing model for needs analysis, we found it more useful as an informing source in the design of an observation checklist for establishing common language needs than any other available alternative (cf. Richterich 1973, 1974). As a working instrument for describing the language tasks facing our students, Munby's 'model' needed extensive revision since it was not an adequate tool for use in empirical observation.

We describe below the data collection instruments we devised for, and the procedures which were subsequently employed in, the series of observations conducted at various institutions which were known to have high intakes of overseas students on certain courses (v. Appendix 3.2.2, pp.690-693).

3.2.1.2 The Observation Method of Data Collection

For each academic course we visited (v. Appendix 3.2.2, pp.690-693) a separate observation checklist was completed (v. Appendix 3.2.1, pp.672-688). This checklist comprised a number of forms dealing with the categories outlined in Table 3B below.

TABLE 3B
CATEGORIES USED IN OBSERVATION CHECKLIST

1. Purpose(s) of Study	(p.672)
2. Events and Activities	
2.1 Lectures	(p.673)
2.2 Seminars/Tutorials	(p.674)
2.3 Practical Classes	(p.675)
2.4 Reference Study	(p.676)
2.5 Written Work	(p.677)
3. Setting	
3.1 Physical Setting: Spatial	(p.678)
3.2 Physical Setting: Temporal	(p.678)
3.3 Psychosocial Setting	(p.679)
4. Interactions	
4.1 Position	(p.679)
4.2 Role Set	(p.679)
4.3 Role Set Identity	(p.679)
4.4 Inventory of Social Relationships	(p.680)
5. Instrumentality	
5.1 Medium	(p.680)
5.2 Mode	(p.680)
5.3 Channel	(p.681)
5.4 Non-verbal Medium	(p.681)
6. Target Level	
6.1 Dimensions	(p.682)
6.2 Tolerance Conditions	(p.682)
7. Communicative Key	(pp.683-688)

During each lecture, seminar or practical class that we observed, an appropriate form for events and activities (2.1-2.3) and interactions (4.1-4.4) was filled out. A form was completed immediately after each event for setting (3.1-3.3) and instrumentality (5.1-5.4).

After each event a form was also filled out for target level (6.1-6.2) and communicative key (7). It was felt that the data on these two categories were extremely subjective even though an attempt was made to keep a record during the course of the observation. Category 7, communicative key, was far too specific and too lengthy to be used systematically during an observation and, of all the categories, proved the most impractical on which to collect data, especially in the light of the level of detail and complexity involved in monitoring and recording dynamic events. As we note in Sections 3.2.1.2.6-3.2.1.2.8 below, we were very unhappy about the quality of the data we were able to collect in these later categories.

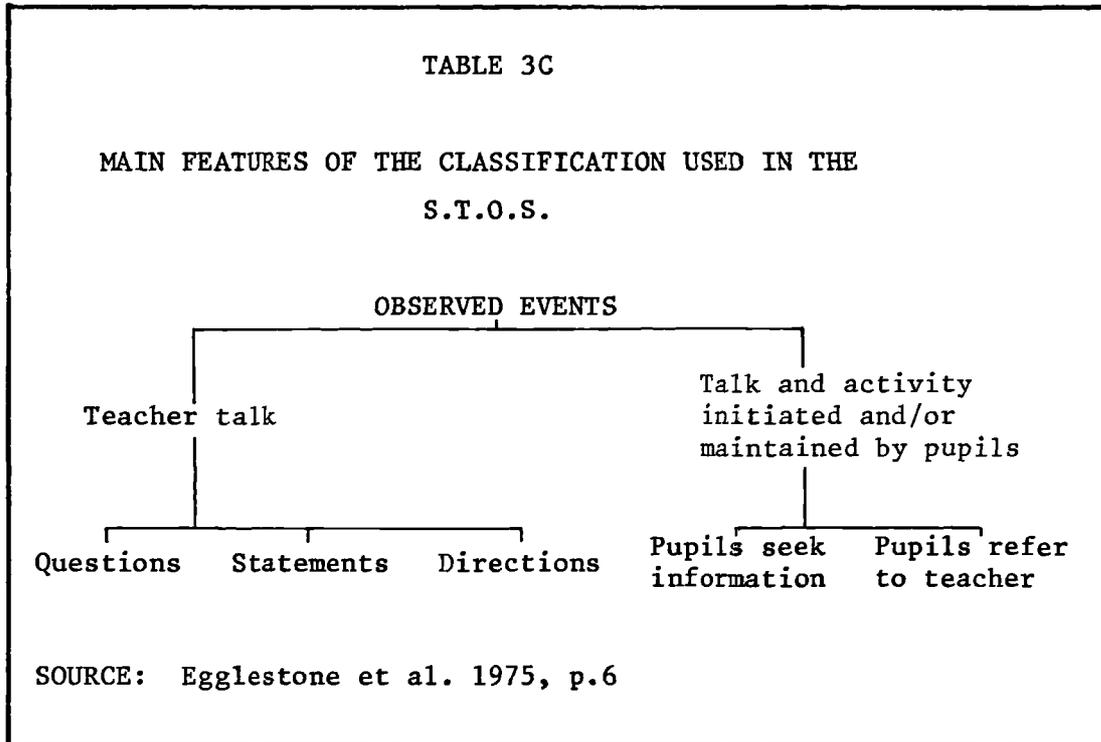
On the basis of information obtained in follow up interviews with staff and students on the course, a form was completed for reference study (2.4) and for written work (2.5).

At the end of all the observations and interviews in respect of a particular course, a final observation checklist was filled out onto which we transcribed a composite of all the data we had been able to collect on each of the study modes.

3.2.1.2.1 Events and activities

As our starting point we needed to devise a method of analysing activities more precisely than Munby allowed for. After a review of the available literature it seemed that, with modification, the system devised by Egglestone et al. (1975) for recording intellectual transactions occurring in science classes, would form a useful starting point for our analysis of events and activities in the academic context. Employing an amended version of their Science Teaching Observation Schedule (S.T.O.S.) (v. Appendix 3.2.1, pp.671-688), we were able to make general comparisons across discipline areas and levels particularly in relation to the linguistic transactions that occur in the study modes: lectures, seminars, practical classes.

Table 3C below illustrates how the activities occurring in these events are divided into more specific categories.



Here was an effective way of ordering linguistic transactions which, with modification, might enable us to record within, for example, the timespan of a lecture, the characteristic activities that took place.

Our adaptations from Munby's model and the S.T.O.S. provided us with a method for quantifying, albeit roughly on a four-point scale, the relative frequency of various exchange/moves during an academic event. Sampling a number of such events in a course gave us an indication of the relative occurrence of these different discourse functions.

Though we relied heavily on Egglestone et al. (1975) in devising our observation checklist, we did not adopt their inflexible method of interaction analysis because this would have limited the nature of the observation to recording mechanically, according to fixed time intervals, the occurrence of a prescribed set of features. We wished to take account of a range of additional features of the situation, e.g. setting, interactions, instrumentality, target level dimensions, target level tolerance conditions, communicative keys and also to capture the overall structure of the interaction rather than a segmentalised version of it. These additional features could not

have been recorded within the rigid constraints of the method of analysis advocated by Egglestone et al. Neither Munby's model (v. Munby 1978) nor S.T.O.S. (v. Egglestone et al. 1975) concerned itself with the non-verbal code, a serious omission which we were in part able to take account of in our observations, albeit on a more impressionistic basis.

As only one observer, this researcher, was involved, the reliability of the more subjective procedure adopted is unproven. The non-mechanical, impressionistic nature of our observational technique as against pure, mechanical, interaction analysis (v. Houston 1974), the researcher's attention span and the limited opportunistic sample of classes under review, all advise caution in the conclusions to be drawn on the basis of our observations alone.

It is, however, worth noting that Gove (1981), in a similar study on 'A' level science students employing our data collection instruments and using the sign system of observation advocated by Egglestone et al. (1975) found (p.39):

"... an overall similarity between the two sets of results derived from observation in respect of the 'A' level science group."

After familiarisation with the categories selected for recording information and a series of trial runs we felt that a reasonable record of activities and interactions in a lesson could be kept. In recording our observations of activities we used a four-point scale:

- H: High frequency of occurrence: a feature takes up a considerable part of the event or occurs a large number of times.
- M: Medium frequency of occurrence: a feature occurs more than occasionally, though does not occur a large number of times or constantly over a long stretch of time.
- L: Low frequency of occurrence: only a few isolated instances of a feature, occurs only occasionally, perhaps one, two or three times in a particular event.
- N: Non-occurrence: feature does not occur at all in a particular event.

Estimates were made in the context of a whole lesson, with a gradual cumulative progression up the scale, where required, as the lesson developed and we gained a better overall impression of frequency of occurrence. Our observations were checked through after each lesson to see if they accorded with the overall impression we had gained of the occurrence of various features.

A check was also made, where possible, with the teachers who had conducted the classes to see if they regarded our observation as a fair representation of what had occurred and whether it differed widely from the norm of that particular study mode as far as his or her own individual teaching style in that course was concerned.

Over a series of observations of a particular study mode, e.g. lectures in a course, it was found that the extent of occurrence of a particular feature might vary considerably according to a number of factors, e.g. the stage reached in the course, the subject content of the course, the complexity of the material being presented, the teacher involved and the methodology employed. Where this spread of occurrence arose, we indicate its range in our data specification in Section 3.4 below.

Having described the general procedures adopted in the observations and examined, in particular, the recording of events and activities, we consider below the other areas on which we collected data.

3.2.1.2.2 Setting

This was an attempt to specify the physical and temporal settings within which language activities were based and to provide a list of possible psychosocial environments in which students might operate.

In the observation schedule and follow up interview we attempted to establish a general picture of the most likely settings the students we were concerned with were likely to find themselves in. The results of the findings are summarised on page 306 below. The information in part 3, the inventory of psychosocial environments, is based largely on the impressions gathered by the researcher

during the observations. We have taken from Munby's (1978) list of psychosocial environments those which proved to be relevant to the observations we were able to make. We would, however, agree with Davies (1981a) about the practical difficulty of being clear about vague categories such as setting or interactions.

3.2.1.2.3 Interactions

This was an attempt to establish the positions in which students had to enact certain roles; the role set, the different people with whom they needed to use English when enacting a particular role; the role set identity, the number, age group, sex and nationality of members of the role set and the social relationship within which the learner had to use the language. Again Munby (1978) provides a set of rather vague categories to describe the type of relationships the member of the 'role set' might find himself involved in. As a result of subjective decisions arising out of our observations we have recorded in Table 3BB, page 308 below, some tentative impressions of the role sets, role set identities and social relationships by which the students in our survey are best characterised. One can only draw up specifications in the light of limited and, perhaps, short-term social relationships for, as Ng'ombe (1981, p.72) pointed out:

"It is not possible to imagine an interlocutor who is perpetually junior or senior in relation to other interlocutors."

One can predict from the observations a range of such relationships but some will apply to a limited number of interlocutors, some perhaps to all.

3.2.1.2.4 Instrumentality

This parameter deals with the medium, mode and channel of communication. All three methods of enquiry, observation, interviews and questionnaires helped us establish the important features here (v. Table 3CC, p.310)

3.2.1.2.5 Dialect

This parameter covers the dialects that the student needed to produce or understand. In our opinion this is of much less importance as a category as compared to interactions or events/activities. Broadly speaking, students on different courses in different institutions were exposed to a whole range of accents, but not normally to marked regional dialect variation. We concluded that students need to be able to cope with a range of accents of an intelligible nature in an academic environment.

3.2.1.2.6 Target level dimensions

As well as collecting information on events and activities, setting, interactions, instrumentality and enabling skills during the observation stage of data collection we also sought information on target level dimensions (v. Appendix 3.2.1, p.682) and tolerance conditions (v. Appendix 3.2.1, p.682).

When we started the series of observations in January 1980 we attempted to collect data of this type in accordance with Munby's specification (v. Munby 1978), but soon realised the sterility of this endeavour. An examination of the examples in B.J. Carroll (1978a) detailed in Spec. 6, Table 3D below, illustrates the problems involved and shows the fallacious nature of examples and figures provided for the tolerance conditions in particular.

We agree with Hawkey (1982) that Munby's categories for establishing target level dimensions (v. Table 3D below): size, complexity, range, speed, delivery and flexibility, are "internally confused". Munby's concept of range must subsume his dimensions of complexity and delicacy for once one includes the "quantitative extent" of the formal, functional and cognitive dimensions under one label - there are not many dimensions left uncovered. Munby (1978, p.165) includes 'discourse-coherence' under his complexity label. This inevitably takes us into the area of how illocutionary acts relate to each other (cf. Kelly 1978 and Widdowson 1978) and thus encroaches on his dimensions of range and delicacy. For our purposes,

we have followed Hawkey and limited complexity to the syntactic level and take it to include cohesion as described by Widdowson (1978).

The delicacy of forms and functions was almost impossible to assess whether one took it as referring to specificity and detail or, as Hawkey (1982) suggested something more akin to subtlety. We agree with Hawkey that it is better handled under the ideas of referential and functional range (v. p.120 above). Speed of communication is better seen as a communicative characteristic and as a criterion for evaluation rather than a task dimension. Flexibility, if one extends its meaning beyond that given by Hawkey to cover role as well as topic switching is best subsumed under the functional range category.

A further problem in attempting to apply Munby is that no allowance is made for the different type of setting the student will find himself in. Are we simply to assume that a task's dimensions will be the same, whether it is performed in a lecture, practical class or seminar? If they are dissimilar, doubt is immediately cast on B.J. Carroll's (1978a) figures in Table 3D below; one would in any case wish to know how the figures he quotes were arrived at. Given the paucity of empirical observation how can we be certain the target levels for the H.N.D. business studies student described in Table 3D below reflect the different demands placed on him by different study modes encountered in his course, let alone those placed on all business studies students, in different courses. The use of a seven-point scale does not, in itself, guarantee precision or objectivity (v. Miller 1956) and in our initial trials we found it very difficult to operate. A greater degree of consistency resulted from using a much simpler scale of high, medium and low (v. Appendix 3.2.1, p.682). This provided a way of broadly characterising the dimensions of a communication activity which might prove informative at the test construction stage.

TABLE 3D

Specification of Communicative Needs

SPEC. 0 The Participant

Age: 20s
Nationality: Nigerian
Language: Hausa
English std.: Intermediate

SPEC. 1 Purpose of Study

Course: HND Business Studies
Polytechnic
Study Areas: Business Studies:
Economics, Law,
Business Accounts,
Statistics,
Marketing,
Purchasing
General Area: Social Sciences

SPEC. 2 Setting for English

Physical: Lecture room, tutorial
room, library,
factories, business
offices
Temporal: Full-time in term,
plus vacations,
Av: 10 hrs. p.d.

SPEC. 3 Interactions

*Learner-instructor
*Outsider-insider
Non-professional-professional
*Non-native-native
*Insider-insider
*Adult-adult
Note: Interactions recorded 3 or
more times are marked with
an asterisk.

SPEC. 4 Instrumentality

Medium: Listening, speaking,
reading, writing
Mode: Monologue, dialogue
(Spoken and written to be
heard or read; sometimes
to be spoken as if not
written.)
Channel: Face-to-face, print,
tape, film

SPEC. 5 Dialect

All sections: Understand British
Standard English dialect. Produce
acceptable regional version of
Standard English accent.

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

Dimensions:	L	Sp	R	Wr
(max=7) Size	6	3	7	3
Complexity	7	4	6	5
Range	5	4	5	5
Delicacy	5	5	6	6
Speed	6	4	5	6
Flexibility	5	5	3	3

Tolerance Conditions:

(max=5) Error	3	4	3	3
Style	4	4	5	4
Reference	3	4	2	2
Repetition	3	4	2	3
Hesitation	3	4	4	3

SPEC. 7 Events/Activities

- 7.1 Lectures: Listen for overall
compr. Make notes.
Ask for clarification.
- 7.2 Seminars/Tutorials: Discuss
given topics. Listen
for compr. Make
notes. Ask for
clarification.
- 7.3 Reference Study: Intensive
reading. Rdg. for
main information.
Assignment rdg.
Assessment rdg.
- 7.4 Writing Reports: Sort out infm.
Factual writing.
Evaluative writing.
- 7.5 Keeping up-to-date: Routine
checking. Rdg. for
intensive. Rdg. for
infm. search.
- 7.6 Indust./Comm. Visits: Discuss
topics. Discuss after
visit. Listening for
infm. Take notes.
Ask for clarification.

As it turned out we found that our recordings, even with the improved dimensions of size, complexity, functional and referential range, varied from low to high in the sum of the classes observed for the various events across the broad discipline areas and levels.

Due to this poor discrimination, we were able to do little more in our selection of texts for the test than ensure that they were regarded as appropriate for 'A' level and undergraduate students in the disciplines under review. To this end we consulted both staff and students from the various discipline areas and levels as to the suitability, in terms of the stated task dimensions, of the spoken and written texts we had selected. A further condition to be met was that native speakers in these groups could handle the texts with limited difficulty. The methods employed for selecting texts in terms of size, complexity, functional range and referential range are more fully discussed in Chapter 4 below.

The apparent applicability of what we have considered as task dimensions to the evaluation of proficiency had been noted both by Carroll, B.J. (1978a, 1980) and Morrow (1979) though both tended to use them to move directly into the specification of assessment scales or bands. For our purposes we do not regard these dimensions as assessment criteria, but rather as objectives in that they describe a communicative event as it could or even should be. We agree with Hawkey (1982) that the dimensions used to describe particular events and activities should be kept separate though, obviously, not divorced from the criteria for the assessment of test task performance. For, as Hawkey (1982, pp.168-169) argued:

"... the tester needs to be balancing the achieved as against the required result ... The dimensions, based on native speaker performance of a task, are concerned with the required result; the assessment criteria are about the communicative effectiveness actually achieved. It is important in relating the required and achieved levels of performance to distinguish between the dimensions inherent in the event itself ... and the criteria for evaluation."

Hawkey's approach seems preferable to that of Morrow and Carroll, in that it is useful to be able to relate in a fairly systematic manner the evaluation of performance and task dimensions rather than

attempting to conflate the two.

3.2.1.2.7 Target level tolerance conditions

Munby's target level was also coded for tolerance in the dimensions of linguistic error, stylistic failure, reference (to dictionary, etc.), repetition (repeat/ask for repeat), hesitation (lack of fluency).

These were seen by Munby as a way of modifying the predicted dimension of target level activity in the light of allowances made by those with whom the candidate has to interact.

In our attempt to collect data on the tolerance conditions applied, we relied on information gathered incidentally during the course of the lesson and more specifically in the follow up interviews. It soon became evident that alterations were made to suit individual circumstances and varied from one study mode to another, e.g. as between seminars and tutorials. Lecturers indicated that they themselves were inconsistent in the allowances they made and these might vary in the case of a particular student even from one occasion to another. It was certainly not possible to quote a single figure for a course when more than one tutor was involved.

We would argue that, in practice, tolerance conditions are almost impossible to establish with any degree of accuracy. Even when we asked in the questionnaire whether allowances were made for overseas as against British students in written work, with no quantitative decisions to be made, no common policy was evident within levels or discipline areas.

Having tried to implement this particular aspect of Munby's model we would be interested in discovering how B.J. Carroll (1978a) managed to specify the tolerance levels included in Table 3D above. How does one, in fact, establish a tolerance level for reading?

3.2.1.2.8 Communicative key

This parameter of Munby's was concerned with how the activities comprising an event are performed. According to Munby (1978, p.39) the categories expressed:

"... the likely attitudes or keys that need to be produced or understood in connection with an event."

We would agree strongly with Davies's (1981a) reservations about this parameter which attempts to specify the attitudinal tone in which a communicative activity is carried out. This is one of the areas where Munby has particularly underestimated the practical difficulty of attaining the kind of specificity he seeks. Armed with a full list of attitudinal tones (v. Appendix 3.2.1, pp.685-688) we attempted in the early stages of the observation programme to record details of communicative keys. This was an extremely subjective process especially in view of our limited sample. It soon became apparent that it was impossible to do systematically whilst the event was taking place, owing to the speed with which activities take place in real life. It was also difficult to decide upon the most appropriate adjective to describe the tone and in some cases, to decide whether any attitudinal tone was in fact being expressed and if it was, whether it was the one the observer believed it to be.

As none of the events observed was tape-recorded it would be impossible to make any claims for reliability with regard to the attitudinal tones employed. What were noted by the observer were subjective impressions of those tones, which had seemed to be important to an understanding of the particular lecture/seminar/lab session and which stood out as having been encountered by students. The result is probably a more restricted list than actually occurred, but it seemed that a certain set of attitudinal tones was important and these are listed as Appendix 3.2.3, page 695.

Despite the fact that our data are empirically based on real as against hypothetical students (v. Carroll, B.J. 1978a), we make no claims for its generalisability or validity, or for whether other instances of the same behaviour would necessarily produce the same attitudinal tones. Because of their unsatisfactory nature these data were ignored

in designing our test.

3.2.1.2.9 Enabling skills

In the second stage of Munby's model the information derived from the Communicative Needs Processor (C.N.P.) is processed into the syllabus content specification. The activities (and their communicative keys) are first established and then decisions have to be made as to which is the most suitable of three alternative ways of processing them - focusing on microskills, microfunctions or on linguistic forms.

Munby (1978) admitted that the examples he gave in the linguistic encoding section were intuitive guesswork. Davies (1981a, p.335) argued:

"... the usefulness stops at the point of linkage from the C.N.P. to the Language Skills Selector when the entries become intuitive and ad hoc. The assumption the book seems to make is that if the complete model is applied then there will be certainty about the language output."

It is difficult to see how Munby's linguistic encoding can realise the microskills in linguistic forms. It is even less clear if one is working with microfunctions where, as Davies (1981a) pointed out, the most interesting functions have a very large number of possible language realisations. It would certainly seem that there is no way of being certain about the language output component of the "model".

In the absence of any explicit (or implicit) guidelines in Munby (1978) about when specifications should focus on microskills and when on microfunctions we decided that the former seemed to suit our purpose better. We attempted to process our information by focusing on microskills, since, if these could be empirically verified, they would have an application which extended beyond any one particular instance of performance and would reflect an aspect of what is generally understood by competence (v. Morrow 1979).

Through our focus on events and activities in the observations, interviews and questionnaires we attempted to gain some insight into the enabling skills our students would appear to need to cope with the various tasks in different study modes. On the basis of the empirical evidence collected, we offer, at the end of each section below, a list of those skills which appeared to be required by science, engineering, social, business and administrative studies students. It seemed that these skills were to a greater or lesser extent, required by all students on all courses.

Further problems arose out of the fact that, although Munby (1978) twice listed his taxonomy of enabling skills, no attempt was made to place them in any hierarchical order. There was no attempt either to suggest how the skills relate to each other; ways in which some skills may be considered superordinate to others or possible implicational relationships of possession of higher order skills for the possession of lower order skills. Ideally, there is a need for some sort of implicational scaling, i.e. an hierarchical arrangement whereby we might be able to say if X gets these factors right he should get the rest. Mead (1981, p.75) pointed out:

"... insofar as the skills represent options of equal weight the system is self-contradictory. Skills 20-25 would appear to presuppose the learning of all other skills, but as the system is organised, it is implied that the learner might wish to manipulate variations in stress without necessarily needing to express himself explicitly or implicitly."

As we mentioned above there is also a danger in considering the skills as discrete, as it is by no means clear whether we use them singly or in combination to complete successfully a particular task.

We decided that in those test tasks which related to a particular study mode, where we focused on either listening or reading skills as against adopting a more integrated approach, we should try as far as possible to include items which focused on each of the

enabling skills identified as contributing to proficiency in these constructs and certainly on those we considered to be higher order skills.

3.2.1.2.10 Conclusions

Through the series of observations we arrived at a more valid and comprehensive description of the purposes, activities, setting, interactions and task dimensions our target population were subject to than is possible by armchair needs analysis alone. In addition, through these empirical data collection procedures, we were able to identify more closely the constituent enabling skills that might be needed by these students for the successful performance of various activities in the academic context (v. Carroll, B.J. 1978a).

Though we have been highly critical of Munby (1978) and Carroll's (1978a) interpretation, we nevertheless feel that there were elements in these studies which contributed to the methodological procedures we adopted. Our point of departure was the need to devise techniques for collecting data from comprehensive observational studies of students actively engaged on their courses. The failings of Munby and Carroll largely stem from this deficiency.

In addition to the data collected through the observations, we carried out follow up interviews with teachers and students whenever possible (v. Section 3.2.2 below). These two procedures helped inform the major data collection method, the questionnaire survey described in Section 3.2.3 below.

3.2.2 The Follow Up Interviews

In addition to the observations an attempt was made, whenever possible, to carry out a follow up interview with staff and students on the course under review. The rationale behind this was that it would enable us to:

- (a) collect information impossible to gather through observation, e.g. data about reading and writing activities: copies of examinations, handouts, students' written work, information about criteria of assessment used in marking written work; information concerning the amount of course time spent in lectures, practicals, tutorials and seminars;
- (b) check the data gathered during the observation for its generalisability to the course under review;
- (c) establish the main English language problems experienced by overseas and British students;
- (d) examine the validity and efficiency of the questions in both staff and student pilot questionnaires.

Data recording consisted of taking notes during the interview with the participant's permission, so that the interview could be written up immediately afterwards.

Walker (1978) in his research, depended solely on interview methods of data collection arguing that it allowed a more flexible question format and meant that the interviewer was able more effectively to select questions and check that the interviewee understood what was being asked of him/her. However, the value of the approach is affected by the difficulties involved in arranging interviews and their time-consuming nature. In addition, we were well aware of the complex nature of successful interviewing (Kahn et al. 1957; Shipman 1972 and Engelhart 1972) and the difficulty of achieving what Walker (1978, p.52) described as:

"... a dynamic interactive experience in a friendly, permissive atmosphere, encouraging the respondent to reveal information and motivating him to keep presenting useful facts."

The interviews conducted must be seen as an opportunistic sample, however, because of the problems of availability of staff and

students. Owing to the limited nature of the conclusions that could be drawn from the interview data alone, we have not recorded them separately.

3.2.3 The Questionnaire Method of Data Collection

The data derived for the series of observations were to be supplemented by a questionnaire enquiry. Various pilot versions of the questionnaire were tried out during 1980 (v. Appendices 3.3.1 pp. 698-737) on staff and students with whom contact was made during the series of visits to various educational institutions in the further and higher education sectors to observe teaching. These questionnaires were constructed on the basis of the information established through the observations together with a review of the relevant literature (cf. Hale 1964; Holes 1972; Cowie et al. 1977a; Candlin 1977 and Ryan 1979). As a result of the feedback obtained in interviews with staff and students, we were able to refine the final versions so that they came closer to asking questions in a way that produced the information we required (v. Appendix 3.3.1, pp.738-787). Questions which had yielded the wrong information or which defied accurate response were either omitted or rewritten. Replies to certain of the open-ended questions led to additional questions or supplementaries being asked in the final version. We were thus confident that we were asking questions which were both relevant and intelligible to the correspondents at least insofar as they reflected the language operations arising out of their courses of study. We attempted to simplify the language, as far as this was possible, without losing any of the delicacy in the questions posed.

The question of length was a serious problem. In the end we cut down the questionnaire to what we felt was the basic minimum of information we needed, but still felt that it was probably off-putting for a large number of the correspondents.

During 1981 we contacted all the university and polytechnic science, engineering and social, business and administrative studies departments and colleges offering G.C.E. Advanced level science, where we knew from earlier research that there were large numbers of overseas

students studying and asked them to assist us in our project. We then asked those who were willing to co-operate to let us have details of the numbers of overseas students, for whom English was not the first language in the country of origin, enrolled on specific courses within their departments, together with numbers of the staff who taught them. The final versions of the questionnaire (v. Appendix 3.3.1, pp.738-787) were then sent to staff and through them to both British and overseas students. The questionnaires were then returned individually in pre-paid envelopes to try to ensure that those completing them would not feel inhibited. In all a total of 5947 questionnaires were sent out. The responses to these questionnaires enabled us to establish a wider basis for our description of the language tasks facing students in a variety of subject areas and provided us with information on the extent of difficulty both overseas and British students encountered in coping with a variety of tasks and constraints in the academic context. Completed questionnaires were received from 940 overseas students, 530 British students and 560 staff, in respect of 43 post-graduate courses, 61 undergraduate courses and 39 Advanced level Centres (v. Appendix 3.3.2, pp.791-800).

We were aware that our survey was limited and that we could only claim to have established some preliminary parameters. We did, however, attempt to describe the teaching situations at the transactional level (Sinclair et al. 1975) as systematically as possible and would argue that, despite the limitations of our endeavours to establish by questionnaire and observation what was common in terms of language skill requirements across disciplinary boundaries, we collected a basic core of empirical evidence on which future research could build and against which our tests could be compared. More immediately we had a specification which attempted a description of the actual academic context that students would have to operate in. We also established in general terms through the questionnaire, those academic activities overseas students experienced difficulty with as compared to their British counterparts and this helped us decide on the types of test tasks we needed to include in our battery in order to assess what Candlin et al. (1976, p. 2) described as:

"... mastery of study skills competence in what might be called 'didactic discourse', the discourse of all teaching/learning situations."

What follows below is an examination of the general descriptive parameters that relate to the academic context, in which the students, in the levels and disciplines under review, were operating. These parameters are used to inform the test task construction phase (v. Chapter 4). We, of course, recognise that they are indeed only parameters in the sense of being quantities which are constant with regard to the particular cases considered, but which are likely to vary in different cases. They are considered under the following headings.

- (a) Purpose(s) (of the participant(s) in the event and of the event itself).
- (b) Activities (sub-tasks involved in achieving the purpose(s)).
- (c) Setting (physical and psychosocial).
- (d) Interaction (role set and social relationships of participants).
- (e) Instrumentality (medium, mode, channel of communication of the event).

3.3 THE PURPOSE OF LECTURES, SEMINARS/TUTORIALS AND PRACTICAL CLASSES

3.3.1 The Lecture

For the purposes of our questionnaire survey, we took the lecture as the teaching period which is mostly occupied with continuous talk by the teacher. There may be some opportunity for questions, but in the main all that students have to do is listen and take notes. Straker-Cook (1975, p.44) pointed out that for most levels and disciplines the lecture may be defined as a monologue situation:

"... regardless of the number of participants, only one is a protagonist: he is responsible for almost all the verbal contribution to the discourse, under social conventions which allow other participants to intervene only very occasionally, only very briefly, and only on the invitation (verbally or non-verbally indicated) of the protagonist."

Our investigation gave us an indication of the relative amounts of monologue and interactive discourse occurring in different types of classes, at different levels, in different subject areas.

The widespread prevalence of the lecture as a teaching method across levels and subject boundaries gave it an important place in our study for, as the literature illustrated, although lectures are not an effective way of stimulating thought, or bringing about a change in the opinions of an audience, they are as effective as other methods for conveying information (v. Bligh 1972).

Our observations confirmed Parish's conclusion (1977, p.75) in her survey of university teaching and learning methods that:

"... the lecture is widely used in all departments."

Its primacy as the teaching technique most frequently employed to aid students in the acquisition of knowledge is borne out by the returns to the pilot questionnaire and the comments advanced in interviews conducted during the visits made to institutes of further and higher education. A majority of students reported having at least ten lectures a week and often substantially more, with only

post-graduate social scientists consistently dropping below this figure.

Academic staff were asked in the questionnaire to comment on what they saw as the purpose of the lecture. Many commented on its value as a method for information transfer, for example:

"The orderly presentation of factual and practical data in a distilled form."

"Give them the most important ideas on any topic."

"One of the main means by which the principles, concepts and biological themes are explored and transmitted."

"It provides them with the basic material of the syllabus."

"They provide the framework within which the students can study and organise their reading."

As well as distilling and informing on the current stage of knowledge it was seen as a means for explaining difficult concepts and their significance in a wider context. As one engineering professor expressed it:

"Getting information across while keeping the students awake. Making them realise that what they are taught does relate to the real world."

Linked to its function of transmitting information orally some viewed its importance as:

"Providing students with a reasonable set of notes on the subject."

Many of the staff emphasised how it provided a "common basis", "a structure", "a grounding" for future work:

"Exposition/explanation of basic principles, giving conceptual framework within which students can develop detailed knowledge."

and it was seen as a means of bringing students of disparate academic backgrounds up to a minimum level of shared knowledge in the subject area.

Several staff referred to its value as a platform for advice on further reading whilst others also pointed out that, on some topics,

it gave the students access to material in areas where there were:

"No good text-books, published material sparse and widely scattered."

or where:

"Material (was) not easily available elsewhere."

or it:

"Highlights important and explains particularly intricate issues in highly specialised areas of study not always adequately covered in published literature."

A biology professor drew attention to a further important dimension:

"(The lecture) creates in the students' minds an idea of the attitudes to the subject of the lecturer, by summarising work leading to a particular conclusion or description, if possible adding those studies that suggest an alternative view. A lecturer should add his own assessment, giving reasons why he favours that conclusion."

The widespread preference for the lecture method often stems from its practical benefits rather than its intrinsic superiority as a teaching technique:

"Much of our teaching is solely via lectures because of the large classes."

"It is the main teaching method used in our year one courses due to large student numbers."

Students were similarly invited to express what they saw as the purpose of the lecture. Most commented on its use for transmitting information, for example:

"To pass on to the student the relevant information."

"To provide us with the knowledge that we would require for our course."

"To introduce the student to the subject and explain, with examples, its applications."

"To convey the main ideas of the subject matter as well as being a means of gaining a set of notes for the course."

Additionally, emphasis was laid by some on the function:

"To give students a better understanding of the subject matter."

and some students commented on its value in directing them to appropriate reading matter:

"To stress the more important areas of the course to be followed up with students' reading."

"Give us the basic knowledge and guidelines for further reading in a particular subject."

"Outline the topics briefly and show where further reading is necessary."

As an ideal we might agree with Parish's (1977, p.75) conclusion:

"The main value of the lecture is that staff can present up-to-date material based on work experience and information from a wide variety of sources, selected, appraised, and presented to the student in a structured way."

3.3.2 The Seminar and the Tutorial

These teaching methods were grouped together in the questionnaire since the terms are often used loosely and overseas students might be unable to distinguish between two teaching modes that were, in the event, quite similar. It became apparent during the observations that, even when considered together, they occupy a small proportion of total contact hours except perhaps for post-graduate social science students and certain groups of undergraduates (v. Table 3AA on setting, p.306 below). In some cases certain groups of students rarely encountered any variant of this technique. For the purpose of the questionnaire an inclusive category for seminars and tutorials was characterised as differing from a lecture in giving much more opportunity for the participation of students, e.g. there may be reading and study of a paper by a student; discussion of topics after a short introduction by students or teacher; a teacher may go through written work or questions prepared by student(s); there may be discussion of any matters or problems on the initiative of student(s) or teacher; the student(s) may work through problems set by the teacher.

From comments made by staff during the observations, the low frequency of this teaching activity in most discipline areas would seem to arise

out of temporal and manpower constraints and this is certainly not a reflection of its potential value as a teaching technique for assisting learners to apply acquired knowledge or skill.

Many of the staff consulted, saw its value as a further means of information transfer:

"To develop a theme derived from the lecture framework in its full complexity and with regard to the full range of empirical material which can be marshalled in support/ opposition of the said theme."

"Clarification of lecture information and any other information acquired by the students through the process of discussion."

"To cover material not given in lectures, perhaps because of insufficient time or because it cannot be suitably dealt with in the form of a lecture."

Others saw it as assisting appreciation and understanding:

"To allow the student to assess the relevance and correctness of theories and modes of analysis expressed in lectures and books."

"Relating materials to whole degree course."

Many took the view of it as a sort of academic first-aid post:

"Going over example papers which they all too often seem incapable of doing without help."

"Surgery for those having difficulty with lecture course and/or mathematical content."

Others commented on its value in applying acquired knowledge or skill:

"To teach them how they can use the subject."

"To improve the capacity of students to analyse issues."

"To reinforce learning through lectures and books by converting it into active rather than passive knowledge."

"Teaching legal method i.e. application of rules to hypothetical situations."

Additionally mention was made of various functions such as allowing the teacher to direct attention to further reading, stimulating interest and in a limited number of cases 'assessing performance'.

Only occasionally was mention made of the 'interactive' nature of the seminar/small tutorial group:

"To try and get an argument going which is neither contrived nor a mere dialogue (or worse still monologue)."

"To discuss the topic in question with as little participation by myself as possible."

More frequently the purpose of a seminar or tutorial was:

"To give the students an opportunity to talk and develop their own thoughts, within the framework of fairly 'strong' guidance of the discussion by members of staff."

As the report of the observations below confirms, this was much more like the true picture of a teacher-centred seminar or small tutorial group with most of the interaction being channelled through and by the teacher. As Parish (1977, p.76) lamented:

"The situation is one of paradox: in each Department importance is attached to the active participation of students in the learning activities; yet seminars (which are ideally suited for this) appear to be conceived in ways that under-involve the students in their education."

For the student, these periods were seen as affording opportunity, at least more than in lectures, to discuss topics and develop ideas:

"The students can discuss a topic using their readings."

"To study topics in greater depth with opportunity to discuss."

"To deepen understanding of specific topics by discussing with students and lecturers."

For some groups of students, most usually post-graduates in the social sciences:

"Most of the basic coursework is discussed or introduced during these sessions."

Many saw them as offering at least the possibility of greater involvement:

"We hope to be involved more in the subject."

"The students could participate and clarify doubts."

The latter point was taken up by many students who looked upon these classes, especially the smaller tutorial group, as being useful for clarifying problems:

"To help you with any problems that you have in lectures, practical classes or even your homework."

"To discuss topics which are unclear."

"Quite useful as they help us understand certain lectures that are quite difficult at first."

"To enable the student to put forward his/her views on the subject, ask questions and bring up any difficulties."

Finally, a number of students also regarded this type of class as a means of applying skills and knowledge:

"Learning by doing, analysis through discussion and argument with the teacher as chairman guiding the discussion."

"Opportunity for student to talk on a prepared subject and then participate in the tutor directed discussion, usually should be conversant with the material at hand and aim to apply it to the issues being covered."

"Apply what is taught in lectures to problems."

"Putting and applying the theory to problems and trying to solve them."

3.3.3 The Practical Class

The function of this particular teaching technique is to assist students in acquiring a variety of skills. For the purpose of our survey we characterised it as the type of class in which students do exercises under the supervision of a teacher, which involve the handling of equipment, instruments, or specimens of some kind, e.g. scientific experiments, learning to use calculating machines, drawing plans, using industrial machinery.

Many of the staff consulted, emphasised its importance for skills development. They variously saw its purpose as:

"Acquisition of engineering feel."

"Development of practical skills."

"Learning of techniques, scientific method, presentation and manipulation of data."

"Develop familiarity with handling material, apparatus, data, etc."

"To develop skills in experimental techniques."

Its other major contribution was that of illustrating theoretical input:

"To get the student to attempt to use theory."

"Personal verification of principles outlined in lectures."

"To enable them to use their knowledge to analyse their experimental results."

Students were also canvassed for their opinions and similarly they saw practical classes as important in skills development as well as enabling them to link the theory derived from their lectures, seminars and background reading with its more tangible practical application.

Many referred to its importance in skills development:

"See how a computer operates."

"Give an idea of how the theory and the equipment fit together."

"To train us to use instruments."

"It's quite useful as it teaches us to handle equipment."

Others saw the practical class as providing a bridge between theory and practice:

"To introduce to us the application of the equation or theory or idea we have, or are going to, study."

"Apply what is gathered from theory into practice."

"To relate subjects studied with reality."

"These are very important as they help us to visualise the concepts we have learned in theory through the lectures."

"They give us the opportunity to really see what's happening. It is easier to understand something when we actually see it in action."

3.4 ACTIVITIES IN THE ACADEMIC CONTEXT

3.4.1 A Cautionary Note

The returns to the postal questionnaire were not a random sample of those we sent out as we had no control over the selection of those who returned it. We could not demonstrate that there was not a correlation between a propensity to return a questionnaire and certain of the characteristics we were intent on measuring which would mitigate the validity of our results.

It is worth noting that while administering the pilot questionnaire under close supervision, at the Universities of Exeter, Southampton and Reading and at Farnborough and Padworth Colleges, the weakest students in the classes nearly always had difficulty in fully comprehending what was being asked of them. Despite great effort being made to ensure clarity and simplicity in the wording of the questionnaire, by constant revision and pre-testing during the observations, its very length meant it was a daunting prospect for the weaker overseas students especially for poor readers. It is not unreasonable to speculate that many of the weaker students from overseas were unable to complete the final version of the questionnaire they received. Those with the worst problems might not therefore be represented in the final breakdown.

Added to this is the evidence that exists showing how the weaker students tend to underestimate the difficulties that they experience (cf. Chaplen 1970; Sen 1970 and Walker 1978). Jordan (1977a, p.14) carried out a series of tests on overseas students and found:

"... in 1972-73 and again in 1974-75 students' self assessment ratings were examined and compared to the students' scores on the Chaplen Tests. Overwhelmingly the results showed that the students at the lower end of the scale in the tests grossly overestimated their language ability ..."

Thus, estimates of the extent of difficulty encountered in the questionnaires might be too low. The questionnaire was also administered from February to April, by which time most of the

students completing it would have had at least six months to improve their language ability as well as become aware of any problems that existed.

We would argue therefore that differences between non-native and native speakers of English, where they exist, are in all likelihood greater at the beginning of their studies than is indicated by the student questionnaire returns.

In the discussion below we merely claim to be establishing descriptive parameters on the basis of the information made available to us in both the observations and returns to the questionnaire. Both methods of data collection had their limitations.

In the observations only a relatively small number of post-graduate courses were visited because of difficulties arising from the small numbers on such courses and the large number of courses without any formal fixed points. We were thus unable to visit any post-graduate engineering programmes and only a very limited coverage was possible for science post-graduate classes. It is also likely that the more courses, at each level, and in each subject area, we had visited, the greater would have been the variation in the activities that we were able to observe. This would have resulted in a wider spread in the frequency of occurrence of certain of the activities in the observation schedule than we have reported.

In the questionnaire survey, we initially contacted all the departments in the country who, according to the figures established in an earlier investigation, had reasonably large numbers of overseas students on their programmes. We informed them of the nature of our project and asked them, if they were willing to co-operate, to let us have appropriate details of overseas and British students and staff who could be sent questionnaires. We were thus limited by the extent to which departments were willing to co-operate in distribution and staff and students in returning the questionnaires to us.

The sample for our questionnaire suffers from this defect. Firstly there are not equal numbers of respondents in each level or discipline area. Overall there are a large number of overseas

science 'A' level students as compared to all other groups. The university undergraduate science replies are limited in number with respondents mainly from mathematics based courses and in addition no replies from chemistry students are represented in these returns. The undergraduate social science returns are influenced by two large groups of Nigerian students on the N.E. Wales Polytechnic B.Ed courses. There are also limited numbers of returns in all the post-graduate courses. Thus we cannot claim that our sample is a random sample of the student population at large.

In total though we received completed questionnaires from:

940 overseas students
530 British students
559 staff concerned with teaching these.

These were completed in respect of the following number of Centres:

43 post-graduate
61 undergraduate
39 'A' level.

The questionnaire thus provided us with quite a wide survey of the activities students might have to engage in as part of their study across subjects and levels, and in addition it gave us a reasonably large population on which to base conclusions concerning the extent to which students had difficulty in carrying out certain tasks in the academic context. For, in addition to information concerning activities, we were also able to establish in general terms through the questionnaire, where the major problems were likely to occur for students coping with these activities in the academic context.

It must be remembered though that degree of difficulty does not necessarily correlate with frequency of activity, or with the importance of that activity for the course in question. The amount of time spent on a particular activity, the 'use hours', as Straker-Cook (1977, p.45) termed it, may not be the only important criterion for:

"Even if reading and speaking, say, were equally demanded in terms of use hours, they would present disproportionate difficulty to a student who has a reasonable competence in receptive skills, but little experience in sustained oral communication."

It may also be the case that relatively infrequent activities are very important and also a major source of difficulty. We must look both at the relative frequency of each activity across subject boundaries and levels, as well as considering how much difficulty this has occasioned for each group of students. While it makes no sense to look at overall aggregate frequency of activities, as different groups are disproportionately represented in the totals, it is possible and worthwhile to aggregate the degree of difficulty encountered in each of the activities (v. Appendices 3.4 and 3.5, pp.802-826). This gives us some indication of the relative amounts of difficulty caused by various activities. By weighing this in evidence with the frequency of an activity for the different subject groups at the different levels, and with the relevant incidences of difficulty encountered in performing this activity, we have the basis for making decisions concerning the desirability of taking a factor into account in our test task design.

3.4.2 Listening Comprehension Activities

3.4.2.1 The Nature of Listening Comprehension

Candlin et al. (1976, p.52) pointed out that as well as being concerned with linguistic descriptions i.e. what we hear, in order to identify the comprehension skills which the overseas student must acquire, we must also consider how we hear. They quote Gurney (1973, pp.96-97):

"Normally in sustained communication the hearer 'loses' between 20% and 40% of an utterance."

and emphasise that:

"We are not aware of any linguistic description that takes that fact into account, it emphasises just how far the 'ideal speaker-hearer' of some linguists' theories is from reality. It also emphasises the importance for the foreign student of understanding not only the syntactic system of the language, but also the discoursal and cohesive systems through which the less-than ideal speaker can guide the errant attention of his hearer to the message elements he considers most important."

It does not matter that we lose part of the soundstream for, as native speakers, we usually only need to understand part of what is said to us to cope satisfactorily. Shannon et al. (1959) estimated that 50% of the English language is redundant in this way. Mastery of the language system brings with it the ability to exploit this redundancy of the code, but many overseas students encounter difficulties because of their inability to do this. They are further hampered by the fact that performance constraints often interfere with speech transmission. Environmental noise, rapid speech, heavy accent can normally be coped with by the native speaker who can supply the lost information from the signal on the basis of his knowledge of the language. The overseas student is unable to exploit this redundancy with the result that his immediate memory systems are overloaded and he cannot retain enough of the signal to process it (v. Miller 1956).

Kelly (1981, p.174) argued that the 'size' or amount of the discourse to be coped with is a performance constraint of some significance for proficiency assessment. Performance on a five minute taped lecturette might not be indicative of an ability to cope with a fifty minute academic lecture where the student has to retain far more information and cope with a possibly more complex coherence structure. In addition there is a difference between discourse spoken ex tempore as in a real lecture and that of the spoken version of a short written text normally to be found in listening tests. The length of the discourse may cause fatigue and affect the attention span with a resultant fall off in comprehension. Thus Ingram's (1968) advocacy of short tests free from any memory load would seem at odds with the demands that are placed on students in the real life situation.

Candlin et al. (1976, p.54) rightly stressed that the task for the overseas student is not to understand isolated, idealised articulations of certain phones, but to learn to hear phonemes as invariant over different words and produced by different speakers. They argued:

"... in first language acquisition this does not require the child's learning to isolate specific sounds (Turner 1975, p.105) and it seems questionable as an assumption for language teaching."

Speech recognition does not take place in terms of minimal units. We can extend this to language testing for, as Wanner (1973) suggested, since adult native speakers are not very good at identifying isolated words in their first language so it would seem unreasonable to expect the foreign language learner to achieve better results on this sort of exercise.

Neither is listening comprehension the passive skill it is often characterised as, but rather it requires active interpretation on the part of the hearer. Lieberman (1965, p.41) noted that in an experiment he carried out with linguists on intonation:

"... the linguist often considers his 'subjective' judgement and fills in the Trager Smith pitch notation that is appropriate to the structure of the sentence, which he usually infers from the words of the sentence and his knowledge of the language."

Thus the way we interpret intonation may depend largely on information from our perception of the syntactic form of a message. It is not sufficient to hear to be able to understand and as Candlin et al. (1976) suggested, contexts of use may be very influential in successful comprehension. This is perhaps an argument in favour of exposing the student to samples of 'authentic' lecture discourse rather than the 'classroomese' of his former E.F.L. experience. The problems of approximating to 'authentic' lecture discourse in the test situation, especially as regards ensuring the presence of appropriate visual information, are of course not inconsiderable.

Many factors affect the extent to which overseas students will experience problems in listening comprehension. The length of time they have been here is one important consideration. Research has shown (cf. Davies 1965; Chaplen 1970; Jordan et al. 1973; Morrison 1974 and Hawkey 1982) that listening comprehension problems are likely to decline after a period of residence in this country with exposure to the everyday English as spoken by people in Britain. However, as Straker-Cook (1977) pointed out, attendance at initial lectures and talks and participation in seminars and discussion groups places a heavy demand on oral/aural skills in the earlier stages for the overseas student. Jordan et al. (1973, p.43) found that although listening comprehension problems declined amongst

their overseas students from an initial incidence of 70%, in students' statements of difficulty they were still mentioned by 39% even after one term's exposure. That they exist to the extent that our questionnaire returns suggest, six months after the courses have started, gives cause for some concern.

The degree of difficulty experienced is obviously dependent on the total context of situation the student operates in. The clarity and loudness of the addresser's speech, the coherence and cohesion of the discourse, the accents, the speed, the number of participants, the amount of interference and other performance constraints (v. Kelly 1981) will all affect the individual's ability to operate. The degree to which comprehension difficulties affect successful transmission can be mitigated if a range of supportive techniques, e.g. duplicated handouts, outlines on the blackboard, are employed. All these techniques would serve to reduce the load on the aural processing capacities of the student.

Overseas students may have more problems in seminars and small group discussions than in lectures. For in multiparticipant interactions, as Kelly (1981, p.174) observed:

"Not only is there the matter of adjusting to the different voices, accents, rates and styles of speech quickly enough to follow the thread of the discourse. There is the further problem for the S.F.L. learner that a discourse built up from the contributions of different participants can be expected to be more disjointed and subject to more digressions and false starts, than that of a single speaker. Overlapping contributions will also further confuse and trouble the learner."

We can identify three orders of error (v. Straker-Cook 1975). Firstly those occurring in the decoding process due to faulty auditory processing, e.g. phonemic discrimination, speech rhythm, stress patterns, unfamiliar grammatical patterns, idiomatic expressions, regional varieties. A higher order of error could be seen as incomplete understanding of discourse organisation, the student missing many of the signals of transition and discourse markers, perhaps oblivious to the effect of certain cohesive devices or insensitive to the lecturer's manipulative strategies.

Even if he copes at these two levels, error may still occur as he might be so taken up in decoding and comprehending that he loses track of the information itself; he understands the words but does not follow the lecture. Rivers (1966, p.199) accurately described this phenomenon:

"... he may recognise the essentials of the message but not be able to remember what he has recognised. This is because he is unable to concentrate his attention on the crucial elements of the message long enough to rehearse them sub-vocally before moving on with the continuing voice. All his attention is taken up with recognition."

3.4.2.2 The Listening Comprehension Tasks Students Perform in the Academic Context

3.4.2.2.1 Informing transactions

In this category we were concerned with how often teachers made statements of fact or principle: of problems, of hypothesis and speculation, and of experimental procedures, across the range of teaching situations in the subject areas and levels under consideration. Egglestone et al. (1975, p.15) had found, this category of transaction to be "the most frequently occurring one".

It became apparent during the observations we carried out (v. Appendix 3.2.2, pp.690-693) that especially in science and engineering classes it is difficult to separate verbal discourse from its often non-verbal context. In many instances a teacher's statements to a class serve merely to illustrate drawings, equipment or other material. Often teachers put material on the board and require the class to copy it and continue with further explanation as a form of running commentary while students are engaged in copying the material provided (v. Egglestone et al. 1975).

Candlin et al. (1976, p.21) in their detailed analysis of the text of a small number of engineering lectures discovered that the lecture:

"... constitutes an example of what Halliday and Hasan (1976) term 'language in action' where the text is likely to contain a high proportion of instances of exophoric reference."

They (1976, p.28) argued that in engineering lectures:

"The diagrams and comments or examples written on the blackboard by the lecturer form the core (if not the entire body) of the notes which students take away. The visual elements of the lectures are closely integrated with the spoken text and have a crucial part in the development and explanation of the information exchange. They contribute to clarifying reference and exemplification in the text making much of the reference exophoric."

Hutchinson et al. (1981, p.6) made a similar point in their study of practical demonstrations at the technical level:

"The importance for the discourse of the presence of a visual display can be easily appreciated. *Language ceases* in effect to be the primary element in the communication: it takes on an interpretative role - explaining, highlighting or contextualising what is visually observable. It is only meaningful when related to the visual ... the visual display and the actions of the teacher relating to it provides the structure for the discourse."

and (p.60):

"... the visual display carries a large proportion of the meaning load. Because the speaker can rely on the visual to provide the specific meaning of what he is saying, the language of the practical demonstration lacks the precision of reference and the rigorously formalised structures traditionally associated with technical discourse."

As a result one finds that in these periods there is extensive use of reference visually accompanied by appropriate gestures or actions which are often the only means of clarifying the frequent, verbally unmarked shifts of reference between text and visual display.

Hutchinson et al. concluded (p.61) that in practical demonstrations:

"With a visual display as the focus of a demonstration there is no need for the teacher to strive for self sufficiency in the verbal text. Vocabulary and form can be varied because a gesture, an action or simply the visual presence of the referent will make the meaning precise. In toto then the language of technical instruction avoids the specific terms and structures that characterise written technical discourse in favour of a more conversational mode, in which non-verbal clues are of paramount importance."

TABLE 3E

THE FREQUENCY OF OCCURRENCE IN OBSERVATIONS OF
TEACHER MAKING STATEMENTS:

1. of fact, principle

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N

2. of problems

3. of hypotheses or speculation

4. of experimental procedure

- a. lectures
 - b. seminars/tutorials
 - c. practical classes
- H: High frequency of occurrence
M: Medium frequency of occurrence
L: Low frequency of occurrence
N: Non-occurrence

- Eng. U Engineering Undergraduate
- Sci. U Science Undergraduate
- Sci. P Science Post-graduate
- Sci. A Science 'A' Level
- S.Sci U Social Science Undergraduate
- S.Sci P Social Science Post-graduate

Whenever a box is shaded in, it represents the degree of frequency noted for that activity in at least one observation. A range of frequencies represents either:

- i) the differential occurrences of that activity in a number of observations in the same or similar programmes. (v. Appendix 3.1.2), or
- ii) the opinion of the lecturer(s) involved as to the general frequency of that activity in similar situations.

Because of restrictions on length and the difficulties foreseen in getting staff accurately to introspect on the frequency with which they employed the various types of statement under consideration, we decided not to seek details of their use of these in the questionnaire.

During the observations conducted at the various institutions in the further and higher education sectors (v. Appendix 3.2.2, pp.690-693) we attended a large number of lectures, seminars/tutorials and practical classes. We noted that informing transactions accounted for a great deal of the activity that took place in the lectures and seminars observed (v. Table 3E above). The most common of these were: the transmission by the teacher of information concerning fact and principle, and to a lesser extent, the statement of problems arising out of the particular subject matter. In practical classes (where these were applicable) statements of this type decreased and there tended to be a higher occurrence of statements of hypothesis or speculation and particularly of experimental procedure.

The observations attested to the central role of information displayed on the board in a large number of science and engineering lectures as well as to the importance that must be placed on the reference to physical objects such as scientific and other equipment especially in practical classes in these areas. In the social science teaching situations there was generally a lower level of exophoric reference and thus without the visual support available in many lectures in science and engineering, the verbal text here was by necessity more self-sufficient and structurally precise. It would have been helpful with hind-sight to have introduced an additional category in this section of the observation schedule: 'teacher makes statements reinforced by demonstration', as this would have enabled us to show more clearly the relative incidences of teachers giving different types of information while using demonstration as an aid, as against teacher simply standing before pupils and telling them the information.

Given the importance of this physical context Candlin et al. (1976, p.39) referred to some possible implications:

"The results of our examination, however incomplete, support the view that the lecture is an interactive situation. They emphasise that in this style of lecture delivery, the integration of text, visuals and gesture does not permit separation of the various elements ... simple audio recordings for use with a tape recorder or in a language laboratory will be inadequate."

If students were only to hear a text on tape without visual support, reference made to factors in the context of situation would be difficult to pick up. The integration of speech and visual materials would thus appear to be a significant factor to be borne in mind in the construction of a listening comprehension test for scientists and engineers in particular. As McEldowney (1976) pointed out, the ability to deal with closely integrated and often non-verbal devices also used for communicating information is an important element in the spoken as well as the written medium. Rivers (1966, p.198) noted:

"Over and above the clues provided by sound sequences we convey further elements of meaning by body movements, facial expressions, slight changes in breathing, length of pauses and degrees of emphasis."

She argues that:

"... without taking paralinguistic and kinesic features into account as further delimitation of the message, no oral communication is complete."

3.4.2.2.2 Directing transactions

The concern here is with the frequency and type of directions the teacher gives the students. These are instances of where the teacher tells the students to do certain things and student behaviour confirms they are acting upon his instructions, e.g. if the teacher puts information on the board, tells the students to write it down, and the students then copy this out, it is categorised as a directive as long as the teacher only comments infrequently during its occurrence. In all cases "where equipment or other illustrative material is used as a source of information to which students make extended reference unaccompanied by teacher comments", these are to be counted as resulting from teacher directives (v. Egglestone et al. 1975, p.23). Directives usually involve some use of the imperative though this form function correlation is inevitably variant.

TABLE 3F1 SUMMARY OF THE RETURNS TO THE QUESTIONNAIRE CONCERNING THE FREQUENCY OF TEACHER GIVING INSTRUCTIONS TO LOOK AT VARIOUS SOURCES OF INFORMATION, e.g. books, handouts, diagrams and scientific or other equipment. (Directing transactions) in: QuAla- lectures, QuAlb- seminars/tutorials and QuAlc- practical classes.

	Qu Al a				Qu Al b				Qu Al c			
	N	S	O	N/A	N	S	O	N/A	N	S	O	N/A
Eng. U	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
Eng. P	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
Sci. U *	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω
Sci. P	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
Sci. A **	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
S.Sci. U	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω	Ω
S.Sci. P	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ

KEY

Overseas students' replies:

φ means 20% to 39%
 φφ means 40% to 59%
 φφφ means 60% to 79%
 φφφφ means 80% to 100%

British students' replies:

Δ means 20% to 39%
 ΔΔ means 40% to 59%
 ΔΔΔ means 60% to 79%
 ΔΔΔΔ means 80% to 100%

Staff replies:

Ω means 20% to 39%
 ΩΩ means 40% to 59%
 ΩΩΩ means 60% to 79%
 ΩΩΩΩ means 80% to 100%

N: Never
 S: Sometimes
 O: Often
 N/A: Not applicable

Eng. U. Engineering Undergraduate
 Eng. P. Engineering Post graduate
 Sci. U. Science Undergraduate
 Sci. P. Science Post-graduate
 Sci. A. Science 'A' Level
 S.Sci. U Social Science under-graduate
 S.Sci. P Social Science post-graduate

* Science undergraduates were mostly Mathematics students (v. Appendix 3.2.3 for breakdown of where replies came from)

** 'A' level classes divided into non-practical (lectures) and practical

We also included in this category those instances where the teacher referred the students to information sources not present at the time of utterance. Obviously we were unable to check whether students actually followed up these references, but that they understood the teacher expected them to conform to his directives was considered sufficiently important to include it under this category of transaction.

In the questionnaire both staff and students were asked in Question A1 to state how often instructions were given to look at various sources of information and their answers are summarised in Table 3F1 above.

In lectures and practical classes (where applicable) there was a high frequency of occurrence of these directing transactions reported by both staff and students, though in the cases of science and engineering there was a noticeably higher occurrence recorded in lectures and seminars at the post-graduate as against the undergraduate level. There seems to be a lower level of occurrence in general in the seminar/tutorial situations possibly because of the lower level of information transfer in these periods, which are ideally concerned with the application of learned skills and knowledge rather than the acquisition of such. This lower level of occurrence is not found in social science post-graduate seminars and tutorials where it is perhaps teacher directives to further external sources of reference which account for the higher incidence reported.

It was noted in the observations (v. Table 3F2 below) that a good deal of the reference to be found in the directing transactions of science and engineering lecture discourse was made to tangible, visible objects and equipment, or information (verbal and non-verbal) recorded on the blackboard.

TABLE 3F2
THE FREQUENCY OF OCCURRENCE IN OBSERVATIONS OF
TEACHER DIRECTING PUPILS TO SOURCES OF
INFORMATION FOR THE PURPOSE OF:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
1. acquiring or clarifying facts or principles						
2. identifying or solving problems						
3. making inferences, formulating or testing hypotheses						
4. seeking guidance on experimental procedure						
5. developing own opinion or line of argument						

KEY AS ILLUSTRATED FOR TABLE 3E

a. lectures
b. seminars/tutorials
c. practical classes

H: High frequency of occurrence
M: Medium frequency of occurrence
L: Low frequency of occurrence
N: Non-occurrence

In science post-graduate and social science lectures we observed that a good deal of reference was often made to sources of information not present at the time of utterance and this was similarly a marked feature of the social science seminars as well. This could perhaps account for the greater incidence of these transactions being reported by post-graduate as against undergraduate science and engineering students in the returns to the questionnaire discussed above.

It is noticeable that particularly in seminars/tutorials and practical classes that the staff generally recorded slightly higher frequencies of occurrence than the students did in most subject and level groupings. It must be remembered in reading these figures that staff had been instructed to limit their replies to cover only those classes in the courses they personally taught on the particular programme we had specified. Students were obviously distilling from a far wider range of experience gained from all the courses in a particular programme.

3.4.2.2.3 Eliciting transactions

In this category we were concerned with the frequency with which staff asked the students questions in the various teaching situations and the kinds of question they asked.

For the purpose of recording the results of the survey we have treated the 'A' level science non-practical class as the equivalent of the lecture at other levels, though defining it separately in the 'A' level questionnaire as the period mainly occupied with talk by the teacher, the students listening and usually taking notes. Students may ask as well as answer questions and there may be some discussion. The teacher may go through written work or questions prepared by student(s). The students may work through problems set by the teacher. The 'A' level non-practical class is far more directly interactive than lectures at the other levels with the teachers more obviously concerned with securing feedback on the extent to which the information transmitted has been successfully transferred.

TABLE 3G1 SUMMARY OF THE RETURNS TO THE QUESTIONNAIRE CONCERNING THE FREQUENCY WITH WHICH STAFF ASK STUDENTS QUESTIONS IN: QuD5a - lectures, QuD5b - seminars/tutorials and QuD5c - practical classes.

	QuD5a			QuD5b			QuD5c		
	N	S	O N/A	N	S	O N/A	N	S	O N/A
Eng. U	φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω	
Eng. P	φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω	
Sci. U	φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω	
Sci. P	φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω	
Sci. A	φ	φ					φ	φ	
	Δ	Δ					Δ	Δ	
	Ω	Ω					Ω	Ω	
S.Sci. U	φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω	
S.Sci. P	φ	φ		φ	φ				
	Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω	

KEY AS PREVIOUSLY ILLUSTRATED FOR TABLE 3F1

In the summary of the returns to Question D5a, in Table 3G1 above, the highest occurrence of teacher initiated eliciting transactions in lectures was not surprisingly in the 'A' level sciences. Though overseas social science undergraduate students report a fairly high occurrence here also, this result is partly due to the large number of overseas students, and the allowances made for them, on the two B.Ed courses which constituted a large proportion of the returns in this category (v. Appendix 3.3.2, p.796).

The highest frequency overall was to be found in the seminar/tutorial situation and in particular, social science students reported a very high occurrence here. The numbers of eliciting transactions were more limited in the practical classes partly because of the lower number of students who have these and staff who teach in them, but perhaps also as a result of the more limited interaction between staff and students as the latter tend to work very much on their own in these periods. It is interesting to note that, as we saw in Section 3.4.2.2.2 above, in both practical and seminar/tutorial classes the staff often claimed a higher frequency of occurrence for the transaction than the students. (It is perhaps worth reiterating that the students are talking about their individual participation in all the seminars and practicals they attend in a programme whereas the lecturers had been specifically instructed to talk only about the body of students in the classes they themselves taught, on the particular programme we specified.)

In Table 3G2 below we have summarised the relative frequency of different types of eliciting transactions as they occurred in our observations across subject areas and levels. In general they corroborate the evidence gathered in the questionnaires and clearly show the more extensive use of these transactions by 'A' level science staff. The observations of other subject areas and levels show the relative scarcity of this type of transaction in the lecture situation. The high frequency of occurrence in the seminar/tutorial situation shows that even in supposedly multiparticipant interactions, the teacher still has a strong controlling influence on the overall structure the discourse takes. In practical classes the nature of the questions alters and where questions occur they are most likely to be concerned with observation and interpretation of observed or recorded information.

TABLE 3G2

THE FREQUENCY OF OCCURRENCE IN THE OBSERVATIONS OF
TEACHER ASKING QUESTIONS (OR INVITING COMMENTS)
WHICH ARE ANSWERED BY:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
1. recalling facts and principles	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
2. deducing solutions to problems	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
3. making hypotheses or speculation	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
4. designing of experimental procedure	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
5. direct observation	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
6. interpretation of observed or recorded information	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N

a. lectures
b. seminars/tutorials
c. practical classes

H: High frequency of occurrence
M: Medium frequency of occurrence
L: Low frequency of occurrence
N: Non-occurrence

KEY AS ILLUSTRATED FOR TABLE 3E

3.4.2.2.4 Formality

We did not seek to gather information through the questionnaire on the formality of the language used in the various teaching situations because of the difficulties involved in standardising the respondents' understanding of the term. The results of the observations concerning this, set out in Table 3H below, must for similar reasons be treated with some circumspection as they are at best impressionistic.

Even within the same lesson there could be wide variation in the addresser's skill in, and care for, articulation and in the level of formality adopted in terms of explicit markers of both grammar and lexis. The very concept is difficult to assess with any degree of consistency and accuracy in that no comprehensive and exclusive criteria have been evolved for describing it in the spoken mode (cf. Joos 1967; Crystal et al. 1975; Quirk et al. 1972 and Kelly 1978). At best we can perhaps describe utterances as being relatively more formal or more informal than others.

Very few lectures might be described as extremely 'distant, rigid' (v. Quirk et al. 1972, p.25) or what Joos (1967, pp.39-41) termed 'frozen', as few lecturers confine themselves to reading out a transcript verbatim, what Halliday et al. (1964, p.91) regarded as 'a special case of written rather than spoken language'. Most lecturers only use a framework of notes and their lectures might be described as 'formal' rather than 'consultative', in Joos's terminology, in the sense of there being extremely limited participation on the part of the students, except in the case of 'A' level science lectures (non-practical classes). This is perhaps, to a certain extent, inevitable when a group being taught reaches a certain size (v. Joos 1967). The 'formal' nature of the lecture is also evidenced by the low frequency of questions asked by staff (v. Table 3G2 above) and by students (v. Table 3W1, p.282 below) during these periods. One unfortunate science student was severely rebuked when he proffered a comment in one science lecture by the words:

"That was a rhetorical question."

not having fully appreciated that in a formal frame, one does not make insertions.

However, during the observations we made, we were often struck by the 'informality' of much of the language that was employed in lectures and even more so in practical classes. Frequently a switch was made into a more informal register in lectures in an attempt to explain a difficult concept or when reference was made to something physically extant in the situation, e.g. a diagram on the blackboard or equipment on the bench. With the onset of this rapid, casual articulation of everyday English and the intrusion of elisions, assimilation, repetition and pause factors it is little wonder that some overseas students are unable to exhibit the same degree of comprehension as with more carefully spoken material (cf. Wijasuriya 1971; Winter 1971; Holes 1972; Edwards 1978 and Kelly 1981).

In the seminar/tutorial periods the style became more 'consultative' with the increased participation of the students in these activities bringing on additional problems arising out of multiparticipant interaction.

In practical classes Hutchinson et al. (1980) illustrated how the language employed by the staff lacks the precision of reference and the rigorously formalised structures normally associated with technical discourse (especially written). In lectures, where reference was made to an external object, present at the time of utterance, we encountered very colloquial pro-forms. Talking of symbols in a chemical equation one lecturer commented:

"This chappie goes walkies over here."

They were even more common in the discourse of the practical class where we also noted a looseness and variety in grammatical structures, a lack of precision in the kind of vocabulary employed; with many synonyms and often general vocabulary being preferred to specific technical terms, particularly where the supervisor felt the need to employ glossing techniques. Explanation was usually given not in vocabulary specific to the subject area, but in what Hutchinson et al. (p.14) described as 'the general language of the technologically-

aware consumer society'. Often the language students need to understand teachers' statements is not that of the academic text-book but that of the everyday media of western society (v. Hutchinson et al. 1980).

In Table 3H below we have summarised our tentative opinions on the extent of the use of informal language. It does show the more informal nature of the seminar/tutorial and practical classes in that an increased use of informal language on the part of the students was observed arising from their increased participation in these activities.

The style seldom slipped into the familiar, 'casual' and 'intimate' categories described by Joos (1967, p.23), what Crystal et al. (1975) characterised as natural, everyday, informal conversation, and Quirk et al. (1972, p.25) as:

"... the intimate, casual or hearty - often slangy - language used between very close friends (especially of similar age), or members of a family, or used when a speaker feels for any other reason that he does not need to bother what the listener thinks of his choice of language."

This is more a feature of social interaction outside the classroom and only occasionally appeared in some practical classes where students were working closely together on a task with peers and especially with intimates.

TABLE 3H
THE FREQUENCY OF OCCURRENCE IN THE OBSERVATIONS
OF THE USE OF INFORMAL LANGUAGE:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
1. by teacher	a	b	c	a	b	c
	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
	a	b	c	a	b	c
	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
2. by student	a	b	c	a	b	c
	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
	a	b	c	a	b	c
	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
	a	b	c	a	b	c

a. lectures
b. seminars/tutorials
c. practical classes

H: High frequency of occurrence
M: Medium frequency of occurrence
L: Low frequency of occurrence
N: Non-occurrence

KEY AS ILLUSTRATED FOR TABLE 3E

3.4.2.3 Problems in Listening Comprehension

3.4.2.3.1 General problems in listening comprehension

The replies to Question A4 of the questionnaire are recorded in Table 3I below. In Part B staff were asked to indicate, for each group separately, the proportion of British and overseas students, on the courses they taught to first year groups in the programme specified, who appeared to experience difficulty in:

QuA4/1 understanding spoken description or narrative

QuA4/2 understanding spoken instructions

QuA4/3 understanding informal language

QuA4/4 understanding the subject matter of the talk.

We were able to ask the students more directly in Part A of Table 3I how much difficulty they experienced in each of these areas.

Overseas students reported having greater problems than their British counterparts in understanding spoken description and narrative (QuA4/1) and instructions (QuA4/2) across subject boundaries and levels. Slightly fewer overseas students had problems in understanding instructions than in following description and narrative and this was a view supported by the returns to the staff questionnaire.

Great difficulty for the overseas students, according to both staff and students, occurs in understanding informal language (QuA4/3), e.g. colloquialisms, idioms, slang, jokes. Over 10% of all the overseas students admitted to having 'a lot' of difficulty with this. A further 28½% of all overseas students admitted to 'some' difficulty. This was the category where there was the most significant difference between British and overseas students with over 80% of all British students experiencing 'no' difficulty at all (v. Appendix 3.4, pp. 803-804).

TABLE 31

PART A) SUMMARY OF RETURNS TO THE STUDENT QUESTIONNAIRE CONCERNING STUDENTS' ESTIMATES OF THE AMOUNT OF DIFFICULTY THEY EXPERIENCE IN THE GENERAL LISTENING COMPREHENSION AREAS OF:

	QuA4/1			QuA4/2			QuA4/3			QuA4/4		
	H	M	N	H	M	N	H	M	N	H	M	N
Eng. U	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/1- descrip- tion or narrative	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
Eng. P	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/2- instruc- tions	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
Sci. U	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/3- informal language	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
Sci. P	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/4- subject matter	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
U: a lot of difficulty												
M: some difficulty												
L: very little difficulty												
N: no difficulty												

PART B) SUMMARY OF RETURNS TO THE STAFF QUESTIONNAIRE CONCERNING STAFF'S ESTIMATES OF THE PROPORTION OF STUDENTS EXPERIENCING DIFFICULTY IN THE GENERAL LISTENING COMPREHENSION AREAS OF:

	QuA4/1			QuA4/2			QuA4/3			QuA4/4		
	H	M	N	H	M	N	H	M	N	H	M	N
Eng. U	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/1- descrip- tion or narrative	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
Eng. P	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/2- instruc- tions	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
Sci. U	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/3- informal language	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
Sci. P	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ	φ
QuA4/4- subject matter	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ
U: a lot of difficulty												
M: some difficulty												
L: very little difficulty												
N: no difficulty												

KEY AS
ILLUSTRATED
FOR TABLE 3F1

H: a lot
M: some
L: very few
N: none

One would accept that linguistic difficulty must, to some extent, add to the difficulty involved in absorbing a technical concept: because such concepts require very accurate use of language. However, as regards the understanding of the conceptual content of the subject as distinct from the language it is expressed in (QuA4/4), no clear distinction between native and non-native speakers emerges and British students admit to having as much and sometimes more difficulty. The overseas student does not see this as being as great a problem as the other more exclusively linguistic difficulties, whereas the British student admits to having greater difficulty with the subject matter per se than he does with the other more purely linguistic categories. In general the staff returns support the view that there is less of a divide between overseas students and British students as regards the difficulties encountered in this category.

It is obviously difficult for staff to assess the proportions of students who are experiencing problems, especially given the low level of interaction due to large numbers in the lecture situation (for all except 'A' level) and the limited use of the seminar/tutorial technique; thus staff returns must be treated with some caution. However, a 'don't know' category was provided and one might 'reasonably' assume that those without information such as the tutors who commented:

"I do not see them enough to know."

"... there are over 200 students attending the lectures. I cannot really answer this."

would have recorded their replies here.

Staff and students were also asked in Question A4/5 to comment on any other general difficulties students had experienced in listening comprehension.

Amongst the features they commented on were:

a) Length of discourse -

"... especially in lectures, their understanding and concentration decreases considerably after a while unless their fatigue is relieved. This is not a fault of their understanding but their concentration and (presumably) my lecturing."

b) Cultural factors -

"... students find difficulty in relating subject matter to own experience."

"... cultural backgrounds introduce difficulties e.g. when one has to study the use of a computer in applications appropriate to European culture."

c) Referential adequacy -

"Occasionally some non-technical words are new to them, which one assumes are commonly used."

"Particular words in normal speech that they do not know will result in a whole sentence or explanation being lost. Technical words are necessarily explained for the whole class."

"... for overseas students scientific vocabulary is often limited and the same word having two meanings, often one in common usage and the other technical, causes confusion."

d) Conceptual problems -

"Many students, particularly from overseas, have difficulty in abstracting from a physical system, the conceptual ideas by which it can be understood and analysed. This is particularly so when three dimensional concepts are required, or where relative motions take place."

Overseas students mainly referred to problems with:

a) Social English -

"When I speak, for example with my English classmates, I really find difficulty in understanding what they say especially if we speak about something not related to academic subjects."

"... difficulties in understanding other students rather than lecturers."

b) Lexis -

"The lecturer uses some words I don't understand."

c) Cultural adaptation -

"Sometimes I don't know or have not seen what is being described to have an idea of it, so I have to make rough guesses. For instance there are no canal locks in Nigeria, so I actually know little or nothing about them."

3.4.2.3.2 Specific problems in listening comprehension

An attempt was made in the student questionnaire (QuA6) to focus on the level of discourse at which the problems were occurring and so students were asked:

QuA6 - Please indicate how much difficulty you have in each of the following:

1. Recognising individual words in what is being said.
2. Recognising where sentences end and begin.
3. Understanding completely what the speaker is saying and linking this to what he has said earlier.

In addition we sought to gather information on the particular constraints that might have affected their performance so we also asked:

QuA5 - Please indicate how much difficulty you have in understanding your teachers or other students when:

1. They talk very fast.
2. They speak quietly.
3. Their accents or pronunciation are different from what you are used to.
4. More than one person is speaking, as in group discussions.
5. There are other problems which interfere with listening comprehension (please specify below).

The students' responses to these two sets of questions are to be found in Table 3J below.

As Table 3J illustrates, great difficulty for the overseas students as compared to the British students, occurs in recognising individual words (QuA6/1) (cf. Larter 1962; Chaplen 1970; Morrison 1974; Walker 1978 and Ryan 1979). The overseas students also find problems in following the discourse as it unfolds while relating this back to what has been said earlier (QuA6/3). The British students seem to have little difficulty in coping with either of these though they do have slightly more in understanding the whole discourse. Both groups have 'very little' difficulty in recognising structure at the sentence level (QuA6/2).

TABLE 3J

SUMMARY OF RETURNS TO THE STUDENT QUESTIONNAIRE CONCERNING THE AMOUNT OF DIFFICULTY THEY HAD IN UNDERSTANDING TEACHERS AND OTHER STUDENTS WHEN THEY:

	QuA5/1			QuA5/2			QuA5/3			QuA5/4			
	H	M	L	H	M	L	H	M	L	H	M	L	N
Eng. U	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Eng. P	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Sci. U	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Sci. P	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Sci. A	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
S.Sci. U	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
S.Sci. P	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ

SUMMARY OF RETURNS TO THE STUDENT QUESTIONNAIRE CONCERNING THE AMOUNT OF DIFFICULTY THEY HAD IN RECOGNISING:

	QuA6/1			QuA6/2			QuA6/3			
	H	M	L	H	M	L	H	M	L	N
Eng. U	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Eng. P	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Sci. U	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Sci. P	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
Sci. A	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
S.Sci. U	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ
S.Sci. P	φ	φ	Δ	φ	φ	Δ	φ	φ	Δ	Δ

QuA5/1 - talk very fast
 QuA5/2 - speak quietly
 QuA5/3 - have different accents or pronunciation than they are used to
 QuA5/4 - are talking in a group

QuA6/1 - individual words
 QuA6/2 - where sentences end and begin
 QuA6/3 - what has been said and relating this to earlier parts of the discourse

KEY AS ILLUSTRATED FOR TABLE 3FI

H: a lot of difficulty
 M: some difficulty
 L: very little difficulty
 N: no difficulty

The categories in Question A5 are not isolates and any combination of these performance constraints may give rise to difficulty. However, as the returns indicate (QuA5/1) it was generally the speed of the utterance which caused the greatest difficulty (cf. Sen 1970; Wijasuriya 1971; Morrison 1974 and Kelly 1981) for the overseas students and this ability to comprehend fast spoken discourse divides sharply native and non-native speakers. There is also a wide gap between the relative amount of difficulty experienced by the two groups when more than one person is speaking at a time, also noted by Rogers (1977), Mackenzie (1977) and Kelly (1981). Thus seminars and group discussions are examples of multiparticipant interactions which place heavy demands on the overseas students' language proficiency (cf. Crystal et al. 1975 and Kelly 1981).

Question A5/2 shows that native speakers have almost as much, and in some cases more, trouble than non-natives in comprehension when the discourse is delivered too quietly perhaps because the overseas student is used to paying closer attention to what is being said and is more often to be found nearer the front of the classroom than his British counterpart. Accent (QuA5/3) also caused a good deal of trouble for many overseas students and in a number of cases for the British students as well, an observation also made by Larter (1962), Sen (1970), Wijasuriya (1971), Jordan et al. (1973), Morrison (1974), Edwards (1978) and Walker (1978).

The students were asked to comment on whether there were any other problems which interfered with listening comprehension. Fewer British students made any contributions here and their replies only echoed the type of problem referred to by the overseas students. The following comments serve to illustrate the replies made by the overseas students concerning the various problems they encountered.

- a) Accent and speech - Students drew attention to the fact that it is the ability to comprehend a wide range of accents that is needed and this does not only refer to native speaker accents:

"... difficulty when I listen to somebody whose native language is not English or with strong local British accents."

"Here, not everybody speaks with B.B.C. accents."

"Some teachers in the Poly are Welsh or Scots. It is so difficult to understand them because of their accent and pronunciation. When I tried to think what a word he says really is I always missed out the other following words and the whole statement is meaningless."

"I don't find difficulty in what the teachers say during the class, but sometimes when one of the English students speaks with the teacher or asks any questions I find difficulty in understanding what he asks about."

One of the tutors also added this comment to his questionnaire:

"... they often find it difficult to understand each other when speaking in English. Most students (and staff for that matter) find it difficult to follow Indian speakers, mainly because they speak so rapidly (being fluent) with a 'strong' accent. This causes problems during discussion in class and particularly in group tutorials."

b) Adverse signal-noise ratio -

"Outside noise."

"Construction engineering department next to our block."

"Bad acoustics."

"Student noise while lecturer is trying to talk."

"Noisy chattering in the lecture."

"Unwanted side discussion by class members."

c) Teacher induced problems - The difficulties in comprehension can in some cases be compounded by the teaching staff themselves (v. Edwards 1978) and students referred to the following failings on the part of their teachers:

"The teacher is writing on the board and speaking at the same time."

"Purpose not defined from the beginning."

"Lecture disorganised."

TABLE 3K1

SUMMARY OF RETURNS TO THE QUESTIONNAIRE CONCERNING THE FREQUENCY WITH WHICH DUPLICATED NOTES WERE PROVIDED IN: QuA2a- lectures, QuA2b- seminars/tutorials and QuA2c- practical classes.

SUMMARY OF RETURNS TO THE QUESTIONNAIRE CONCERNING THE FREQUENCY OF OCCURRENCE OF OTHER METHODS OF TAKING NOTES: QuA3/1- copying, QuA3/2- dictation and QuA3/3- note taking from speech delivered at normal speed.

	QuA2a			QuA2b			QuA2c			QuA3/1			QuA3/2			QuA3/3		
	N	S	O/N/A	N	S	O/N/A	N	S	O/N/A	N	S	O/N/A	N	S	O/N/A	N	S	O/N/A
Eng. U	φφ	φφ		φ	φ		φφ	φφ		φφ	φφ		φφ	φφ		φφ	φφ	
	ΔΔ	ΔΔ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	
Eng. P	φφ	φφ		φ	φ		φ	φ		φ	φ		φ	φ		φ	φ	
	ΔΔ	ΔΔ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	
Sci. U	φ	φ		φ	φ		φ	φ		φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	
Sci. P	φφ	φ		φ	φ		φ	φ		φ	φ		φ	φ		φ	φ	
	ΔΔ	ΔΔ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	
Sci. A	φ	φ		φ	φ		φ	φ		φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	
S.Sci. U	φ	φ		φ	φ		φ	φ		φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	
S.Sci. P	φ	φ		φ	φ		φ	φ		φ	φ		φ	φ		φ	φ	
	Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ		Δ	Δ	
	Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω		Ω	Ω	

KEY AS ILLUSTRATED FOR TABLE 3F1

3.4.2.4 Note Taking Activities

As we have seen above, listening activities are complex and if students have to take notes as well the difficulties are compounded. Candlin et al. (1976, p.63) noted that the students they observed:

"... have to be able to receive a verbally and visually transmitted message; to decode and memorise parts of it after reception; to relate the newly received parts to the already perceived parts; to select from these the elements they are to re-encode; and when this has been done, the re-encoded parts are written or copied down."

Before students can begin to take notes they have to develop considerable skill in speech perception. They need to be able to separate the main points from supporting details, to understand the organisation of the argument, to match different comprehension strategies to the requirements of the text and be able to integrate information from a variety of sources, e.g. talk with visuals, written text, or diagrams.

In the questionnaire and through the observations we sought to establish how students gained a written record of the information transmitted. The findings are described below in Section 3.4.2.4.1.

3.4.2.4.1 Methods of acquiring a written record

In Question A2 both staff and students were asked how often duplicated notes were provided in a) lectures, b) seminars and tutorials, c) practical classes and their replies are recorded in Table 3K1 above.

Overall, in most disciplines and at most levels, the frequency with which duplicated notes were distributed is mainly in the 'sometimes' category. In lectures though a large number of engineering and social science undergraduate courses and science and engineering post-graduates recorded a high occurrence of the teacher providing notes in this form. In those subjects where there were practical classes, duplicated notes were 'often' provided. In seminars and tutorials the situation varied much more but in general there appeared to be less frequent provision of duplicated notes.

In general the staff estimates of the frequency of provision are lower than the students because they are only talking about the particular course they themselves teach on the programme the students are taking. The observations (v. Table 3K2, Section 4) showed that when they were used, duplicated notes could be a central feature of the information transmission, especially in engineering and science.

Staff and students were also asked in the questionnaire to specify the frequency with which the following note-taking tasks were performed:

QuA3/1 copying diagrams, charts, graphs, written notes, etc. from the blackboard.

QuA3/2 taking notes dictated by the teacher.

QuA3/3 extracting and re-encoding information from spoken discourse.

Their answers are summarised in Table 3K1 above.

Taking Question A3/1 first, in science and engineering lectures at all levels there is a very heavy incidence of copying information directly from the blackboard and/or the overhead projector transparency. This was not as frequent an activity for social science students. This picture is reinforced by the observations as can be seen in Table 3K2, Section 1, below. Comments made to me during the observations indicated that many lecturers in engineering and science were satisfied if the students only took away a set of notes accurately transcribed from the board.

One has only to compare normal university science and engineering lecture theatres with those of the social sciences to find a physical embodiment of this; the former normally possess multiple blackboards rising to the ceiling, whereas the latter often only have one single board.

The second category, dictation, was not a common feature in most teaching according to the questionnaire returns (Question A3/2, Table 3K1 above) except for some 'A' level science students. There is evidence, however, across subjects and levels that it is not a device as infrequently employed in the teaching situation as staff would claim. Again it must be remembered that staff are talking only of the particular courses they teach on the specified programme whereas students' replies are influenced by the whole teaching programme.

TABLE 3K2
THE FREQUENCY OF OCCURRENCE IN OBSERVATIONS OF
VARIOUS METHODS FOR GAINING A WRITTEN RECORD
OF SPOKEN DISCOURSE:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
a	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
b	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
c	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N

1. Copying diagrams, charts, graphs, written notes, etc. from the blackboard.

2. Taking notes delivered at dictation pace.

3. Extracting information from speech.

4. Using duplicated notes relating to the teaching situation.

KEY AS ILLUSTRATED FOR TABLE 3E

a. lectures
b. seminars/tutorials
c. practical classes

H: High frequency of occurrence
M: Medium frequency of occurrence
L: Low frequency of occurrence
N: Non-occurrence

In the observations (v. Table 3K2, Section 2) only at 'A' level was its frequent use noted, but here again there is evidence in all subjects at all levels of the rate of articulation being slowed in order to allow students to record a verbatim transcript of (often very small) parts of the discourse.

The third category of extracting information from discourse spoken at normal speed (v. Table 3K2, Question A3/3 above) was a common activity for some undergraduates and a lot of post-graduates in most disciplines, but especially social science. It was not such a frequent activity for 'A' level science students, undergraduate engineers and scientists because so much is copied from the board, particularly in engineering.

The observations (v. Table 3K2, Section 3) show a similar pattern with the greatest frequency of this form of note-taking occurring in social science lectures. It was not a frequent activity for most 'A' level science students, but for those taking biology 'A' level in the classes observed, the volume of subject matter to be covered for the examination meant that the teacher spoke at normal speed and expected students to extract and record information from this discourse. In fact, only a few of the students observed, e.g. mathematics undergraduates and 'A' level mathematics students, escaped having to perform this activity at all.

3.4.2.5 Problems in Note Taking

To try to describe more closely what was causing difficulty here apart from those problems in listening comprehension listed above, we asked students in Question A7 to indicate how much difficulty they had in:

QuA7/1 recognising what is important and worth noting

QuA7/2 being able to write down quickly and clearly all you want to

QuA7/3 thinking of and using suitable abbreviations

QuA7/4 organising the notes they took down so that they could understand them when reading them later.

The results are summarised in Table 3L below.

TABLE 3L

SUMMARY OF RETURNS TO THE QUESTIONNAIRE CONCERNING Qua4/5 (STAFF) THE PROPORTION OF STUDENTS ENCOUNTERING DIFFICULTY IN NOTE TAKING AND A7 (STUDENT) THE AMOUNT OF DIFFICULTY EXPERIENCED IN: Qua7/1 - recognising what is important and worth noting, Qua7/2 - being able to write down quickly and clearly all you want to, Qua7/3- thinking of and using suitable abbreviations and Qua7/4 - organising notes taken down so they could be understood when read later.

	Qua4/5				Qua7/1				Qua7/2				Qua7/3				Qua7/4				
	H	M	L	N	DK	H	M	L	N	H	M	L	N	H	M	L	N	H	M	L	N
Eng. U	φ	φ	Δ	Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ
Eng. P	φ	φ		Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ
Sci. U	φ	φ	Δ	Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ
Sci. P	φ	φ	Δ	Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ
Sci. A	φ	φ	Δ	Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ
S.Sci. U	φ	φ	Δ	Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ
S.Sci. P	φ	φ	Δ	Δ		φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ	φ	φ	Δ	Δ

H: a lot
M: some
L: very few
N: none
DK: don't know

Amount of difficulty -
H: a lot
M: some
L: very little
N: none

KEY AS ILLUSTRATED FOR TABLE 3F1

We also asked the staff in Question A4/5 what proportion of the students currently taking the courses in the programme specified had difficulties with note-taking. Obviously unless they regularly checked the notes (as some did) their impressions are highly subjective, but it is interesting to observe (v. Table 3L above) that they felt that British students had only slightly less of a problem with this in general and almost as much in the case of science 'A' level, engineering and science undergraduate, students.

In all categories (QuA7/1-4) the overseas students admitted slightly greater degrees of difficulty than the British students though normally most saw themselves as having only slight difficulty.

Both overseas and British students seemed to experience little difficulty in organising the way they took notes down so that they could understand them later (QuA7/4).

In recognising what is important and worth noting (QuA7/1) overseas social science students, especially post-graduates, seemed to have less difficulty than overseas students in other categories probably in part due to the fact that they have to rely on this method of note-taking to a far greater extent than other students (v. Table 3K1, QuA3/3 above). Overall there was only a slight difference between the difficulties reported by British and overseas students.

Most overseas students ran into some problems in being able to write down quickly and clearly all they wanted to (QuA7/2) and in using suitable abbreviations (QuA7/3). It is in these two categories that there is the largest gap in relative amounts of difficulty reported by British and overseas students across the board.

In general one might conclude that there was, to some degree, a study skills problem for both British and overseas students here.

The students were also given the opportunity to comment on any other difficulties they had experienced in note-taking (QuA7/5), many pointed to the difficulties of listening, decoding, re-encoding and writing at the same time. We have made a selection from the comments of the overseas students:

"I miss some important things the lecturer says while making notes because I cannot write and listen at the same time."

"Simply because there is no time. You have to choose between understanding or writing in clear notes."

"To follow the argument and write it down at the same time."

"... we cannot write down what the lecturer says when he simultaneously uses the blackboard. So often in our notes we can only find what was written on the blackboard and we are missing other useful information."

Others took the opportunity to expand on the difficulties they had already been asked about earlier in the questionnaire:

a) Speed of the utterance (Question A5/1, Table 3J)

"The lecturer speaks too quickly and I can't catch up."

"Lecturers do not allow enough time for notes and diagrams to be taken down."

"Most lecturers become faster when they read out quotations."

b) Trying to take too much down (Question A7/1, Table 3L)

"Trying to write down every important point the lecturer said."

"Difficulty in taking down on paper almost everything given in the lecture."

"Sometimes the subject is so dense that you have to write down almost everything that is said by the lecturer, and that's a task I can't tell you."

c) Problems with lexis (Question A6/1, Table 3J)

"... some difficult words the lecturer did not write on the board and I did not know how to spell them."

"I don't understand some words the lecturer uses and therefore it slows my ability to take notes."

d) Discourse organisation (Question A6/3, Table 3J)

"Subject matter not often presented in logical order."

"Lack of stress by lecturer on importance of items."

"Teachers don't specify the main points."

"... too much rambling."

"... the build up of the lecture remains obscure."

One student put the problem of taking notes succinctly (Question A7, Table 3L):

"Keeping up is sometimes hard due to the fact that not only is the subject in question hard to understand, but there is the added problem of understanding the language and the problem of speed."

3.4.2.6 Constituent Enabling Skills

On the basis of the investigations described above and a survey of the literature, we drew up a list of the enabling skills that we might wish to test discretely in addition to the procedures we might adopt for assessing listening comprehension in an integrated framework with reading and writing. Munby (1978) proved a valuable informing source in this exercise.

1. Deducing the meaning and use of unfamiliar lexical items through understanding word formation and contextual clues.
2. Understanding relations between parts of a text through cohesion devices especially grammatical cohesion devices such as reference.
3. Understanding relations between parts of a text by recognising indicators in discourse especially for introducing, transition and conclusion of ideas and for anticipation of objection or contrary view.
4. Understanding the communicative value of sentences and utterances with and without explicit indicators, e.g. definition, example.
5. Understanding conceptual meaning, e.g. comparison, degree, cause, result, purpose.
6. Skills concerned with understanding and meaning, especially the ability to recognise the speaker's attitude towards the listener and topic of utterance, as conveyed mainly by intonation.

7. Distinguishing the main idea ~~for~~ supporting detail, e.g. by differentiating the whole from its parts, fact from opinion, statement from example, a proposition from its argument.
8. Understanding explicitly stated information.
9. Understanding information in the text not explicitly stated, e.g. through making inferences.
10. Interpreting text by going outside it: relating information in the text to information not contained in the text, e.g. through picking up exophoric reference.
11. Skimming - (a) listening to obtain the gist,
(b) listening for specifics.
12. (a) Extracting salient points to summarise the whole text, a specific idea or topic, the underlying idea or point.
Selective extraction of relevant key points from a text especially involving the co-ordination of related information.
- (b) Reducing text through rejecting redundant or irrelevant information or items especially determiners, repetition, compression of examples, use of abbreviations, use of symbols denoting relationships between states, processes, etc.

3.4.3. Reading Comprehension Activities

3.4.3.1 The Nature of Reading Comprehension

Oller (1972a, p.313) commented:

"Of all the skills required by students of English as a second language, surely none is more important to success in college level coursework than the ability to read at a reasonable rate and with comprehension."

In a recent survey of 200 American staff in a variety of disciplines, A.M. Johns (1981) found that they considered reading and listening skills as being the most essential to non-native speaker success in university classes. The National Association for Foreign Student Affairs (N.A.F.S.A.) carried out a study into what students from developing countries considered the most important language skills in their situation. The N.A.F.S.A. survey (Lee, M.Y. et al. 1981) reported that reading text-books and journals, understanding spoken English and writing papers and a thesis were generally considered to be the most important skills. They considered that oral interactional linguistic skills were of lesser importance and they rated the skills to converse with faculty and other students far lower than listening, reading and writing. James (forthcoming) in an extensive survey of lecturers' attitudes in Britain, found a similar situation with spoken ability being considered as far less important than reading, listening and writing abilities.

The value of reading activities is well documented. Mackay et al. (1975, p.44) pointed out:

"The growing stress on reading skills in English in various parts of the world ... is understandable in the light of the fact that most scientific research is published in English whether or not the native language of the authors is English."

and Mackay et al. (1974) provided evidence of the widespread use of English as the vehicle of a body of scientific and technological information where in many cases translations into the vernacular do not exist. D.N. Wood (1967) showed that over half the world's scientific and technological literature is published in English.

Thus in global terms and in terms of the specific needs of students arriving to pursue courses of study through the medium of English, reading is an activity which, as we will show below, they will have to cope with as a major academic activity in all courses at all levels. There is a problem however, as Mackay et al. (1975, p.46) pointed out:

"Most so-called reading comprehension tests currently in use for foreign learners of English are based on no clear statement of exactly what they are measuring and are therefore extremely blunt tools for the job they are intended to perform. Nevertheless, they are used to pass or fail students according to their performance in the test."

It is beyond the scope of this work to analyse fully the psycholinguistic process which constitutes reading comprehension, but a short examination of the literature has been made in order to provide a background for our discussion of the reading activities the students in our survey were involved in and the problems they encountered in operating in that medium.

From an operational viewpoint reading comprehension might be seen as an increase in the amount of information an individual is able to exhibit as a consequence of reading a passage of verbal materials. Reading is also considered as a process of creative synthesis (v. Eskey 1973) or as Jones (1974, p.11) described it:

"... the content of a message (the conceptually structured display of information items from a knowledge store) is matched with the reader's existing state of knowledge by a set of conceptual processes capable of judging what is new and relevant so this can be extracted for storage in the long-term memory or knowledge store of the individual."

Goodman (1967) suggested that reading was a process which involved deriving meaning from a printed text through simultaneous use of grapho-phonetic, syntactic and semantic information, a 'psycholinguistic guessing game' involving interaction between thought and language (p.127):

"Efficient reading does not result from precise perception and identification of all elements, but from skill in selecting the fewest, most productive cues necessary to produce guesses which are right the first time."

Many writers stress this need to relate the linguistic value of what is read to its communicative value. J.B. Carroll (1972, p.13) argued that:

"... comprehension of a message is adequate or satisfactory to the extent that the language receiver apprehends, at least provisionally, whatever linguistic information is present in the message and is able to relate that information to whatever context is available at a given time."

Davies et al. (1974, p.167) similarly considered that it was necessary for the reader to understand:

"... the meaning of the linguistic forms and the communicative function they fulfil in the text concerned."

Rivers (1968) also made a division between comprehension of elements of the code and assimilation of the message.

Gray (1960), is quoted in Clymer (1972, p.60 et seq.) as stating that, an analysis of the evidence available showed that understandings, skills and attitudes common to most reading activities can be classified under four headings:

- (i) Word perception, including pronunciation and meaning.
- (ii) Comprehension, including 'a clear grasp of what is read'.
- (iii) Reaction to and evaluation of ideas the author presents.
- (iv) Assimilation of what is read through fusion of old ideas and information obtained through reading.

These aspects operate simultaneously and reading is considered by Gray as a 'unitary act'.

There seems to be a dilemma at this point which emerges in the literature. Is reading a single indivisible skill or can we break it down and test component parts of it? Dakin (1969) suggested that there are three elements involved in reading: recognition of the visual input, structuring of the input into meaningful stretches (first words then sentences) and then interpretation of this input. The relationships between these processes are seen as being complex,

often occurring virtually simultaneously with each involving the others. Though this tripartite division does not necessarily accord with psychological reality it is useful for purposes of analysis.

Mackay et al. (1974) also argued that reading is not a single skill but a process comprising a complex set of inter-related skills involving (p.7):

- "1) word recognition and the mastery of basic vocabulary and such technical or specialised vocabulary as may from time to time be required.
- 2) the ability to see in the material the structures of the sentences, paragraphs and longer passages that constitute the thought units.
- 3) the intelligence necessary to follow the thought development thus presented and make any relevant deductions, inferences or critical assessments.
- 4) the ability to concentrate on the reading task."

Though reading might be an activity that is made up of a number of skills, both of a motor and a cognitive sort, Davies et al. (1974, p.155) warned:

"... it is as well to bear in mind that for the reader, whether hesitant or fluent, what he is doing is engaging in a single co-ordinated activity ... The analysis of reading skills has been very detailed, but we seem to be a long way from understanding the manner in which different features of the skill relate to form one process. We have had a good deal of analysis but little synthesis."

This is a view supported by the Schools Council Project described in The Effective Use of Reading (Lunzer et al. 1979) which found that reading could not be broken down into a number of distinct sub-skills with the evidence in fact pointing strongly to a single aptitude.

Until more evidence is available which shows how the different skills relate to form the reading comprehension process, it would seem prudent not to rely solely on tests which merely evaluate performance in isolated reading skills which are believed, in some way, to contribute to the overall comprehension process, but to include also measures for assessing the latter in a more integrative fashion.

3.4.3.2 The Reading Tasks Students Perform in the Academic Context

In both staff and student questionnaires we asked how often the students on a particular programme had to perform various reading tasks. Details of these are given below along with staff and student responses to them. The nature of the reading tasks assigned to students was also the subject of scrutiny during the observational visits.

In the questionnaires we asked how frequently students carried out the following tasks:

- Question B1
- 1) Reading carefully for comprehension of all the information in each of the following:
 - 1.1 duplicated notes given to you by the teacher
e.g. photocopies, printed notes, etc.
 - 1.2 written questions done either in class or
at home
 - 1.3 laboratory worksheets
 - 1.4 examination questions
 - 1.5 some text-books: whole or part
 - 1.6 any other (please specify below)
 - 2) Reading to get a general idea of the main information about a topic, e.g. general background reading, as follow up to lectures or in preparation for seminars, etc.
 - 3) Search reading to get information specifically required for particular written assignments, e.g. for homework tasks, project work, etc.
 - 4) Critical reading to establish and evaluate the author's position on a particular topic
 - 5) Reading to check sources of new information such as articles in recent journals, new books, etc. to see how useful they are to your course of study
 - 6) Reading quickly to find out how useful it would be to study a particular text more intensively

The results of the returns to this part of the questionnaire are summarised in Table 3M1 below.

Question B1/1 Reading intensively for comprehension of total text in:

1.1 duplicated notes relating to the teaching situation - The greatest frequency of occurrence of this type of reading activity was in science and engineering at both undergraduate and post-graduate level. For science 'A' level and social science students it occurred less frequently though still quite heavily in the 'sometimes' category. Very few students recorded that they 'never' received these.

The findings from the observations (recorded in Table 3M2 1.1 below) confirm the results of the questionnaire. In engineering undergraduate and science undergraduate and postgraduate classes, this activity occurred far more frequently than it did in the social sciences or science 'A' level (particularly mathematics).

1.2 written questions done in class or assigned for homework - For all disciplines, at all levels, this was an activity which occurred 'sometimes' or 'often' for most students according to the questionnaire returns.

The observations (Table 3M2 1.2 below) showed a mixed result with some science undergraduates and social science students not having to do this at all in some of the limited number of classes we observed.

If we link this category with the returns to question B1/1.4, the comprehension of examination questions, then we could argue that at one time or another this form of intensive reading activity is an important one for all students.

TABLE 3M2 THE FREQUENCY OF OCCURRENCE OF VARIOUS READING TASKS ESTABLISHED ON THE BASIS OF INFORMATION GATHERED DURING THE OBSERVATIONS

REFERENCE STUDY	ENG. U	SCI. U	SCI. P	SCI. A	S.SCI U	S.SCI P
1. Reading intensively for comprehension of total text	H M L N	H M L N	H M L N	H M L N	H M L N	H M L N
1.1 duplicated notes related to the teaching situation						
1.2 written questions done in class or assigned for homework						
1.3 laboratory worksheets						
1.4 examination questions						
1.5 prescribed texts						
2. Reading to extract main information from the text to get a general idea of a topic						
3. Reading to extract specific assignment oriented information						
4. Reading to establish and evaluate writer's position on a particular issue						
5. Reading for purpose of monitoring sources of new information and assessing relevance to course of study						
6. Reading to assess desirability of text for intensive study						

H: High frequency of occurrence
M: Medium frequency of occurrence
L: Low frequency of occurrence
N: Non-occurrence

KEY AS ILLUSTRATED FOR TABLE 3E

1.3 laboratory worksheets - Engineering undergraduates in particular seem to be faced with this form of reading comprehension (see also QuA2c, Table 3K1). The science results are more variable because of the different balance in the returns of overseas and British students on different courses. Those students studying mathematics (a large number of our respondents, as is shown by the detailed breakdown in Appendix 3.3.2, pp.794-799) do far less of this type of reading, but students in other subject areas within the sciences may make a lot of use of these as the science 'A' level and post-graduate returns show. However, unlike description or narrative, following written instructions is not an activity most social science students need to cope with.

These findings are largely borne out by the observations (v. Table 3M2 1.3 above), with engineering undergraduates in the degree courses observed, making the greatest use of practical class worksheets and social science students none.

1.4 examination questions - Obviously frequency is a difficult yardstick to apply here as even when they have to do it relatively infrequently, as was the case for post-graduate mathematicians at Exeter, inability to cope with this form of reading can be fatal in terms of academic success. It is a task most students have to perform at some stage in their programme even if it occurs very infrequently in the first year, as appears to be the case in the social sciences.

1.5 prescribed texts - This is a vital activity for all post-graduates especially in the social sciences and for undergraduates in this area as well. For 'A' level science, and science and engineering undergraduates, it is a less frequent activity, the majority of the replies falling in the 'sometimes' category. Overall there is a very low occurrence of replies in the 'never' column across disciplines and levels. This was seen by both staff and students as one of the most important reading activities for most students at least in terms of frequency of occurrence. Few students can escape performing this task.

These results are largely borne out by the observations (v. Table 3M2 1.5) which confirm that reading intensively in one form or another is a vital activity for all students at all levels in all disciplines.

Question B1/2 Reading to extract main information from the text to get a general idea of a topic:

This again is a task which most students have to perform 'sometimes', though some science and engineering undergraduates and science 'A' level students seem to manage to avoid it. It is particularly important for post-graduates especially in the social sciences and for undergraduates in that area as well.

Question B1/3 Reading to extract specific assignment oriented information:

This appears to be the most important extensive reading task for all students at all levels. It is very frequent at post-graduate level in science and engineering and even more so at both undergraduate and post-graduate levels in the social sciences.

It is interesting to note that this is an activity which students claim to do far more than the staff recorded. A similar discrepancy occurs in a linked form of reading activity, Question B1/6 concerning the frequency with which texts are read to assess their desirability for intensive study. This is perhaps to be understood in the context of the staff recording their replies only in connection with the students in the courses they teach in the programme specified. They are thus talking only about their own particular courses, whereas the students are talking about the overall programme and are exposed to a far wider set of teaching activities. It is also an activity that perhaps the staff are less likely to know about than some of the others.

The observations (v. Table 3M2, 3) support the view arising from the questionnaire returns that this search reading to extract information for assignments is the most important extensive reading activity for students particularly those in the social sciences.

Question B1/4 Reading to establish and evaluate writer's position on a particular issue:

This was not a frequent occurrence in science and engineering subjects though it did happen more frequently at post-graduate level according to the questionnaire returns. In general staff considered it hardly ever occurred at 'A' level or undergraduate level except in the social sciences; even at post-graduate level it occurred 'sometimes' rather than 'often'.

This is the picture we get from the observations (v. Table 3M2, 4) when only in the social sciences was any real frequency of occurrence of this activity noted.

Question B1/5 Reading for purpose of monitoring sources of new information and assessing relevance to course of study:

The returns in the 'never' column indicate that most science and engineering students do not have to perform this activity below post-graduate level. Even at the post-graduate level it is only really a frequent activity (in principle at least) for some social science students. It would seem to be an activity which overseas students claim to do more often than British students.

This picture is the one we also get from the observations (v. Table 3M2, 5).

Question B1/6 Reading quickly to assess desirability of text for intensive study:

Together with B1/2 and B1/3 this was an extensive reading task which a number of students across disciplines and levels claimed they performed frequently. Overall, the replies were more varied on this activity perhaps reflecting differences in preferred learning styles amongst the student body as a whole. Engineering and science undergraduates and science 'A' level students again recorded the highest number of entries in the 'never' column, and likewise social scientists in the 'often' column. The replies of the post-graduate social science students again indicate that extensive reading tasks form an essential part of their learning activities. This is borne out in the observations (v. Table 3M2, 6) where social science students, particularly at the post-graduate level, were thought to perform this task the most frequently.

It is interesting to note, as mentioned above, in connection with Question B1/3, that staff consistently thought that the students they taught did less of this type of reading than the students replying to the questionnaire claimed. This might be because of the lesser need for this in the courses one teacher contributes to a programme than the students are exposed to as a whole.

Both staff and students were asked in Question B1/7 in the questionnaire to specify any other types of reading tasks performed in respect of academic programmes. No new types of reading emerged in the responses, but the replies do provide an idea of the wide variety of different reading materials that students can be exposed to. We have quoted from their replies where these illustrate or supplement the description of the reading tasks we have outlined above. We have grouped them in their subject categories under the broad headings of intensive and extensive reading. The selections are not systematic and serve to illustrate rather than define reading material encountered across courses. They include contributions from both British and overseas students.

Engineering Undergraduate:

- Intensive - "specifications for surveys and drawings; administrative details, e.g. explaining course examinations etc.; computer printouts and error diagnoses; own and friends' notes taken in lectures; application forms for jobs."
- Extensive - "newspapers; periodicals/journals; reading about the social role of engineers; reading for pleasure; technical journals and bulletins."

Engineering Post-graduate:

- Intensive - "project briefs."
- Extensive - "reading to appreciate current events and trends in the economy and industry; other research dissertations; leisure reading; relevant papers, journals and articles; reference works; reports."

Science Undergraduate:

- Intensive - "reading proofs of theorems or studying worked examples relating to the course; lecture notes own and others."
- Extensive - "reading for general knowledge (New Scientist); newspapers; journals."

Science Post-graduate:

- Intensive - "old lecture notes from undergraduate courses; reading of notices for general information on campus life."
- Extensive - "review articles; periodicals; journals; manuals prepared by instrument suppliers."

Science 'A' Level:

- Intensive - "copied notes; own notes; revision notebooks, e.g. key facts; physics comprehension passages in examination papers; special notices; science dictionaries; reference books."
- Extensive - "manufacturers' manuals; fiction; newspapers; magazines."

Social Science Undergraduate:

- Intensive - "interpreting numerical or graphical information; indexes; library files; statistics; other students' notes; questions and specimen answers."

Extensive - "the business section of a good Sunday or daily newspaper; interest reading; past projects; other students' essays."

Social Science Post-graduate:

Intensive - "statutes, plans, policies, circulars; statutory instruments; case studies, notes from other students."

Extensive - "reading daily newspapers; novels and magazines; articles in journals; newspapers, periodicals; original source material."

3.4.3.3 Reading Difficulties Encountered by Students

In the final version of the questionnaire we decided that, given the receptive nature of this skill, it would be better to ask only the students where they experienced difficulty. Thus in Question B3 we asked students to indicate how much difficulty they had in each of the following (where applicable):

QuB3/1 Reading carefully to understand all the information in a text

QuB3/2 Reading to get the main information from a text

QuB3/3 Search reading to get information specifically required for assignments

QuB3/4 Critical reading to establish and evaluate the author's position on a particular topic

QuB3/5 Reading quickly

QuB3/7 Reading texts where the subject matter is very complicated

QuB3/8 Any other reading difficulties (please specify below)

and the results are recorded in Table 3N below.

Question B3/1 Reading carefully to understand all the information in a text:

Overall the overseas students experienced slightly more difficulty with this than the British. Few overseas students admitted to experiencing 'a lot' of difficulty although roughly a third admitted to 'some difficulty'. The majority of overseas students recorded 'very little' difficulty in this type of reading activity which, as the returns to Question B1/1 above show, is likely to be quite a

frequent activity for a lot of students and therefore any difficulty is likely to lead to some problems. Overall less than 20% claimed 'no' difficulty here.

Question B3/2 Reading to get the main information from a text:

Again the British students experience slightly less difficulty here than the overseas students and a majority of both consider they have 'very little' or 'no' difficulty.

Social science post-graduates, both British and overseas, would seem to have the least reading difficulties of any group with regard to all the questions in this section.

TABLE 3N SUMMARY OF RETURNS CONCERNING THE AMOUNT OF DIFFICULTY EXPERIENCED BY STUDENTS IN VARIOUS READING TASKS

H: a lot of difficulty M: some difficulty L: very little difficulty N: no difficulty

	B3/1				B3/2				B3/3				B3/4				B3/5				B3/7																			
	H	M	L	N	H	M	L	N	H	M	L	N	H	M	L	N	H	M	L	N	H	M	L	N	H	M	L	N												
Eng. U	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ						
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				
Eng. P	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ		
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				
Sci. U	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ		
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				
Sci. P	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ		
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				
Sci. A	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ		
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				
S.Sci. U	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ		
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				
S.Sci. P	φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ			φ	φ		
	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ	Δ				

KEY AS ILLUSTRATED FOR TABLE 3F1

Question B3/3 Search reading to get information specifically required for assignments:

The overseas students seem to experience more problems than their British counterparts in this task and the majority of their replies fall into the 'some' or 'very little' difficulty categories. Overall, under 20% claimed they had 'no' difficulty at all in this.

Many students reported a high frequency of occurrence of this activity (v. Table 3M1, QuB1/3) so that any difficulties encountered here might be important.

Question B3/4 Critical reading to establish and evaluate the author's position on a particular topic:

Both groups of students, overseas and British, experience a noticeably higher level of difficulty with this than they did in the previous reading tasks, though against this must be weighed the evidence (v. Tables 3M1 and 3M2 above, QuB1/4) that this is a relatively infrequent activity for all except social science students. It is interesting that overseas social science undergraduates admit to having a lot more problems with this as against other reading tasks and this can be put down to the difficulties involved as well as the frequency with which they perform the task.

Question B3/5 Reading quickly:

There is quite a difference here between the British students, who largely record 'very little' or 'no' difficulty, and the much higher numbers of overseas students admitting to 'some' or 'a lot' of difficulty. There is a connection with the difficulties experienced in Question B3/3 and given the call made upon the extensive reading skills evidenced above (v. Table 3M1, especially the returns to QuB1/2, B1/3 and to a lesser extent QuB1/5 and B1/6) there must be cause for concern in this area. Less than 20% of all overseas students claimed they had 'no' difficulty here.

Question B3/7 Reading texts where the subject matter is very complicated:

It is not surprising that about 60% of all students admitted to having 'some' or 'a lot' of difficulty here. British undergraduates in engineering and science experienced nearly the same amount of difficulty as their overseas counterparts in these circumstances. Very few students, either overseas or British, recorded that they had 'no' difficulty in this case.

Students were also given the opportunity (QuB3/8) to specify any other reading difficulties they had encountered. We have attempted to categorise the replies, made by overseas students only, which expand upon or add to the difficulties noted above. A much smaller number of comments were made by the British students and they did not raise any problems other than those referred to by overseas students below.

Lexis:

"Meaning of specialist terms."

"Difficulty with words not encountered before."

"Speed is affected by having to look up obscure words in appropriate dictionary."

"Understanding of specific biological or chemical terms when they are new to me or are in an unusual context."

Size:

"Difficulty in understanding long sentences."

"Too much reading to do. I think it would help if we were provided with notes so we didn't have to waste so much time on irrelevant information."

"Reading very large texts."

Complexity:

"Difficulty in reading abstract subjects like sociology text-books."

"Too literary articles."

"Where the passage is very dense i.e. too many important points in two or three lines."

"No pictures."

Operational:

"Books with very small print."

"When the book has no subject index."

"Poorly written books or articles."

"Questions are often badly worded so that it becomes difficult to interpret what the question means."

"Blackboard writing is strange."

"Reading the lecturer's handwritten notes."

Speed:

"There is not much difficulty in reading quickly but the difficulty is to understand what has been read if read quickly."

"Reading quickly and taking in relevant facts."

In the pilot version of the questionnaire we had also asked tutors whether overseas students found difficulty in carrying out written instructions in practical classes and a selection of their answers specifying the difficulties are recorded below. We felt that only tutors responsible for the practical classes would have experienced, at first hand, the reading problems overseas students might encounter.

Need for verbal support:

"... inability to obtain overall picture of practical from just written instructions and constant verbal clarification needed, e.g. in relating one instruction to a previous one which has been completed."

"The printed words - if I translate, or just repeat what is written it is usually understandable. They don't seem to trust their own reading ability."

Lexis:

"Usually it is only the meaning of some vocabulary."

"They sometimes can't understand textbook problems when either an unusual word is used, e.g. beam or ridge, or when situations unfamiliar to them are referred to, e.g. cricket, spin driers."

"Misunderstanding the particular use of language in a certain context."

Structural:

"Difficulty in understanding formal written English."

"Understanding long complex sentences and following an argument through."

Knowledge and Culture:

"Instructions for many experiments assume that the student has done something similar previously and also that he has practical experience (he often does not)."

"Their troubles generally stem from weak technical background. How to decide whether a valve is open or closed for example."

Procedural:

"My own handwriting causes problems."

Overall comprehension:

"Inefficiency through lack of precise interpretation of instructions."

"Experience in reading through and following an ordered list of instructions seems sometimes lacking."

"When the set of instructions are too long they can carry out a single instruction, but are often 'floored' by a set of them."

3.4.3.4 Importance of Note Making

In pilot versions of the questionnaire we asked staff and students whether they considered making notes from text-books important and to expand on the reasoning that lay behind their answers.

Constraints, such as length and the time it would take to complete, prevented us asking a similar open ended question in the final version of the questionnaire. We have attempted to categorise the positive replies we received to these earlier versions below, treating staff and student responses separately. As regards the staff returns, a majority of the replies in favour of note making were received from social science staff, but there were sufficient in other disciplines to warrant treating the staff replies together. It was perhaps only mathematics teachers as a group who did not see the need of this activity for their students, e.g.:

"No, solving problems is the essence of learning mathematics."

"They are given concise notes on each topic by means of handouts - textbooks are used mainly as a source of examples."

Obviously there was a great deal of variation in staff estimates of the importance of this activity even within the same department, though we would argue that it is possible in the comments we have recorded below to see some general trends in what is regarded by staff as being important in making notes on written source material.

Improved understanding:

"Enables them to fix ideas dealt with during lectures."

"Textbooks and other sources give perspective to the subject."

"In order that students have further information that may clarify information not fully covered in lectures, or not fully understood by the students."

"If the student fails to understand my lectures, he may be able to grasp the same material from a book."

Additional information source:

"Provides experience in extracting essential information which they may use for purposes of essay writing and revision."

"... to supplement ideas and fill in information they may have missed."

"The lecture course defines the material they are expected to know. Textbooks are advised to reinforce this material or present an alternative approach but not to act as a primary source of material."

"Very often notes are the only adequate source of detail a student will get as lectures have to be a broad sweep and it's important that students understand the background and detailed argument of people mentioned in passing in lectures."

"This is the principal means by which the basic lecture material is fleshed out and made applicable to a wider range of examples."

"Pressure of time obliges the lecturer to omit certain topics, which the students must then read up from the textbook."

"To amplify parts of the syllabus."

"Gives an awareness of alternative views."

"It adds to my notes and gives a basis for comparison."

"The textbook provides an alternative to my lectures not a copy of them."

"It is useful preparation for seminar work."

Pedagogic value:

"To teach them to be selective and reorganise information."

"To encourage the discipline of selection of information and original research of material."

"They read to develop an ability to summarise succinctly."

"The ability to extract and précis information is important, also valuable to collate information from different sources."

"In order to summarise for themselves the meaning of what is said, and to interpret it later in written reports and assignments."

"The course is far too lecture directed as it is. Without individual reading in a critical manner they can never develop intellectually."

"In post-graduate education private reading is the main and preferable mode of study, it is also the most rapid way for students to acquaint themselves in depth with their field of specialisation."

"It is a means of encouraging independent search for materials and acquisition of a habit of enquiry as opposed to spoon feeding and slavish following of lectures."

"They have to exercise a critical faculty and learn to compress essentials into understandable and useful forms."

"Extracting information is important in the sciences. Making their own notes helps them distinguish the important facts. Unfortunately many students simply copy extracts from books."

"Assists comprehension and memory."

"Useful for revision."

"As a means of condensing main points to provide a conceptual framework which summarises a topic and is easier to remember than full details, but also easy to expand, as I find that for many students, if they can remember the main headings of the framework, expansion in detail is relatively easy in exams, essays, tutorials or for general use in later life."

"It helps them to become familiar with scientific vocabulary."

A number of tutors made adverse comments, considering note-making unimportant for their particular group of students:

"They should have sufficient notes from their 'A' level classwork."

"No, not required, the lecture content is sufficient."

"They learn mathematics by doing examples."

"Books largely mathematical."

"Not at post-graduate level. Should be past that one."

"At a post-graduate level most students will have the basic knowledge acquired from text-books and will be acquiring additional information which need not necessarily be annotated."

"There is no suitable text-book for my course. I cover the necessary material verbally and provide handouts as needed."

We have attempted to lay out the overseas student replies in a similar fashion to that adopted for the staff replies above. As mentioned above, British students made far fewer contributions here and did not experience any other problems than those we have listed below from the replies of the overseas students.

Improved understanding:

"To clarify some concepts or to understand better with some other examples."

"Because sometimes I cannot follow the lectures so I have to read the book to understand the topic."

"Basic theory has to be understood. In writing notes one familiarises oneself with the theory."

"It helps to make the principles clearer and improves memory of fact."

"It gives me a chance to revise the lecture and try and form notes in my own way which I may understand better."

Additional information source:

"Sometimes what the lecturer says is only very brief. He has no time to cover everything on a particular topic during the lecture. The handouts also are very brief most of the time - only the important points are written. So it is essential to take notes from text-books to have a clearer and better understanding of the subject."

"To gain more information and different ideas and be able to keep these for future reference."

"By using a text-book we can go into more detail."

"Because I believe by doing this you broaden your knowledge on a particular topic."

"Some of the important definitions and theory have to be taken from text-books."

"Because the lecturers never dictate notes and the notes I take down during lectures are inadequate."

"To supplement poor lectures."

"A way of recording knowledge from the experience of others."

Value as a learning device:

"It is important because we learn more and usually they are a lot more thorough than lecture notes."

"Text-books show the way that materials should be presented and the spelling of new vocabulary."

"You get used to using strange technical words and phrases."

"Summarising concepts often helps to understand them better and to have a clearer and schematic use for them."

"To be able to see how things relate to each other as opposed to putting information into already demarcated categories."

"It is important for isolating key facts through an accurate précis of the material which provides a more useful tool for everyday writing and exam revision."

"Because if I don't take notes I forget what I studied after half an hour because it's not in my language, so I forget it very quickly."

"It is easier to memorise notes since it is in shorter form and in my own sentences."

"Being the only source of information which can be gone through without wasting time during the examination period."

Only a small number of the respondents considered it unimportant:

"Because it's a waste of time, after taking down points from the lecture."

"It is time and energy wasting because the text-books are still there to be consulted from time to time."

"No, because I don't think notes can give one enough information."

"Because it would mean a sort of copying the author, word by word into the notebook which of course doesn't make any sense at all."

"Some text-books are so dense that they are impossible to summarise."

"... because sometimes the information required for note-making on a particular topic may not be available in only one text. We don't have the time to read all of them and make notes."

3.4.3.5 The Frequency with which Students Make Notes from Text-books

In the final version of the questionnaire we sought to establish how often students made notes from text-books and the results are recorded in Table 30 below. Science and engineering undergraduates seem to perform this task the least. In the case of the former, this is partly explained by the large numbers taking mathematics among our respondents. The engineers, as we said above, received a lot of duplicated notes and took very few notes for themselves even in lectures, apart from what they copied off the blackboard.

Only in the social sciences are there large numbers recorded in the 'often' category: approximately 60% and 70% for undergraduates and post-graduates respectively. However, at least 20% of all engineering and science post-graduate and 'A' level science students claimed it was also a frequent activity for them. The majority of 'non-social science' students reported in the 'sometimes' category.

TABLE 30 NOTE MAKING ACTIVITIES

	B2			B3/6			
	N	S	O	H	M	L	N
Eng. U	φ	φ					φ
	Δ	Δ					Δ
Eng. P	φ	φ	φ				φ
	Δ	Δ	Δ				Δ
Sci. U	φ	φ					φ
	Δ	Δ					Δ
Sci. P	φ	φ	φ				φ
	Δ	Δ	Δ				Δ
Sci. A	φ	φ	φ				φ
	Δ	Δ	Δ				Δ
S.Sci. U	φ	φ					φ
	Δ	Δ					Δ
S.Sci. P	φ	φ					φ
	Δ	Δ					Δ

Question B2 Summary of questionnaire returns concerning frequency with which students make notes from textbooks

N: Never
S: Sometimes
O: Often

Question B3/6 Summary of returns concerning the amount of difficulty students experienced in making notes from textbooks

KEY AS ILLUSTRATED FOR TABLE 3F1

H: a lot of difficulty
M: some difficulty
L: very little difficulty
N: no difficulty

This was the picture we gained from the observations (v. Table 32, p.235 in Section 3.4.4 below) with social scientists claiming to make notes a lot though some science post-graduates and some of the science undergraduates (non-mathematics) had to do this as well.

3.4.3.6 The Extent of Difficulty F countered in Note-Making

We also asked the students how much difficulty they had met in making notes from text-books and their replies are also recorded in Table 30 (QuB3/6).

Overall the British students seem to have had less of a problem here, but the gap is quite narrow. The biggest difference between the British and overseas students occurs in science 'A' level, social science undergraduate and the post-graduate groups, who also had to cope with this task quite frequently. The least difficulties were experienced by the social science post-graduates who had to perform this task the most.

Overall, many overseas students would seem to experience 'some' difficulty in this area.

3.4.3.7 The Specific Problems Encountered in Note-Making

In the pilot questionnaire we were able to ask students what they found difficult in note making. Constraints such as length, and the time it would take to complete prevailed against us including this in the final questionnaire. Despite the comments being from a very limited sample, they do throw some light on the problems students encounter.

Picking out the main points:

"Deciding the most important points to copy out without including irrelevant details."

"It is difficult to decide which points are relevant."

"To summarise an argument."

Condensing into note form:

"To express in a concise but exhaustive manner all that I read."

"To write the ideas in our own words in such a way that no quality is lost."

"Trying to express the point in note form yet still maintain the same meaning."

"To put points into own words without copying straight from the book."

Unfamiliarity with lexis or subject matter:

"If the sentences contain vocabulary which I can't understand and if the subjects are not familiar I'm not sure which points to emphasize so I find it difficult."

"I can't understand what the passage really meant."

"I don't understand some of the words and I have to keep looking them up."

Size of text:

"My course is very broad in subject matter so I find it difficult to go through so many references."

3.4.3.8 Constituent Enabling Skills

On the basis of the enquiry described above and a survey of the literature we drew up a list of skills which would seem to be involved in academic reading comprehension tasks. As well as testing reading comprehension in an integrated fashion by linking it with listening and writing tasks we may well wish to test discretely a student's competence in these more individual constituent enabling skills.

We have concentrated on what Davies et al. (1974) have called the 'structuring' and 'interpretation' stages since if the candidate was not beyond the 'recognition' stage there would be little point in attempting any test of competence at this level.

We have set out below those skills which we feel were important to our target population in the light of the foregoing discussion. A list of further informing sources is attached to each skill.

A) Reading Skills

1. Reference skills:

- (a) Reacting appropriately to typographical features, e.g. punctuation, titles, headings, sub-headings (cf. Mackay et al. 1974 and Munby 1978).
- (b) Skills needed when selecting texts or books and deciding whether contents are relevant to needs i.e. establishing background ethnographic information, e.g. by use of table of contents, preface, index, bibliography (cf. Beard 1970; Geddes 1977; Munby 1978 and Hawkey, M. 1979).

2. Word perception, decoding:

Deducing the meaning and use of lexical items through understanding word formation and contextual clues. The concern is not with the specialist technical vocabulary of a particular discipline, these having limited and defined meanings, but rather with what we might call sub-technical vocabulary, high frequency context independent words occurring across disciplines and with what Martin (1976, pp. 92-96) terms 'academic vocabulary', words which have a common focus in research, analysis and evaluation; the activities which characterise academic work. As Perren (1963) noted a long time ago it is the vast, shifting, ill-defined mass of common words and structures which cause the real problems (cf. Davies 1944; Gray 1960; Perren 1963; Swales 1971; Clymer 1972; Quirk et al. 1972; Mackay et al. 1973; Swan 1976; Mackay et al. 1975; Martin 1976 and Hawkey, M. 1979).

3. Understanding relations within the sentence:

This especially involves an understanding of sentence structure, modification structure, negation, complex embedding (cf. Quirk et al. 1972; Mackay et al. 1973; Davies et al. 1974; Swan 1976; Munby 1978 and Widdowson 1980).

4. Understanding relations between parts of a text:

- (a) Through awareness of grammatical cohesion devices especially reference.
- (b) Through awareness of lexical cohesion devices especially lexical set/collocation (cf. Bormuth et al. 1970; Widdowson 1971; Wijasuriya 1971; Allen et al. 1974b; Davies et al. 1974; Jones 1974; Mackay et al. 1974; Sim 1974; Halliday et al. 1976; Brumfit 1977; Morrow 1980a and Widdowson 1980).

5. Understanding relations between parts of text by recognising indicators in discourse:

Recognition of 'indicators' (Munby 1978), 'clues' (Yorkey 1976), 'linking signals' (Leech et al. 1975), 'signalling devices' (Heaton 1970), especially those used for introducing an idea, transition to another idea, concluding an idea and anticipating an objection or contrary view (v. Munby 1978).

6. Understanding the communicative value of sentences with and without explicit indicators:

This includes what Davies et al. (1974, p.166) termed the 'modal, metalingual and contact functions' of certain linguistic devices (cf. Mackay et al. 1973 and Munby 1978).

Widdowson (1971) and Bates et al. (1976) argued that these functions are particularly relevant to the understanding of text in activities of a scientific nature, especially the way they are used to develop various methods of planning and organising information in expository language (cf. Widdowson 1971; Lackstrom et al. 1973; Jones 1974; Selinker et al. 1974; Bates et al. 1976 and Selinker et al. 1976a).

Munby (1978, pp.185-189) also offered a list of language micro-functions that can occur in this category and lists them under the following broad headings:

1. Scale of certainty.
2. Scale of commitment.
3. Judgement and evaluation.

4. Suasion.
5. Argument.
6. Rational enquiry and exposition.

7. Understanding conceptual meaning:

This involves in particular an understanding of quantity and amount, definiteness and indefiniteness, comparison and degree, time, location and direction, means and instrument, cause, result, purpose, reason, contrast, condition.

There is a great need among science and engineering students in particular to understand the ways in which these basic notions are expressed in English in their various grammatical and lexical realisations (cf. Strevens 1971a; Davies et al. 1974; Jones 1974; Leech et al. 1975; Dudley-Evans 1977 and Munby 1978).

8. Understanding explicitly stated ideas and information:

This gives rise to what Davies et al. (1974) termed 'direct reference questions' concerning details, main ideas, etc. (cf. Clymer 1972; Mackay et al. 1975 and Munby 1978).

9. Understanding ideas and information in a text not explicitly stated:

(a) Through making inference, e.g. concerning the context in which it was written, causes, reasons, conclusions, opinions main ideas.

(b) Through understanding figurative language (cf. Carroll, J.B. 1972; Clymer 1972; Mackay et al. 1973; Swan 1976; Munby 1978; Hawkey, M. 1979 and Widdowson 1980).

10. Separating the essential from the non-essential in a text:

Distinguishing the main idea from supporting detail by differentiating especially the whole from its parts, statement from example, fact from opinion, a proposition from its argument (cf. Dudley-Evans 1977; Munby 1978 and Hawkey, M. 1979).

11. Transcoding information presented in a non-linguistic form,
e.g. tables, graphs, diagrams:

These methods of presenting information are used frequently in many disciplines because they can convey information in a clear and concise way (cf. Swales 1971; Fortune 1977; Munby 1978; Hawkey, M. 1979 and Widdowson 1980).

12. Skimming:

Skimming a text i.e. not reading every word (cf. Fry 1963 and Beard 1970).

- (a) Surveying to obtain the gist of a text or a general impression (cf. Swan 1976; Munby 1978; Hawkey, M. 1979; Morrow 1980a; Wallace 1980 and Widdowson 1980).
- (b) Scanning the text to locate specifically required information on a single point, multiple points or complete topic.

Whereas surveying does not require close scrutiny of the text, in scanning we require both rapid reading followed by intensive study depending on 'size' of information sought (cf. Hawkey, M. 1979; Wallace 1980 and Widdowson 1980).

13. Note-making:

- (a) Extracting salient points for summary - This could be a summary of the whole text, a specific idea or topic in the text or the underlying idea or point of the text (cf. Beard 1970; Bright et al. 1970; Barrett's taxonomy in Clymer 1972; Munby 1978 and Hawkey, M. 1979).
- (b) Selective extraction of relevant points from a text - This could involve the co-ordination of relevant information, the ordered rearrangement of contrasting items or the tabulation of information for comparison and contrast (cf. Munby 1978; Hawkey, M. 1979 and Widdowson 1980).
- (c) Reducing a text through rejection of redundant or irrelevant information or items, e.g. determiners, repetition, compression of examples, use of abbreviations (cf. Heaton 1975b; Munby 1978 and Hawkey, M. 1979).

14. Critical evaluation:

Assessing the worth of a text and the way information in it has been organised and expressed (cf. Clymer 1972; Davies et al. 1974 and Mackay et al. 1975).

3.4.4 Writing Activities in the Academic Context

3.4.4.1 The Nature of Written Production

We are not concerned here with the conversational or idiomatic English necessary for social interaction or with the type of English found in literature, as neither is directly relevant to English medium study in the areas under consideration. McEldowney (1976, p.5) argued:

"What is considered to be central is a proficiency in the more expository, neutral, transactional type of English that is the medium of education in English speaking countries."

Brooks (1980, p. 4) made a similar point when she distinguished between the:

"... creative composition which is usually expected to be of a personal nature."

and the:

"... public, impersonal essay which is expected to be largely factual and instrumental ... This second essay has a much more clearly defined content and a more formal structure. It is usually designed to test the candidate's ability to present the appropriate information clearly and concisely."

Academic writing in the areas under review usually takes the form of institutional writing circumscribed by conventions. In many E.F.L. examinations one finds candidates being asked to produce writing of a more personal type: to write an 'essay', to write a letter to a friend, to recount or describe something that has happened to them; in each case providing an addressee out of their own heads and working out the imaginary social ramifications of the situation. Overseas students often have great difficulty in coping with these creative writing tasks, and fail to convey 'intentionality', a term

Searle (1965, 1971) used to describe expressing and communicating something in a particular situation to specifiable persons. We would argue that creative writing is not central to the needs of students we are concerned with, as it does not appear to serve any function in the academic context in which they operate.

Davies et al. (1974) noted that though writing is often the most difficult of the language abilities to acquire it is less often explained why this should be so. They isolate three stages in the process of writing: manipulation, structuring and communication. These correspond to the divisions of recognition, structuring and interpreting noted above in connection with the reading process.

At the manipulative stage there are difficulties for those students unaccustomed to writing in the Roman script.

At the structuring stage the wealth of text-books available on guided composition (cf. Spencer 1967; Alexander 1971; Jupp et al. 1972 and Moody 1976) seems to presuppose difficulties occurring at this level also. Examiners' reports tend to support this view and there does seem to be a certain amount of difficulty in writing at this level even in the first language (L1). Viewed from the communicative perspective though, the value of the sentence as a testing unit perhaps needs reassessing, as the communication of messages in the academic context takes place over larger units, i.e. text, for most students.

According to Davies et al. (1974, p.178) the real difficulty with writing appears at the communication stage because:

"The circumstances in which written communication takes place and the social purposes which it serves, are not the same as those of spoken communication."

In written exchanges the addressee is physically absent with the attendant loss of the paralinguistic and suprasegmental features inherent in spoken communication. Written language needs to be far more explicit in its presentation of referential content because the addressor cannot rely on shared knowledge between him and the addressee. Sharwood-Smith (1973, p. 51) pointed out the need for the writer to:

"... be constantly aware of the receiver as being psychologically present and secondly, since the receiver provides no feedback, to build into the message a large number of features which ensure that the receiver recovers most of the message. This entails a much more conscious organisation of linguistic items than is necessary in informal spoken exchanges ..."

The result is the increased importance which attaches to such things as: word order, carefully formed sentences often complex in structure, a use of non-defining relative clauses virtually absent from speech, tighter lexical and grammatical cohesion, and compositional organisation; since a potentially ambiguous meaning cannot be clarified by tone and stress as it can in speech. In addition there are fewer redundancy features because the reader always has the text to refer back to (v. Jordan 1980).

To be able to communicate effectively in writing, students need to be able to employ the rhetorical conventions appropriate to different kinds of discourse, e.g. narrative, argumentation, description, as well as being able to produce grammatically accurate sentences (cf. Davies et al. 1974; Johnson 1976, 1977a and 1981).

Allen et al. (1974a, p.3) pointed out that two kinds of ability are involved:

"The first is the ability to recognise how sentences are used in the performance of acts of communication, the ability to understand the rhetorical functioning of language in use. The second is the ability to recognise and manipulate the formal devices which are used to combine sentences to create continuous passages of prose. We might say that the first has to do with the rhetorical coherence of discourse and the second with the grammatical cohesion of text. In practice, of course, one kind of ability merges with the other ..."

If we take the case of 'scientific English' as an example, it is obvious that this can no longer be simply characterised in formal terms, as exhibiting a high frequency of grammatical forms such as the passive and the present simple, or as consisting of a particular set of technical lexis, because this approach according to Allen et al. (1974^a, p.4):

"... does little or nothing to indicate what kind of communication it is."

They argued that our aim should rather be to see whether students can understand and produce writings that are (p.5):

"... representative of what we conceive to be certain basic communicative processes which underlie, and are variously realised in, individual pieces of scientific writing ..."

Given sufficient time and resources the best approach would have been to analyse a broad cross section of the texts in the discipline areas under review and characterise these in functional terms. In this way we could have drawn up a profile of the most frequently occurring functions across discipline boundaries. This would have furnished us with a description of what Johnson (1976, p.1) described as the:

"... ways in which these functions are characteristically sequenced to produce coherent discourse."

The elaborate and extensive nature of the analysis that this would demand precluded it from our enquiry though and we were left to construct a framework for our test task description from secondary informing sources.

Taxonomies of writing used for course design in the past have, as Candlin et al. (1978) pointed out, often contained categories which characterise the elements of discourse in quite different ways. Johnson (1976) is helpful here as he illustrated how two distinct parameters for use in the analysis of written discourse can be isolated. First there are the categories of 'communicative function' (v. Wilkins 1973, 1976) which relate the elements to the communicative operations they perform under labels such as 'definition, classification, cause and effect' (for a complete list v. Munby 1978). The second parameter is characterised by a concern with how utterances relate to each other within the framework of the discourse and elements are labelled according to their rhetorical functions, e.g. introducer, developer, modifier (v. Imhoof et al. 1975).

In a taxonomy for testing purposes it would seem sensible to take account of both communicative and rhetorical functions of constituent elements in the texts that students have to cope with in an academic

environment. We need to generalise and identify which communicative functions characteristically operate within which discourse positions, e.g. definitions and classifications might occur most commonly in the introductions to exposition while contrasts and comparisons are more common in developmental paragraphs. Imhoof et al.'s writing course From Paragraph to Essay was developed along these lines.

Johnson provided the most useful informing sources for our purposes in that his research, and the course book this resulted in (1981), are based on extensive experience with university students from overseas attending pre-sessional courses prior to embarking on a wide variety of academic courses in this country, Johnson (1976, p.4) stated two aims:

"The first is to ensure that the student can recognise and produce exponents associated with the communicative functions most commonly found in academic writing; the second is to offer him practice in sequencing these functions to produce coherent discourse."

With this in mind Johnson sought to establish the important rhetorical and communicative functions that students would need across disciplines.

Johnson's findings (1976) are summarised below in Table 3P. To a great extent he manages to realise this format in the later Communicate in Writing (Johnson 1981) which is divided into three parts:

- (i) describing things and ideas
- (ii) describing processes and events
- (iii) developing an argument

though he takes care to point out that this is a crude division in that these are not discrete categories exclusive of each other. He is also cautious (1976) over the extent to which norms can be prescribed, in this case how far communicative functions can be associated with certain discourse positions.

TABLE 3P

JOHNSON'S OVERVIEW OF THE RHETORICAL AND COMMUNICATIVE FUNCTIONS TO BE FOUND IN ACADEMIC WRITING.

DISCOURSE TYPE	DISCOURSE POSITION	EXAMPLE FUNCTIONS
Exposition a) describing phenomena and ideas.	Introduction	Defining Classifying Identifying
	Development	Contrasting Exemplifying
	Conclusion	Summary
b) describing processes	Introduction	Describing purpose Describing means
	Development	Sequential description Instructions
	Conclusion	Summary Describing results
Argumentation	Introduction	Stating a proposition Stating assumptions
	Development	Induction Deduction Substantiation Concession
	Conclusion	Summary Generalisation Speculation

SOURCE: Johnson, K. (1976)

McEldowney (1976) drew up a similar list of discourse types:

- a) straight description which 'occurs in some form in almost all academic disciplines'
- b) description of process
- c) narrative which outlines two or more steps in a sequence and is often accompanied by 'narrative padding' providing a descriptive background to the sequence of events. This is seen by McEldowney as 'important in the writing of reports of various types and in the historical and developmental aspects of most disciplines.'
- d) instructions which 'are a common part of the conduct of education in most disciplines'.

She also outlines the structural exponents that are most likely to be used to effect these core functions.

As regards subject matter Johnson (1981, p.4) argued:

"The underlying assumption is that there are important similarities in the way English is used in the various subject specialisations. If the students learn these common elements they will have gone a good way towards learning to write for their specific needs."

Hutchinson et al. (1979, p.31) made a similar point:

"...we need to make a distinction between the Performance Repertoire of the target situation and the Competence required to cope with it. The Competence providing the generative basis for further learning, irrespective of the target subject, is the proper concern for ESP."

We would agree with McEldowney (1976, p.13) that:

"It is not the purpose of the English test to assess the candidate's knowledge of any subject area but to see how he can handle straight-forward ideas in appropriate English."

She argues that topics should be chosen which are of a 'central, neutral semi-technical' nature, from sources of 'semi-technical writing for the educated layman rather than for the specialist'. In this way it is hoped that no one group of candidates would be given an unfair advantage.

3.4.4.2 The Writing Tasks Students Perform in the Academic Context

In the questionnaire both staff and students were asked to indicate how often students, on the programme specified, were expected to produce the types of written work listed below.

Question C1 1) Writing short introductions or connecting sentences in numerical calculations or mathematical arguments during:

- a) coursework
- b) examinations

2) Writing short connected answers to questions demanding a restricted response, e.g. structured questions or short answer questions where the questions specifically define the limits and nature of the response required (i.e. not more than a paragraph in length) in:

- a) coursework
- b) examinations

3) Producing extended writing (i.e. continuous connected writing greater in length than a single paragraph) in:

- a) coursework
- b) examinations

4) Any other types of written work produced by students (please specify below)

The returns to these questions from both the staff and student questionnaires are recorded below in Table 3Q1.

TABLE 3Q2 THE FREQUENCY OF OCCURRENCE OF VARIOUS WRITING TASKS ESTABLISHED ON THE BASIS OF INFORMATION GATHERED DURING THE FOLLOW UP INTERVIEWS

KEY AS ILLUSTRATED FOR TABLE 3E	ENG. U		SCI. U		SCI. P		SCI. A		S.SCI. U		S.SCI. P	
	COURSE- WORK	EXAMI- NATIONS										
1. Selecting/noting information for use in writing tasks	H	M	L	N	H	M	L	N	H	M	L	N
2. Sequencing and writing up of experiments	H	M	L	N	H	M	L	N	H	M	L	N
a) description of results												
b) tabulation of results												
c) summarising and drawing conclusions												
d) using symbols/ graphs/charts/ labelled and unlabelled diagrams												
3. Writing short verbal introductions to and/or connecting sentences in, mathematical arguments or numerical calculations	H	M	L	N	H	M	L	N	H	M	L	N
4. Short answers in connected writing to restricted response questions	H	M	L	N	H	M	L	N	H	M	L	N
5. Extended connected writing	H	M	L	N	H	M	L	N	H	M	L	N
5.1 practical report												
5.2 project work												
5.3 essays												
5.4 dissertations												
5.5 thesis												

H High frequency of occurrence M: Medium frequency of occurrence L: Low frequency of occurrence N: Non-occurrence

Question C1/1 Writing short verbal introductions to and/or connecting sentences in mathematical arguments or numerical calculations in:

a) Coursework - This was a frequent activity for science and engineering students particularly at undergraduate and 'A' level. It was a far less frequent activity for social scientists and more students in this discipline area reported never having to perform this task at all. However in science and engineering less than 10% of the students recorded answers in the 'never' column. Overall the vast majority of respondents reported that they had to carry out this activity 'sometimes' or 'often' (Table 3Q1, QuC1/1a).

This general picture is confirmed by the observations (v. Table 3Q2 above).

b) Examinations - There seems to be agreement, that in most subject areas surveyed, except the social sciences, written tasks of this type occurred 'sometimes' or 'often' in examinations (Table 3Q1, QuC1/1b).

Question C1/2 Writing short connected answers no more than a paragraph in length in:

a) Coursework - In coursework this was a less frequent writing activity for most of the science and engineering students. Well over half the total responses fell into the 'sometimes' category and there were far fewer students who recorded in the 'often' category than in Question C1/1a. It is noticeable that quite a number of the staff, with the exception of the 'A' level science teachers, reported that their students 'never' had to perform this activity at all on the courses they were responsible for. In general it was the 'A' level students who had to perform this activity most frequently during their courses (QuC1/2a).

Overall, of the three types of writing we asked about, writing of this length appears to be the least important.

The picture was largely confirmed by the observations (Table 3Q2, 4).

b) Examinations - It appears that there was a slightly greater frequency of occurrence recorded for this type of writing in examinations than in coursework in the science and engineering groups (Table 3Q1, QuC1/2b). For most groups it was still only likely to occur 'sometimes', recurring frequently only amongst some of the 'A' level science students.

This was largely confirmed by the observations (Table 3Q2, 4).

Question C1/3 Producing extended writing (i.e. continuous connected writing greater in length than a single paragraph) in:

a) Coursework - In the science and engineering groups there is a good deal of variation but overall both staff and students thought that this was an activity which occurred 'sometimes' or 'often' during the course. The student replies tended to record a higher frequency of occurrence in these categories because although a teacher might set a piece of extended written work very infrequently, students were likely to be set written work by a variety of different tutors, in respect of the different courses that they were taking. Thus very few of the students, with the exception of science undergraduates (largely mathematics students, v. Appendix 3.3.2, p.794), reported that they 'never' had to perform this activity at all in their coursework. For social science students it was a major activity and the returns illustrate they had to complete more written work of this type than most of the other groups (Table 3Q1, QuC1/3a).

The view that extended writing is an activity that very few of the students replying to the questionnaire could avoid, is borne out by the frequencies of occurrence recorded in the observations (Table 3Q2, 5). This shows in general terms, for a limited sample, how frequently different forms of extended writing took place, according to the information

gathered during visits to the various departments in the subject areas under review.

b) Examinations - On the whole staff considered that students had to perform this activity more frequently in examinations than in coursework. The students mainly considered that they would have to do this activity either 'sometimes' or 'often' in examinations. Only a small number of the science and engineering students, particularly undergraduates considered that they avoided this activity altogether in examinations (Table 3Q1, QuC1/3b). Again the students recorded a higher frequency of occurrence for this activity than the staff but the students of course are referring to all the examinations they have in connection with their programme, whereas the staff are answering in connection with the courses they are responsible for. The social science students would seem to have produced this type of writing the most as for coursework.

Staff and students were also asked in Question C1/4 to specify any other type of written work that was produced in the programme. We have combined the replies of both staff and students and grouped them together according to discipline and level. In no sense is this to be regarded as a systematic breakdown of the frequency of various writing activities that occur across level and subject boundaries, but rather as a descriptive account, which is meant to illustrate the possible range of activities that may occur even in the sciences and engineering. Often the contributions that we received could have been subsumed under the three general categories C1/1-3 we looked at above, but they are nevertheless included below as they serve to illustrate the variety of writing tasks that the students under review might be confronted with.

Engineering Undergraduates:

"Very short descriptions with emphasis on structure."

"Keeping a clear and concise log of laboratory experiments not necessarily in complete sentence form."

- "5000 word report of a formal meeting."
- "A long technical essay is required in connection with project work."
- "Design courses assessed on four (or more) reports each of more than four pages in length."
- "Field trip report."
- "Programming involves writing reports on the functions of various programmes."
- "Laboratory reports which require a different general style to the general type of extended writing in QuCl/3."
- "Writing in association with a diagram."
- "Production of operating instructions for programmes produced; comment cards on computer programming."
- "Descriptions in surveying books (i.e. site descriptions)."
- "Instructions on an engineering drawing."
- "Operating instructions for a device or a machine."

Engineering Post-graduates:

- "Dissertation."
- "Thesis on students' individual research projects."
- "Technical reports."
- "Design project."
- "Report with calculations and drawings."
- "Laboratory reports on practical work."
- "Seminar papers."
- "Evaluations of their own calculations and solutions to coursework problems."
- "Procedure for calculations."
- "Monitored log book of professional training."

Science Undergraduates:

- "Connecting sentences which have to convey logical argument and require accurate use of language."
- "Laboratory reports."
- "Annotations to drawings."
- "Numerical statements logically connected by symbols."

"Projects, mini-projects."

"Dissertations, 5000 words."

"Writing up practical/case study reports, short accounts stating main conclusions, difficulties encountered, assumptions made etc."

"University science students are always asked to produce at least two long essays by tutors."

"Written assignments, occasional essays, long essays."

"Preparation for talks during tutorials."

"Brief notes for use during discussions."

Science Post-graduates:

"Practical reports."

"Laboratory work reports."

"Long essays on given topics (2500 - 3000 words) emphasis given to language (2 each term)."

"Literature reviews."

"Project reports."

"Tables of data, diagrams of equipment."

"Computer programming (Simula)."

"M.Sc. dissertation."

"Dissertations are usually 50-100 typewritten pages in length and must be in reasonable English. Nearly every overseas student I have had (about 12) has had great difficulty with this and has needed a tremendous amount of help."

"Answers to questions expressed mainly in mathematical symbols."

"Writing a mathematical argument, which although containing mathematical symbols, is mainly in English. A simple slip in English here could totally alter the meaning of the argument."

Science 'A' Level Students:

"Essays/written assignments."

"Extended essays 3000-5000 words."

"Zoology project 2000 words."

"Book review 4000 words"

"Field work project 4000 words."

"Summary notes on subjects."

"Expansion of lecture notes to provide a better understanding."

"Paragraphs giving definitions and examples of scientific terms."

"Exam length answers for practice in 40 minutes."

"Short answer tests."

"Critical evaluation of the students' own programs."

"Laboratory reports."

"Accounts of practicals involving calculations."

"Graphs."

"Annotated diagrams."

"Critical comparisons and creatively produced diagrams."

"Mathematical arguments are expressed in a symbolic language, this appears to be a very specialist use of language symbols."

"Written solutions are mainly symbolic, and numerical: saying written words in the form of a formula."

"Description of dimensional geometrics."

Social Science Undergraduates:

"Essays/long essays."

"Tutorial essays."

"Seminar papers."

"Individual notes made to clarify problem areas."

"Writing reports for industrial studies."

"Numeric questions in costing presented in a clear form for use by management programmes."

"Dissertations of individual or group project work."

"Project for industrial training."

"Student's project; 5000-7000 words, assignments which test virtually all the skills so far enumerated."

"Outline summaries of current literature."

"Flow charts of accounting system."

Social Science Post-graduates:

"All students on this course must do a 5000 word project involving a financial analysis of a chosen company. All the foreign students report a great deal of effort in preparing this."

"Preparation for tutorial discussions/reports from syndicate groups."

"Seminar papers."

"M.A. dissertation and fairly long course essays (5000-10000 words) involving statistical analysis and description and economic tools such as diagrams and graphs accompanied by appropriate terminology."

"Mini theses."

3.4.4.3 Problems Experienced in Written Production

In both staff and student questionnaires we sought to establish, in terms comprehensible to the respondents, where particular problems had occurred in written work.

Students were asked in Question C2 to indicate how much difficulty they had experienced in their written work with:

1. Writing grammatically correct sentences.
2. Using a variety of grammatical structures.
3. Using appropriate grammatical structures.
4. Using appropriate vocabulary.
5. Using a wide and varied range of vocabulary.
6. The subject matter.
7. Expressing what they wanted to say clearly.
8. Arranging and developing their written work.
9. Spelling.
10. Punctuation.
11. Handwriting.
12. Tidiness.

Their replies are summarised in Table 3R below.

Staff were asked to indicate the proportion of the students they taught, on their courses in the programme we had specified, who displayed the following characteristic defects:

1. Grammatical error.
2. Lack of variety in grammatical structures employed.
3. Use of inappropriate grammatical structures.
4. Use of inappropriate vocabulary.
5. Limited range of vocabulary.
6. Inadequate understanding of the subject.
7. Inability to express themselves clearly.
8. Poor arrangement and development of written work.
9. Poor spelling.
10. Poor punctuation.
11. Poor handwriting.
12. Untidiness.

Their answers are recorded in Table 3S below.

It should be noted that whilst the students were asked how much difficulty they had with each area, the staff were not asked about the degree of difficulty they felt students experienced, but how many of their students displayed each defect. It follows that a direct comparison, in quantitative terms, cannot be made between these two sets of responses.

We also asked staff what importance they attached to the following criteria in their assessment of written work in an attempt to put the difficulties experienced into some sort of perspective.

1. Grammatical accuracy.
2. Variety in grammatical structures employed.
3. Appropriateness of grammatical structures employed.
4. Appropriateness of vocabulary.
5. Range of vocabulary.
6. The subject content.
7. Clarity of expression.
8. Arrangement and development of written work.
9. Spelling.
10. Punctuation.
11. Handwriting.
12. Tidiness.

We have summarised the staff's opinion on the relative importance of various aspects of written work in Table 3T below.

We will now consider each element of writing in terms of:

- a) the difficulties it caused for both British and overseas students (Table 3R)
- b) the proportion of each group of these students the staff saw it causing problems for (Table 3S)
- c) the importance the staff claim they attached to it in their assessment of a student's written work (Table 3T).

1 Grammatical accuracy

Only a quarter of all the overseas students claimed that they had no problems here as compared with over half the British students. The bulk of the overseas students thought that they had 'very little' difficulty and the social science students as a group claimed they had the least problems (Table 3R, QuC2/1).

Staff considered that in the sciences and engineering, higher proportions of the overseas students had difficulty than their British counterparts (Table 3S, QuC2/1). Only the British undergraduate engineers seem to have been really troubled. Social scientists in general were seen to have slightly less of a problem. It is noticeable that only a very small percentage of all students (overseas and British) were seen as having no problem with grammatical accuracy.

As regards the importance attached to this assessment criterion (Table 3T, QuC3/2) the picture is varied: only the engineering undergraduate tutors attached any great importance to it, the majority seeing it as having 'medium' to 'low' importance.

2 Using a variety of grammatical structures

There is quite a difference between the amounts of difficulty experienced by the British and the overseas students (Table 3R, QuC2/2) with over half the British students claiming 'no' difficulty here at all. Only the overseas post-graduate social science students had as few problems.

The staff returns again indicate a sizeable difference between the proportion of British and overseas students experiencing difficulty in this area (Table 3S, QuC2/2).

On the whole the staff attached 'low' or 'no' importance to this criterion in assessment (Table 3T, QuC3/2). It was considered the least important of all the criteria we sought information on.

3 Using appropriate grammatical structures

Most overseas students admitted 'very little' or 'some' difficulty here, whereas most British students claimed 'very little' or 'no' difficulty (Table 3R, QuC2/3). Of the overseas students the social scientists again experienced least difficulty.

Staff replies (Table 3S, QuC2/3) indicate that they considered 'a lot' of overseas students had problems with the use of appropriate grammatical structures, more so than their British counterparts.

In general the staff attached only 'medium' to 'low' importance to this criterion in their assessment of written work (Table 3T, QuC3/3).

According to the overall staff returns (v. Appendix 3.4, p.810) it was these three grammatical categories which caused difficulty to the greatest proportion of the overseas students.

4 Using appropriate vocabulary

This seemed to pose a problem for many overseas students especially science post-graduate and science 'A' level students (Table 3R, QuC2/4). All overseas students other than a small number of engineering post-graduate and social science students experienced 'some' difficulty in using appropriate vocabulary whereas very large numbers of the British students claimed to experience none.

Staff on the whole estimated that fewer of their students had a problem with this than they had had with the grammatical categories. They still however saw a gap existing between the proportions of British and the proportions of overseas

students who experienced problems in this (Table 3S, QuC3/4).

On the whole, staff thought this criterion had 'medium' to 'high' importance in the assessment of written work. It is noticeable that about a third of the staff claimed that they gave 'high' importance to this criterion in their assessment of written work (Question C3/4). Noticeably fewer social science staff considered it of 'high' importance though.

5 Range of vocabulary

This was the category in which the degree of difficulty experienced by the overseas students was greatest, as compared with the British students, the majority of whom considered that they had 'very little' or 'no' difficulty (Table 3R, QuC2/5). It seemed to be less of a problem for the overseas social science students particularly post-graduates.

Staff considered there was a gap in performance between the British and the overseas students though they saw both groups as having less of a problem with this and the appropriacy of lexis employed, than they had had with the grammatical categories (Table 3S, QuC2/5). The problem was seen as being substantial for science 'A' level and science post-graduate students from overseas but not so much of a problem for the social science post-graduates.

Very few of the staff regarded this as being of 'high' importance, the majority considering it as of 'medium' or 'low' importance as an assessment criterion (Table 3T, QuC3/5).

6 The subject matter

Very few students, either overseas or British, experienced 'a lot' of difficulty with this, most claiming 'very little' or 'no' difficulty. The British students seemed to have had only slightly less of a problem in this area (Table 3R, QuC2/6).

Of all the criteria listed this was claimed to be the lowest cause of difficulty by both British and overseas students.

In the staff questionnaire (Table 3S, QuC2/6) the difference in the proportion of overseas as against British students experiencing difficulty in this area is very small, except for science post-graduate students where there appears to be quite a large difference. According to staff, more overseas students had greater difficulty with their handwriting, spelling and punctuation than they did with problems arising out of the subject matter. This seemed to be at odds with our intuitions but the question does refer to the written work staff received. These returns perhaps point to the difficulties involved in attempts to make any clear-cut division between difficulties over the subject content as against the language it is expressed in.

Nearly all the staff claimed that this criterion was of 'high' importance and overall it was claimed to be the most important criterion in their assessment of written work (v. Appendix 3.4, p.812)

7 Clarity of Expression

An increased number of British students particularly in science and engineering saw themselves as having problems here though the majority still considered that they had 'very little' or 'no' difficulty (Table 3R, QuC2/7). The majority of overseas students claimed 'very little' or 'no' difficulty in writing clearly. Although there is still a gap between the relative amounts of difficulty experienced by the two groups, it is smaller than was the case with some of the other criteria.

Staff teaching science, engineering undergraduate and social science undergraduate students thought a lot more of the overseas students had difficulty here than their British counterparts. Nearly half the staff teaching the overseas science post-graduate students thought 'a lot' of their overseas students had difficulty with this matter (Table 3S, QuC2/7).

Clarity of expression was seen by the majority of staff as the second most important criterion of assessment next to subject content (Table 3T, QuC3/7). Social science staff regarded it as of slightly lesser importance than the other staff groups.

8 Arrangement and development of written work

Very few students, overseas or British, saw themselves as having 'a lot' of difficulty here. The majority of overseas students felt they had 'some' or 'very little' difficulty (Table 3R, QuC2/8).

Staff in general considered that 'some' or 'a lot' of their overseas students had difficulty in arranging and developing their written work. There are quite large differences in some cases between the relative proportions of overseas and British post-graduate students that they see as experiencing difficulty in this area (Table 3S, QuC2/8). This is perhaps partially explained by the fact that post-graduate students are far more likely to have to produce longer pieces of extended writing than the undergraduate students and these organisational features become more important the larger the size of text that is being produced.

Next to subject content and clarity of expression, this feature was the one most commonly cited as being of 'high' importance; being judged so by almost half the staff who answered the questionnaire (Table 3T, QuC3/8). Very few staff regarded it as having 'very little' or 'no' importance.

9 Spelling

On the whole, the majority of students claimed they had 'very little' or 'no' difficulty with spelling (Table 3R, QuC2/9). Overseas students usually admitted to having slightly more difficulty than the British students but in some cases less. The British students on the whole considered that spelling was their greatest cause of difficulty in writing.

Staff thought that more overseas than British students experienced difficulty with spelling and, in the case of science 'A' level students and science post-graduates, a lot more (Table 3S, QuC2/9).

Very few staff thought spelling of 'high' importance except at engineering undergraduate level. The majority thought it was of 'some' or 'very little' importance (Table 3T, QuC2/9).

10 Punctuation

Most students, overseas and British, claimed that they had 'very little' or 'no' difficulty here, the overseas students admitting to only slightly greater problems in this area than the British students (Table 3R, QuC2/10).

Staff thought that a large number of overseas students had difficulty in using punctuation correctly. The staff in post-graduate and 'A' level science and engineering considered that notably higher proportions of overseas students suffered from this difficulty (Table 3S, QuC2/10).

The majority of staff thought that it had only 'some' or 'very little' importance in terms of assessment (Table 3T, QuC3/10). Very few staff considered that it had 'high' importance.

11 Handwriting

In the main British students admitted to having more problems with this than the overseas students. The vast majority of overseas students claim 'very little' or 'no' difficulty at all (Table 3R, QuC2/11). Science post-graduates admitted to the most difficulties among the overseas students.

Slightly more overseas students than British were seen by the staff as having problems with this across most subject areas and

levels except social science undergraduate (Table 3S, QuC2/11).

Staff in general attached 'very little' importance to this criterion in their assessment of written work (Table 3T, QuC3/11), though it did assume slightly greater importance for some engineering undergraduate and science 'A' level staff.

12 Tidiness

This is the only category in which British students clearly admit to having more problems than the overseas students. Next to spelling British students claimed that overall this caused them the greatest difficulty. A greater number of overseas students than British claimed that they had no problems in this area (Table 3R, QuC3/12).

The staff thought that few overseas students in general had problems here. They thought slightly more overseas than British experienced problems in this category in general, except in social science and science undergraduate, and science 'A' level classes, where the reverse was true (Table 3S, QuC2/12).

A surprisingly large number of staff regarded this criterion as being of 'high' importance. The majority of the staff saw it as of either 'high' importance or 'medium' importance (Table 3T, QuC3/12). It was considered to be especially important by engineering undergraduate and science 'A' level staff.

Staff were also given the opportunity in Question C2/13 to identify any other characteristic defects displayed in students' written work as a whole. We have listed below those answers which serve to extend or illuminate the discussion above.

Grammar

"...their grammar is often much better than that of the British students who often cannot tell an adverb from an adjective!"

"Inability to write long sentences."

"Overcomplex syntax."

Appropriacy

"Lack of appropriate English phrases connecting mathematics statements."

"Imprecise use of words in a strict scientific sense (the 'Life on Earth' syndrome). Using words that are catchy but are scientifically inadequate."

Conciseness

"The writing is not concise. Many words are used when it is not necessary."

"Some overseas students are not concise enough in their answers and may suffer when they are constrained by time limits."

Instructions

"Students do not always follow instructions on presentation of work."

"Some write too much."

Style

"Some students do not make sufficient use of the available literature in the field and are unfamiliar with the conventions for quoting references."

Orderliness

"Logical explanation of facts in an orderly fashion."

"Inability to produce coherent written work."

"Discussing a science topic at length means a systematic recall of all relevant information which must be explained and organised into a logical sequence."

Relevance and adequacy

"Tendency to copy verbatim from the text without necessarily understanding."

"Incorrect answers to questions posed."

"Overseas students seem to be more likely to submit written work which contains an unacceptable proportion of irrelevant information."

"Tendency to copy verbatim from texts rather than use their own words."

Compositional features

"Setting out i.e. paragraphs, sub-headings, identification of which answer is for which part of a question."

"Format."

Cultural

"Overseas students have sometimes been taught in a different tradition of writing."

Labelling

"Poor illustrations, no labels."

"Appalling use of illustrations, e.g. labels by our students."

"Graphs which are not labelled, diagrams which are not labelled."

Handwriting

"Iranian and Arabic students have difficulty writing clear English because they are writing in an unnatural direction across the page."

"Illegible handwriting though full of character."

Students were also asked in Question C2/13 to specify any other difficulties that they had had with written work. Few British students contributed here and as their replies were covered by the comments made by overseas students we have taken an illustrative sample from the latter, where these supplement or illustrate the description of the problems outlined above.

Speed

"Writing with speed in examinations."

"I write slowly."

"When obliged to write quickly results in serious spelling mistakes."

"Difficulty in writing neatly at fast speeds."

Functional adequacy

"Putting argument in a sequential order."

"... summarising a long argument and drawing conclusions."

Referential adequacy

"I have to use a dictionary when writing coursework essays. When I am not allowed access to a dictionary i.e. in exams, the quality of the language I use drops dramatically."

Relevance

"Writing long paragraphs which are not irrelevant to the title or to what has been asked."

Conciseness

"Writing short sentences which are direct and to the point."

"Shortening a long explanation."

Organisation

"Paragraphing."

"Writing long paragraphs."

Legibility

"Writing at speed in a way that others can read it."

"My handwriting is terrible which means a lot of lost marks when it is not understood."

Staff were further asked in Question C3/13 to specify any other criteria that they considered to be important in assessing the written work of the students on the programme specified. We have tried to categorise various suggestions that were made insofar as they serve to supplement or illustrate the description of the assessment criteria referred to above.

Grammar

"Grammatical accuracy must be sufficient to convey meaning."

Referential adequacy

"... careful use of defined terms, e.g. work, energy."

"There should be no ambiguity in technical descriptions."

"Technical accuracy."

"Accuracy of chemical names."

"Correct answers."

"Use of the correct abbreviations."

"Relevance to the question asked."

"Keeping strictly to the point of the question."

Labelling

"In engineering reports concise labelling of graphs etc. is essential."

Administration

"Absence of collusion between students."

Ability

"Analytical abilities; critical analysis, reflectiveness."

"Willingness to think for oneself."

"Effort made."

"Originality of the approach."

"Initiative in searching out source material."

"Originality in the sense that they have used their own words. Many overseas students borrow large quotations to bolster their English."

Conciseness

"Ability for succinctness while presenting all the relevant points."

"Description of complex ideas briefly and accurately."

"The ability to be brief and avoid 'flannel'."

Legibility

"Within certain limits of legibility handwriting is unimportant, but outside these limits it becomes important and in the extreme cases almost critically so, since some few students present work which is almost impossible to decipher."

"Any work which I cannot read is automatically marked wrong."

Organisation

"Layout and ordering different sections of a report."

"Arrangement and development of mathematical concepts. Coherent and logical structure in the solution of a problem."

"The ability to marshal facts into a logical sequence in essay type answers."

Applicability

"Most of the criteria in section C2/3 are not really important in mathematics. Main criteria for mathematical work for me is: to give a brief accurate statement, to the point - the meaning must be clear even if the English is not correct."

3.4.4.4 The Standards Expected of British and Overseas Students

Staff were asked in the questionnaire:

Question C4 Do you expect the same standards of written work from overseas students as from British students?

* If no, what allowances do you make?

Question C5 Do these allowances you make vary, as between coursework and examinations?

* If yes, please specify in what way(s) they vary.

The staff replies to these questions are recorded in Table 3U below.

TABLE 3U Question C4 - SUMMARY OF QUESTIONNAIRE RETURNS CONCERNING WHETHER THE STAFF EXPECT THE SAME STANDARDS OF WRITTEN WORK FROM OVERSEAS AS FROM BRITISH STUDENTS

Question C5 - SUMMARY OF RETURNS ON WHETHER THESE ALLOWANCES VARY AS BETWEEN COURSEWORK AND EXAMINATIONS

	QuC4		QuC5	
	Y	N	Y	N
Eng. U	11	0	11	0
Eng. P	11	0	11	0
Sci. U	11	0	11	0
Sci. P	11	0	11	0
Sci. A	11	0	11	0
S.Sci. U	11	0	11	0
S.Sci. P	11	0	11	0

Y: Yes
N: No

KEY AS ILLUSTRATED FOR TABLE 3F1

3.4.4.4.1 Variations in the allowances staff claim they make in marking the written work of overseas students

In general about two-thirds of the staff said that they expected the same standard from overseas students as from British students and a third said they did not. Thus, the comments below only represent the views of the minority of staff consulted.

About half of the science post-graduate tutors and two-thirds of the social science tutors said that they did not expect a similar standard. This unfortunately was a slightly ambiguous question in that 'expect' did not necessarily equate with them accepting lower standards of written work as it might be taken as meaning 'did they get it'. However the follow up question:

'If no, what allowances do you make?'

should have helped to remove any misunderstanding here.

The main point to be borne in mind is the variation in allowances that were made by those who did not expect the same standards. We have tried to give an idea of this variation in our selection below of some of the comments that were made by staff in the different subject and level groupings. Though one could gather from these comments that staff were generally prepared to make allowances as regards manner of expression, as long as this did not interfere with the meaning of what was being conveyed, the very variety of the allowances they were prepared to make effectively prevents further generalisation.

We would argue strongly that this variety precludes the possibility of making any valid generalisations concerning tolerance levels that operate in the written medium on the part of staff and must bring into question the findings of both Carroll (1978) and Munby (1978). In practice establishing these tolerance conditions is by no means as easy as they had assumed from their pre-theoretical position.

We would argue that any attempt to specify these tolerance conditions is at best a specious activity when more than one lecturer is involved on a course and sometimes a single lecturer is by no means consistent in the application of these allowances. In any case, at least two thirds of the total staff consulted claimed they made no allowances at all.

The concern below is not to show systematically any differences between subjects and levels, but rather to illustrate how across these there is a good deal of variation in the type of allowance that is made.

Engineering Undergraduate:

"Allowances made in the first term only."

"I am prepared to spend longer going through their work."

"Read for sense rather than for meaning."

"Often the subject matter and the examples they produce are not familiar to overseas students."

"More weight given to the subject content rather than the language."

"Less control over lexis."

"Allowances made for 'quaint' but still comprehensible words and idioms."

"Literary style."

"Punctuation."

"Small allowances made particularly with clarity of work and expression."

"... allowance made for grammar and spelling if it is clear that the student has understood the subject matter."

"... slightly greater tolerance of unusual style or structure."

"I merely accept without penalty a lower standard of English. If the English standard is below that acceptable I refer them to remedial classes."

Engineering Post-graduate:

"Same standard is required in content and clarity, but not in language."

"Answers accepted in note form."

"Allowances made for grammar, spelling and punctuation etc."

"Allowances made are difficult to quantify."

"Allow for the fact that they are working in a foreign language."

"Lack of background information relevant to subjects covered."

Science Undergraduate:

"The same standard of maths is expected but ambiguous connecting sentences are more readily accepted from overseas students."

"Language limitations accepted from overseas students but not from British students."

"Lack of background due to educational and cultural differences."

"Grammatical accuracy sometimes overlooked."

"Fortunately in mathematics this question scarcely arises."

"In the first year considerable tolerance is shown - expect them to be almost directly comparable by the third year."

"Examinations are marked blind so one is less able to make allowances. However even in examinations overseas students can usually be identified by their defective grammar and might thereby have some allowances made."

Science Post-graduate:

"Make more allowances for language difficulties, e.g. we accept a slacker definition or theorem statement provided we could see some evidence of clear mathematical thought."

"They are given longer to do the exams."

"Grammar, spelling, vocabulary."

"It depends on my assessment of how much the student is really understanding."

"... will tolerate lesser clarity if I consider it is caused by linguistic difficulties. Also tolerate less coherence in project report."

"I do not penalise overseas students at all for poor English as long as content is good, or so long as I can understand it sufficiently to be fairly clear of what the student is writing."

"I make allowances for use of language but not comprehension of content or organisation."

"Allowances in level of grammatical accuracy, vocabulary used, spelling, punctuation, etc."

"I do not penalise overseas students for poor English either in coursework or examinations, but I do not accept dissertations until the English is satisfactory."

"One is prepared to make some allowance for unclear expression provided one is satisfied that the material is properly understood. I do not think this is always done very consciously but I think it arises nevertheless."

Science 'A' Level:

"Lecturer spends longer marking work of overseas students."

"Dependant on work and ability of student."

"Practicals assessed on content only."

"The work is acceptable if the meaning is clear."

"Less severe penalties but work is always corrected."

"Lack of vocabulary, grammar and punctuation and spelling."

"Grammar, spelling, and calligraphy for Arabic students."

"The same standard is not expected from overseas students but they will lose marks for incomprehensible work."

"Allowances made for different style of presentation of maths from overseas students."

"The English is not as good but frequently the work is of a higher standard."

"Allowance for clarity of expression, correct grammar and appropriate technical vocabulary - by accepting a lower standard without adverse comment."

"Account is taken of attitude to subject."

"In the early part of the course allowance is made for grammar, punctuation and spelling because the whole class (all foreign) have similar problems to a greater or lesser degree. Towards the end of the first year in 'A' level, little allowance is made for these faults."

"I try to correct spelling and grammatical mistakes and encourage them to improve - in the 2nd year I would expect the same standard."

"Allowance may be made for quaint phraseology and difficult spellings."

"I expect the same information in an essay but expect the style of the essay to be simple, e.g. short uncomplicated sentences; e.g. concepts explained in poor English."

"Often, overseas students, accustomed to a foreign script have problems with handwriting. Additionally I am personally sympathetic to errors of spelling or punctuation."

"Spend more time correcting grammar - but, in main, do not penalise students for their use of English."

"This of course depends very much on the ability of the foreign student. Some are practically bilingual and I would expect an equal standard with British students. Those who do have problems, I make allowances in word order, spelling, grammar, and sometimes in expression, clarity, dependent on content."

"I make allowances for lack of knowledge of the structure of English, spelling, grammatical errors. I decide whether or not they have understood the basic physical principles involved in the course."

"Am prepared to allow a 'period of grace' - to improve."

"Prepared to put up with more eccentricities of spelling."

"Will allow - for a while - poorer handwriting."

"I accept clumsy, wrongly phrased statements if the meaning is quite clear."

"They have to take the same exam."

"Only the most glaring 'language' errors are corrected for overseas students unless their English is deemed to be fairly good, in which case same standard as from British students is expected."

Social Science Undergraduate:

"Within reason."

"It always has been so on this course."

"Perhaps less clarity of expression - but marginal - in coursework only not exams."

"For grammar, spelling, punctuation and to a certain extent clarity of expression."

"Misuse of words and often use of shortened sentences."

"Poor vocabulary, difficulty in precise expression."

"Use of stilted English by overseas students as long as it can be understood."

"Provided that academic content is understood and there is evidence of effort."

"Understandable errors are ignored in assessment grade, but pointed out to student."

"Grammatical lapses etc. are viewed more sympathetically if and when they occur, though sympathy would diminish as time passed - less allowance made for substantive errors. Generally less concerned at weakness."

"Content judged, rather than expression, organisation etc."

"Appropriateness of grammatical structures, spelling."

"I may be inclined to overlook errors in grammar if content is O.K. provided the English is sufficiently clear for me to be able to grasp the student's argument."

"I would try to correct faults in English but I'm usually reluctant to deduct marks for poor style."

"Lack of clarity of expression by British students often results from inadequate understanding of basic concepts which may be incorporated into written work from notes or memory without really being properly understood. Lack of language fluency can sometimes create this impression in work presented by overseas students who have in fact grasped the essential principles and the 'benefit of the doubt' may be necessary."

Social Science Post-graduate:

"Complete allowance made for poor English - so long as clarity of expression is achieved and there is no doubt that concepts and information are understood."

"Grammar, spelling, sophisticated style etc."

"Acceptance that work will take longer to assess. Making special effort to deduce intuitions."

"Lack of sophistication in answers."

"Grammatical faults as long as understanding of subject is demonstrated."

"There are no British students in the programme, but some (e.g. Indians) use English as their first language while others (e.g. Latin Americans) do not. I tend to pay far less attention to grammar, spelling, etc. with the latter and to focus on the content."

"Poorer English expected."

"I try to guess what they mean."

"Same standard in content, but not necessarily in formal presentation."

"Encourage them by practice to improve."

"In terms of expression and language I am more lenient."

"The main criterion is that I can understand what the student is trying to say."

"Only in disregarding expression and giving benefit of doubt where expression unclear."

"Overseas students may have lower standards but not sufficiently low as to bring the University Degree into disrepute."

"In the early stages of the course overseas students' written work tends to be evaluated according to my assessment of what it would be like if expressed in competent English."

"Foreign students tend to write long essays full of mere irrelevant material. This gets knocked out of them early on but some don't learn."

"I allow for problems with English - it is a technical course so it is content that is important."

"Allowance made in use of English not over actual content."

3.4.4.4.2 Variations in the allowances staff claim they make in assessing the written work of overseas students, as between coursework and examinations

Staff were asked, if they made allowances, to specify whether these differed as between coursework and examinations. Of the staff answering the question the great majority said that they did not vary. The replies of the social science post-graduate staff indicated that they felt that there was some variation in the allowances that were made (Table 3U, QuC5).

We have attempted to give an impression of the form these variations might take by categorising the replies under the two groupings below.

Allowances made in coursework but not in examinations

"Students must be relevant in exams - they don't get away with long rambling answers."

"More allowance in coursework, dissertations can be rewritten, exams cannot."

"Coursework provides an opportunity for correction of language difficulties as well as difficulties with course content. By the time of the exam these should have improved."

"One expects better performance in exams."

"Allowances made during the year in report writing but no allowance made in the exams."

"... do not always know who are overseas students in exam papers so cannot make allowances."

"Examinations are marked to the same standard."

"The standard in examinations is expected to be higher as no allowance will be made in external examinations."

Allowances made in examinations but not in coursework

"Greater allowance made in exams where time is limited."

"Coursework does not have the same time and pressure constraints."

"Less attention is paid to grammatical and spelling error in respect of all examination work done at high speed."

"In examinations the precise extent to which allowance must be made must be formalised. In particular one has to distinguish carefully between inability to express ideas and inability to understand ideas."

"Overseas students go to pieces in exams. It takes them longer to read and to write. I do not expect the standard to be as bad in assessed coursework."

"Poorer English is acceptable from overseas students in exams because under pressure they cannot be expected to perform as well as native speakers."

"Content takes precedence over expression in exams."

"In exams I assess the student's understanding of the subject as opposed to his presentation."

"Allowances made in exams for students who have misunderstood the question."

"Only rarely is one penalised for English in mathematics exams provided that one can understand what has been written."

"I have with others been instructed to give overseas students a longer exam to allow for their inadequacies in English, 5 questions in three hours instead of 4 questions in two."

"With all students, I think one makes more allowance with examination work, for obvious slips made under pressure of time (e.g. if a student says something which is incorrect, evidently due to hasty writing, while having shown elsewhere in his answer that he has perfectly well understood the same matter)."

Thus there is also a sharp contrast between those who make allowances in coursework but not in examinations, and those who make allowances in examinations but not in coursework. This lends further weight to our earlier contention: it is an unrealistic task to speculate on the tolerance conditions that will apply in the assessment of written work.

3.4.4.5 Constituent Enabling Skills

Though a majority of our candidates will be entering scientific or engineering courses, continuous writing will be needed to some extent by all of them. Whilst we feel that our aim should be primarily to test productive skills in an integrated form, it might be that we would feel it necessary to test constituent enabling skills as well. This would help satisfy considerations such as objectivity in marking and would also enable us to control the sample of the language skills produced and provide an additional basis for comparison with the more subjective assessments of the largely uncontrolled samples of connected writing produced by the candidate himself.

Even if we chose not to test these constituent enabling skills discretely, they may nevertheless have a part to play in our assessment of any pieces of connected writing that candidates are asked to produce.

On the basis of the enquiry described above and a survey of the literature, we list below those skills which we think contribute to competence in the production of written English.

1. Manipulating the script of the language: handwriting, spelling, punctuation (cf. Munby 1978; Byrne 1979 and Jordan 1980).

2. Expressing relations within the sentence especially
 - (a) elements of sentence structure
 - (b) modification structure
 - (c) negation
 - (d) modal auxiliaries
 - (e) complex embedding
 - (f) focus and theme(cf. Swales 1971; Quirk et al. 1972; Munby 1978; Jordan 1980 and Johnson 1981).

3. Expressing relations between parts of a text through lexical cohesion devices, e.g. lexical set/collocation (cf. Halliday et al. 1976 and Munby 1978).

4. Expressing relations in a text through grammatical cohesion devices of
 - (a) reference
 - (b) comparison
 - (c) substitution
 - (d) ellipsis
 - (e) time and place relaters
 - (f) logical connectors(cf. Halliday et al. 1976; Quirk et al. 1972; Byrne 1979; Munby 1978; Jordan 1980 and Johnson 1981).

5. Using indicators in discourse for
 - (a) introducing an idea
 - (b) developing an idea
 - (c) transition to another idea
 - (d) concluding an idea
 - (e) emphasising a point, indicating the main or important information

- (f) explaining or clarifying a point already made
 - (g) anticipating an objection or contrary view.
6. Expressing the communicative value (function) of sentences
- (a) using explicit indicators
 - (b) without explicit indicators
- (v. Munby 1978, pp.185-189 for list of accompanying micro-functions).
7. Expressing conceptual meaning, e.g. quantity and amount, comparison and degree, causes, result, purpose, reason, condition and contrast (cf. Leech et al. 1975; Swales 1971 and Jordan 1980).
8. Expressing information
- (a) explicitly
 - (b) implicitly
- (cf. Munby 1978; Imhoof et al. 1975; Jordan 1980 and Johnson 1981).
9. Planning and organising information
- (a) narrative
 - (b) description of phenomena and ideas
 - (c) description of process and change of state
 - (d) argumentation.

This would involve understanding relations between parts of text and employing patterns of organisation such as definition, classification, sequence, listing, cause-effect, comparison-contrast, generalisation, exemplification, speculation, summary, concession, induction, deduction, substantiation, instructions (cf. Swales 1971; Widdowson 1971; Allen et al. 1974a; Imhoof et al. 1975; Jones et al. 1975; Bates et al. 1976; McEldowney 1976; Johnson 1976, 1981; Munby 1978 and Jordan 1980).

3.4.5 Speaking Activities in the Academic Context

3.4.5.1 Introduction

A decision was taken by the Board at an early stage in the project not to include initially a speaking test in the T.E.A.P. battery. Given the limited time and resources at our disposal for the construction of T.E.A.P. and the vast extent and complexity of the literature on spoken production, we have not provided a review of the literature. We did however collect the basic data on the speaking tasks students had to cope with in an academic context in order that the information would be available on which a future oral test might be based.

James (forthcoming) contains an extensive review of the literature on spoken production and details how an E.A.P. oral test might be devised on the basis of the specification described below.

3.4.5.2 The Speaking Tasks Students Perform in the Academic Context

In this section we are only concerned with the speaking activities that take place in events occurring in the academic context.

In any consideration of speaking activities in this context we must take account of several factors which might constrain a student's speaking activities. Firstly, in reciprocal speech situations the student has the additional task of monitoring what other people are saying as well as making his own personal contribution. Thus listening comprehension skills will inevitably affect a student's performance in the more productive mode and the difficulties students admit to in speaking activities might, in part, be caused by failings in the more receptive ability. Secondly, it must also be remembered that part of the overseas students' difficulties in speech situations in the academic context might stem from the subject matter itself, although the evidence presented in the earlier sections above suggest that this is as great a problem for British as overseas students.

3.4.5.2.1 Informing transactions

In the questionnaire we asked in Question D1 how often students had to give oral reports or short talks during their courses. The returns from both staff and students are summarised in Table 3V1(a) below.

TABLE 3VI(a) SUMMARY OF RETURNS TO THE QUESTIONNAIRE CONCERNING THE FREQUENCY WITH WHICH STUDENTS GAVE ORAL REPORTS OR SHORT TALKS (Question D1)

TABLE 3VI(b) SUMMARY OF RETURNS TO THE QUESTIONNAIRE CONCERNING THE FREQUENCY WITH WHICH STUDENTS WORKED TOGETHER WITH OTHERS USING ENGLISH AS A MEANS OF COMMUNICATION IN QuD2a - lectures, QuD2b - seminars/tutorials, QuD2c - practical classes

KEY AS ILLUSTRATED FOR TABLE 3FI

	QuD1			QuD2a			QuD2b			QuD2c		
	N	S	O	N	S	O	N	S	O	N	S	O
Eng. U	φφ	φφ		φ	φ		φφ	φ				
	ΔΔΔ			Δ	Δ	Δ	ΔΔΔ					
	ΩΩ	Ω		ΩΩ			ΩΩ					Ω
Eng. P	φφ			φφ			φφ					
	ΔΔ	ΔΔ		ΔΔ	Δ		ΔΔΔ					
	ΩΩ	Ω		Ω		Ω	ΩΩ					Ω
Sci. U	φφ	φ		φ	φ		φφ	φ				
	ΔΔ	Δ		ΔΔ	Δ		ΔΔΔ					
	ΩΩ	Ω		ΩΩ		Ω	ΩΩ					Ω
Sci. P	φφ	φφ		φφ			φφ					
	ΔΔ	ΔΔ		ΔΔ			Δ	ΔΔΔ				
	ΩΩ	ΩΩ		ΩΩ		Ω	ΩΩ					ΩΩ
Sci. A	φφ	φφ		φφ	φ		φφ					
	Δ	Δ		ΔΔΔ			ΔΔΔ					
	ΩΩ	Ω		ΩΩ		Ω	ΩΩ					ΩΩ
S.Sci. U	φφ			φφ			φφ					
	Δ	Δ		ΔΔ	Δ		Δ	Δ				
	ΩΩ	Ω		ΩΩ		Ω	ΩΩ					Ω
S.Sci. P	φφ	φ		φ	φ		φφ					
	Δ	ΔΔ		Δ	Δ	Δ	Δ	Δ				
	Ω	Ω		Ω	Ω	Ω	ΩΩ					Ω

For engineering and science students this was an activity which they were not called on to perform regularly. About half of the students surveyed in these subject areas claimed this was an activity they 'never' had to do at all. If they did have to perform it, they only did so 'sometimes'. For the social science students on the other hand, this was an activity that many felt they had to perform 'sometimes' or 'often' and far fewer recorded answers in the 'never' column. This is the picture we get from the staff returns as well (Table 3V1(a), QuD1).

During the visits we made to educational institutions a variety of constraints resulted in us observing only a small number of seminars or tutorial type classes. Except at the post-graduate level, and in undergraduate social science subjects, students were seldom called on to give oral reports or short talks, and when they did, these were limited in size (Table 3V2(a)). When the activity occurred it usually took the form of a student reading out aloud from a paper he had prepared in writing beforehand. Alternatively, in some of the social science classes each of the students was given a copy of the paper before the period and the author gave a brief synopsis at the beginning of the class. Post-graduates were normally expected to give at least one seminar during a course and, dependent on the size of the group, sometimes more. They were also expected to contribute in other people's.

The range of classes subsumed under this composite seminar/tutorial heading - individual tutorials, problem classes, small group tutorials, small seminars, large seminars, etc. - were in most cases, normally marked by the dominant role played by the member of staff or supervisor conducting it.

TABLE 3V2(a) THE FREQUENCY OF OCCURRENCE IN OBSERVATIONS OF ORAL REPORTS OR SHORT TALKS IN WHICH STUDENTS MAKE STATEMENTS:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
a	H M L N □ □ □ □					
b	H M L N □ □ □ □					
c	H M L N □ □ □ □					
a	H M L N □ □ □ □					
b	H M L N □ □ □ □					
c	H M L N □ □ □ □					
a	H M L N □ □ □ □					
b	H M L N □ □ □ □					
c	H M L N □ □ □ □					
a	H M L N □ □ □ □					
b	H M L N □ □ □ □					
c	H M L N □ □ □ □					

1. of fact, principle

2. of problems

3. of hypotheses or speculation

4. of experimental procedure

KEY AS ILLUSTRATED FOR TABLE 3E

a. lectures
 b. seminars/tutorials
 c. practical classes

H: High frequency of occurrence
 M: Medium frequency of occurrence
 L: Low frequency of occurrence
 N: Non-occurrence

3.4.5.2.2 Students working together using English as a means of communication

In 'A' level science lectures (non-practical classes) this, as one would expect, was a fairly frequent activity given the more interactive nature of these classes. What was surprising, having observed the teacher centred nature of most classes in our visits, was the frequency of this activity reported in lectures by large numbers of overseas students across other groups and levels (Table 3V1(a), QuD2a). In the case of the social science undergraduate returns it may well be that the students on two B.Ed. courses, who make up large numbers of the overseas students in this group, are involved in different types of activities in their lectures from the students who participated in the lectures we observed. However this does not explain away the other high occurrences of this activity. It may well be that the activity the overseas students are referring to is of a more covert kind, in that they are checking with each other more frequently than the British students during lectures, to see that they are getting an accurate record of the information that is being conveyed.

The staff returns indicate that except for the 'A' level science students, staff did not see this as being a frequent activity in the lectures they gave. This would tend to support the view that student-student interaction is not a formalised aspect of most lectures. This is a view supported by the information gathered during the observations (v. Table 3V2(b) below) where a low frequency of occurrence was recorded for this activity in lectures in most subjects and levels with the exception of 'A' level science.

In seminars and tutorials this activity was seen as occurring far more frequently than in lectures according to most staff and students. What is noticeable though is that except for the social scientists there are quite high entries in the 'sometimes' column for the other groups and only in the social sciences is there any heavy concentration in the 'often' area

(Table 3V1(b), QuD2b). With hindsight one feels that there is a certain degree of overlap between some of the replies to this question and those to Question D3b. Group discussion is perhaps a better way of characterising the reciprocal speech activities that take place in seminars especially in the social sciences. However in certain classes under this broad seminar/tutorial heading, e.g. those devoted to problem solving in the sciences and engineering, students may well work with each other rather than engage in group discussion in an open format.

Working together using English to communicate is a more appropriate description of reciprocal speech activities occurring in practical classes, for, as we have seen, it is seldom an integrated feature of lecture activities and it is more easily viewed in terms of group discussion in tutorials and seminars, especially in the social sciences.

In practical classes (Table 3V1(b), QuD2c) science and engineering students generally recorded high frequencies of occurrence for this activity. Very few scientists or engineers claimed that they 'never' had to work with other students using English as the means of communication in these classes. This view is supported by the staff returns.

The frequency of this activity in practical classes is also borne out by the observations. The most frequent occurrences of students consulting each other for various purposes were recorded in this type of class. In Table 3V2(b) below we have recorded the frequency with which students consulted with each other for various purposes in the practical classes we observed. The frequencies recorded are to a certain extent distorted in that they also contain reference to those occasions when students consulted other sources of information such as worksheets or books for the same purposes. In general though they substantiate the view that it is in practical classes that students are most likely to perform this particular activity. This is not to say that overseas students who work with fellow nationals do not also use their own language to communicate with each other, however much this may be discouraged by staff.

TABLE 3V2(b) THE FREQUENCY OF OCCURRENCE IN OBSERVATIONS OF PUPILS SEEKING INFORMATION OR CONSULTING FOR THE PURPOSE OF:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
1. acquiring or clarifying facts or principles	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
2. identifying or solving problems	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
3. making inferences, formulating or testing hypotheses	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
4. seeking guidance on experimental procedure	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
5. developing own opinion or line of argument	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N

a. lectures
 b. seminars/tutorials
 c. practical classes

H: High frequency of occurrence
 M: Medium frequency of occurrence
 L: Low frequency of occurrence
 N: Non-occurrence

KEY AS ILLUSTRATED FOR TABLE 3E

3.4.5.2.3 Students actively taking part in discussions involving the teacher and the class as a whole

Again there are some odd results concerning the extent to which this activity occurs in lectures. One would have expected it to do so in the more loosely structured 'A' level lecture (non-practical class) but according to the returns from some of the social science students it is also recorded as having a high frequency in their lectures as well.

This is not a view shared by the majority of social science staff (Table 3W1, QuD3a). In general, except in staff returns concerning 'A' level science students, it was an activity that was likely to happen only 'sometimes' or 'never'. It was far less likely to happen in undergraduate as against post-graduate lectures.

This raises the question of size as a determinant of activities in the various events under review. At the risk of generalisation one might argue that the smaller the group the greater the possibility of interaction between staff and students. The large numbers in undergraduate lectures (v. Table 3BB, p.308 below) usually prevented any direct interaction taking place. In those lectures specifically for post-graduates the numbers were often far smaller and lectures would often deviate from the more formal lecture frame, particularly if it was taking place in a classroom as against a lecture theatre.

Student participation in discussions involving the teacher and the class as a whole was an activity which took place to a far greater extent in seminars/tutorials than lectures, for all groups, and particularly social scientists (Table 3W1, QuD3b). In general the staff considered that students took part in these discussions far more than the students themselves claimed to, though they are referring to the class as a whole here. Most of the students claimed that this was an activity they engaged in 'sometimes' or 'often' in seminars, and it was a particularly

frequent activity in this type of class for those students in the social sciences.

This view is to a certain extent supported by the information collected in seminars during the observations and recorded in Table 3V2(b) above.

In practical classes (Table 3W1, QuD3C) this was a fairly frequent activity for those who had them, responses falling mainly in the 'sometimes' or 'often' column.

Overall we noted that there was far less argumentation and group interaction, involving the class as a whole, than one might have supposed in potentially reciprocal speech situations, especially seminars. At undergraduate level and below students are far more likely to be answering and asking questions than taking part in group discussions involving the teacher and the class as a whole.

3.4.5.2.4 Eliciting transactions

Our concern here was with the frequency with which students directed their questions to their teacher rather than to fellow students. For the majority of students in lectures this activity occurred 'sometimes' rather than 'often' (Table 3W1, QuD4). At undergraduate level, with the exception of the overseas social scientists, a large number of the students claimed that they 'never' asked teachers questions in lectures. Except for 'A' level science lectures (non-practical classes) this was not a frequent activity, confirming the view expressed above, that the lecture situation is not a directly interactive teaching situation for the majority of students in our survey. Staff on the whole considered that students asked far more questions in lectures than the students themselves admitted to. Their answers were for the student group as a whole though, whereas the students were only answering for themselves.

In general, any questions that were asked, except in the case of 'A' level science and certain post-graduate lectures observed, were normally addressed to the teacher at the end of the session rather than punctuating the monologue during the teaching period itself.

The overall picture of the infrequency of this activity in lectures is supported by the observations (v. Table 3W2 below).

In seminars and tutorials a large number of students claimed that they frequently asked the teacher questions. This was particularly so in the social sciences and undergraduate science and engineering. In general, undergraduates were more likely to ask questions in these periods than post-graduates. Staff returns indicated a greater frequency of occurrence than those of the students, but again their answers relate to all the students in the teaching situation, whereas the students had been answering only in respect of whether they carried out the activity themselves. Very few staff or students viewed this as an activity which 'never' took place.

Both overseas and British students saw themselves as asking teachers questions 'sometimes' or 'often' during practical classes where these applied. It was seen by staff to be a particularly frequent activity at engineering undergraduate level and science 'A' level.

These returns accord with the information we recorded during the observations. In Table 3W2 below, we have summarised the frequency of the various types of eliciting transaction that took place in the various teaching events.

3.4.5.2.5 The frequency with which the teacher asks the class questions

This has been dealt with earlier in Section 3.4.2 on listening comprehension activities (v. Table 3G1, p.169 above).

TABLE 3W2 THE FREQUENCY OF OCCURRENCE IN OBSERVATIONS OF STUDENTS REFERRING TO THE TEACHER FOR THE PURPOSE OF:

	Eng. U	Sci. U	Sci. P	Sci. A	S.Sci U	S.Sci P
1. acquiring or clarifying facts or principles	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
2. seeking guidance when identifying or solving problems	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
3. seeking guidance when making inferences, formulating or testing hypotheses	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
4. seeking guidance on experimental procedure	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N
5. developing own opinion or line of argument	a	H M L N	H M L N	H M L N	H M L N	H M L N
	b	H M L N	H M L N	H M L N	H M L N	H M L N
	c	H M L N	H M L N	H M L N	H M L N	H M L N

a. lectures
 b. seminars/tutorials
 c. practical classes

H: High frequency of occurrence
 M: Medium frequency of occurrence
 L: Low frequency of occurrence
 N: Non-occurrence

KEY AS ILLUSTRATED FOR TABLE 3E

It is worth reiterating that the highest frequency of occurrence of this activity was found to occur in seminars and tutorials, to a lesser extent in practicals, and in general seldom featuring other than rhetorically in lectures.

This further confirms the view, expressed above, that only in seminars/tutorials, and to a lesser extent practical classes, is there any demand placed on oral interactive skills. Even in these classes a greater part of the discourse appears to be initiated by the teacher, or alternatively students address their comments and questions to him, rather than to their fellow students.

The returns in this section of the questionnaire and the information gathered from the observations point to the teacher-centred nature of academic events across disciplines and the noticeably low level of student-student interaction except in those cases where students are working with each other as in practical classes. The high frequency of teacher making statements, even in seminars and tutorials, that we noted in Table 3E, page 162 above, also supports this contention.

One might conclude that oral interactive skills are not as important in terms of frequency of occurrence as the other three skills in the academic context. This is a view which is supported by the most recent N.A.F.S.A. survey (Lee, M.Y. et al. 1981) in the United States and by James (forthcoming).

3.4.5.3 Problems Experienced in Spoken English

In the student questionnaire we were able to ask the students directly (QuD6) about the amount of difficulty they had experienced because of language in the skills listed below:

- QuD6/1 Giving oral reports or short talks.
- QuD6/2 Asking teachers questions.
- QuD6/3 Asking other students questions.
- QuD6/4 Answering questions asked by teachers.

- QuD6/5 Answering questions asked by other students.
- QuD6/6 Working with other students using English to communicate.
- QuD6/7 Expressing their own opinions in discussions.
- QuD6/8 Explaining their opinions when they are not immediately understood in discussions.
- QuD6/9 Expressing counter-arguments to points raised by other students in discussions.
- QuD6/10 Expressing counter-arguments to points raised by teachers in discussions.

The students' replies to these questions are recorded in Table 3X below.

The staff were asked to indicate the proportions of both British and overseas students on the courses they taught who seemed to experience difficulty in:

- QuD6/1 Giving oral reports or short talks.
- QuD6/2 Asking the teacher questions.
- QuD6/3 Asking other students questions.
- QuD6/4 Answering the teacher's questions.
- QuD6/5 Answering questions asked by other students.
- QuD6/6 Working with other students using English to communicate.
- QuD6/7 Expressing their own opinions in discussions.
- QuD6/8 Explaining their opinions when they are not immediately understood in discussions.
- QuD6/9 Expressing counter-arguments to points raised by other students in discussions.
- QuD6/10 Expressing counter-arguments to points raised by the teacher in discussions.

The staff replies are summarised in Table 3Y1 below.

Because of the different nature of the questions we had to ask staff and students, their responses are not directly comparable; the students are talking about the amount of difficulty they individually experienced with each activity, whereas the staff are talking about the proportion of the students on their courses who experienced difficulty (amount unspecified).

TABLE 3Y2 NUMBERS OF STAFF WHO REGISTERED ANSWERS IN THE 'DON'T KNOW' CATEGORY OR LEFT THE BOX BLANK FOR QUESTIONS IN THIS SECTION:

KEY AS ILLUSTRATED FOR TABLE 3F1

DK: do not know

B1: blank

	QuD6/1	QuD6/2	QuD6/3	QuD6/4	QuD6/5	QuD6/6	QuD6/7	QuD6/8	QuD6/9	QuD6/10
	DK B1	DK B1	DK B1	DK B1	DK B1	DK B1	DK B1	DK B1	DK B1	DK B1
Eng. U	φ Δ		φ Δ		φφ ΔΔ			φ Δ	φ Δ	
Eng. P	φ Δ Δ		φ Δ		φ Δ					
Sci. U	φ Δ Δ		φφ Δ		φφ ΔΔ	φ Δ Δ	φ Δ	φ Δ	φ Δ Δ	φ Δ
Sci. P	φφ ΔΔ		φ Δ Δ		φφ Δ Δ	φ Δ	φ Δ		φφ Δ	φ Δ
Sci. A	φ Δ				φ Δ				φ	
S.Sci. U			φ		φ					
S.Sci. P	φ Δ Δ		φ Δ	Δ	φ Δ Δ	φ Δ				

In the discussion below we make reference to the returns from both the student and staff questionnaires concerning each of the activities, in spoken English, looking in turn at:

- a) the amount of difficulty students experienced in each activity (Table 3X).
- b) the proportion of their students that staff considered to have had difficulty with the same activity (Table 3Y1).

Question D6/1 Giving oral reports or short talks:

Most overseas students claimed that they experienced 'some' to 'very little' difficulty with this. The overseas social science students seem to have experienced least difficulty of all the groups. However overall this was the speaking activity with which overseas students claimed they had the most difficulty (v. Appendix 3.4, p.813) though this must be weighed against its relative infrequency for a large number of the overseas students. (With hindsight it might have been better to include a 'not applicable' or a 'don't know' category here because of this relative infrequency, as a number of the estimates may have been mere guesswork). The British students do not seem to have been unduly troubled by this activity (Table 3Y1, QuD6/1).

Staff considered that greater proportions of overseas students than British experience difficulty in this activity, (Table 3Y1, QuD6/1) though a lot of the staff recorded answers in the 'don't know' category or left the box blank (v. Table 3Y2). Overall, it was for this activity that the lowest number of staff recorded that 'none' of their students had problems.

Question D6/2 Asking teachers questions:

This was seen as causing 'very little' or 'no' difficulty by most overseas students and again students in the social sciences claimed to have experienced the least difficulty. The overseas students admitted to having slightly more difficulty than the British students (Table 3X, QuD6/2).

The majority of the staff considered that only 'very few' or at most 'some' of their overseas students and 'very few' or 'none' of the British students had any difficulty here (Table 3Y1, QuD6/2).

Question D6/3 Asking other students questions:

This was one of the activities which overseas students claimed to have had the least difficulty with in this section (Table 3X, QuD6/3). It was the activity which created the fewest problems for the British students.

A large number of staff did not feel able to answer this question (Table 3Y2, QuD6/3) but those who offered an opinion considered that it did not create difficulty for most of the students they were responsible for.

Question D6/4 Answering questions asked by teachers:

This was seen by the majority of the students from overseas to create 'very little' or 'some' difficulty. It was claimed by the majority of British students, that 'answering questions asked by teachers' created 'very little' or 'no' difficulty (Table 3X, QuD6/4).

On the whole, staff thought that 'very few' to 'some' of their overseas students had problems in this area. They thought that a larger proportion of British students had difficulty with this than they had had in most of the other categories though still less than the overseas students.

Question D6/5 Answering questions asked by other students:

The student returns show that this caused 'very little' or 'no' difficulty for the majority of students, overseas or British. The overseas students experienced a slightly higher level of difficulty with this (Table 3X, QuD6/5).

A large proportion of the staff felt unable to answer this question (Table 3Y2, QuD6/5) but the majority of those who did so considered that 'answering questions asked by other students' created difficulty for 'some' or 'very few' of the students, overseas or British, with slightly more of the former having problems (Table 3Y1, QuD6/5).

Question D6/6 Working with other students using English to communicate:

The overseas students claimed only slightly more difficulty than the British students with this and it was seen as causing 'very little' or 'no' difficulty by a majority of the students in this group (Table 3X, QuD6/6).

Of those staff replying to this question the majority of staff thought that only 'very few' or 'some' of the overseas students had difficulty here. They thought that 'none' to 'very few' of the British students had problems in this. Overall, the staff considered that, next to 'asking other students questions', the problem of 'working with other students using English to communicate' affected fewest students (Table 3Y1, QuD6/6).

Question D6/7 Expressing their own opinions in discussions:
All the activities which relate directly to group discussions (D6/7,8,9,10) seemed to create great problems for the overseas students (v. Appendix 3.4, p.813).

As regards 'expressing their own opinions in discussions,' the majority of overseas students across disciplines and levels saw it as causing them 'very little' to 'some' difficulty. More overseas students in the social sciences than in any other subject group claimed that it caused them 'no' difficulty. A majority of the British students claimed that it caused them 'very little' or 'no' difficulty (Table 3X, QuD6/7).

Of those staff responding to this question the majority saw 'very few' or 'some' of their overseas students as having difficulty in 'expressing their own opinions in discussion'.

They thought that fewer British students had problems with this overall (Table 3Y1, QuD6/7).

Question D6/8 Explaining their opinions when these are not immediately understood in discussions:

This caused quite a problem for a number of the overseas students and for some of the British students as well (Table 3X, QuD6/8). It was one of the spoken English activities that overseas students felt caused them the most difficulty (v. Appendix 3.4, p.813).

Overall, staff considered that, across the range of speaking activities, a very large proportion of overseas students found difficulty in explaining their opinions when they were not immediately understood (v. Appendix 3.4, p.815). It was felt to cause problems for a lot more of the overseas students than for the British ones.

Question D6/9 Expressing counter-arguments to points raised by other students in discussions:

Next to giving oral reports or short talks this oral activity was seen by overseas students as the one causing them the greatest difficulty (Appendix 3.4, p.813). It was seen to be far less of a problem by the overseas social science group who recorded the highest returns in the 'no' difficulty column. The British students seemed to have far less difficulty with this, the majority recording their answers in the 'very little' or 'no' difficulty columns (Table 3X, QuD6/9).

Of those staff replying to this question, the majority saw the ability to express counter-arguments to points raised by other students in discussion as causing 'some' or 'a lot' of their students difficulty. It seemed that more students in the 'A' level group than in any other suffered problems here. It was mainly regarded as a source of difficulty for 'some' or 'very few' of the British students by the staff (Table 3Y1, QuD6/9).

Question D6/10 Expressing counter-arguments to points raised by teachers in discussions:

This was seen by overseas students as one of their greatest problems (Appendix 3.4, p.813). It was the activity which fewest of the overseas students claimed to have had 'no' difficulty with. It was also seen by the British students as their greatest problem though the overseas students admitted to a far greater degree of difficulty than did the British students. (Table 3X, QuD6/10).

Overall most of the staff thought that 'some' or 'a lot' of their overseas students had problems in this area. Again it was the 'A' level science teachers who saw the largest proportion of their students as having a problem. This was also the area where the largest number of staff thought that 'a lot' of their students had problems (overseas and British) (Appendix 3.4, p.815).

We were also able to ask students a further series of questions about factors affecting their spoken English performance. In Question D7 of the students' questionnaire we asked respondents to indicate how often the following caused difficulty in spoken English activities:

- QuD7/1 Thinking out how to say what they wanted to say quickly enough.
- QuD7/2 Worrying about saying something in case they made a mistake in their English.
- QuD7/3 Not knowing how to say something in English.
- QuD7/4 Not knowing the best way to say something in English.
- QuD7/5 Not knowing the subject well enough to answer questions.
- QuD7/6 Finding it hard to enter the discussion.

The summary of returns to these questions are recorded in Table 3Z below.

Question D7/1 Thinking out how to say what they wanted to say quickly enough:

Overall, this seemed to be a frequent cause of difficulty for the overseas students and under 20% said that they 'never' had difficulty with this whereas nearly 40% of the British students surveyed claimed that this was 'never' a problem. Nearly 20% of all the overseas students considered that this was a problem which occurred 'often' (v. Appendix 3.4, p.817). Overseas science post-graduates and 'A' level science students experienced this difficulty the most often (Table 3Z, QuD7/1).

Question D7/2 Worrying about saying something in case they made a mistake in their English:

This was the category in which most overseas students thought they 'often' had a problem. Only about 30% claimed that they 'never' worried about making a mistake in their English as against over 80% of the British students (v. Appendix 3.4, pp. 817-818). Again it was the overseas science 'A' level and post-graduate students, who experienced this problem the most often (v. Table 3Z, QuD7/2).

Question D7/3 Not knowing how to say something in English:

For the majority of the overseas students this was only a problem 'sometimes'. This was the category that the largest number of overseas students said that they 'never' had a problem with and fewest said that they had one 'often'. British students rarely found this to be a cause of difficulty and nearly 90% claimed they 'never' had any difficulty here (v. Appendix 3.4, p.818). The overseas social science students seem to have had less trouble with this than students in other areas (Table 3Z, QuD7/3).

Question D7/4 Not knowing the best way to say something in English:

This was a much more frequent problem for the overseas students than the British students. The majority of the overseas students felt they encountered this problem 'sometimes' or 'often' and only about 20% overall 'never' had any difficulty with it (v. Appendix 3.4, p.817).

Question D7/5 Not knowing the subject well enough to answer questions:

This was the category in which there was very little difference in the degree to which British and overseas students felt they had this problem (Table 3Z, QuD7/5). Overall, fractionally more overseas students than British said that they 'never' had a problem with this (v. Appendix 3.4, pp.817-818). It was seen as being a problem 'sometimes' by the majority of the students in the survey.

Question D7/6 Finding it hard to enter the discussion:

A number of the British students thought they had a problem in doing so 'sometimes' though overall it was a more common problem for the overseas student (Appendix 3.4, pp.817-818). It seemed to be less of a problem for overseas post-graduates and undergraduate social scientists (Table 3Z, QuD7/6).

We also asked staff and students to specify in Question D6/11 whether they had experienced any other general difficulties in spoken English. Staff replies were collected together and an anthology of their comments is listed below:

Socio-Cultural

"The difficulties of using language of some overseas students is not merely technical. Some are lucid in written work yet silent in class; this seems to stem in large part from the rather passive role students are expected to adopt in some countries' institutions."

"Many foreign students have been totally discouraged from discussing/arguing with a teacher. A sign of disrespect."

Reticence

"Reluctance to volunteer information or opinion or to ask questions either of me or of other students in class."

"Their main problem seems to be in communicating to others. They do not readily talk to other students and are sometimes hesitant in talking to me!"

"Many students, British and overseas, are reluctant to enter into long discussions."

"Students do not like answering questions in lectures though they do become less terrified as time progresses. Foreign students are noticeably less anxious to draw attention to themselves by answering."

"Overseas students lack confidence in expressing their views."

Working together

"Overseas students tend to work in lingual groups."

Speed

"There is a greater problem in spoken English (as opposed to written) because of the difficulty in thinking quickly and then phrasing in English something which they have not been able to prepare i.e. responding quickly in the discussion/argument/counter-argument situation."

Pronunciation

"A strong accent may mean they are not understood even though their actual English is good."

"Difficulty in pronouncing many common English words."

"From the lecturer's point of view, it is sometimes difficult to understand the English spoken by students from the Far East. Some seem to find difficulty with some of the sounds - such as 'r' and 'l'. On the other hand, I am always surprised at how little difficulty foreign students seem to have in coping with the various accents of the British staff and students."

Subject difficulties

"Both British and overseas students experience difficulties in the use of mathematical grammar and vocabulary."

"Difficulty in giving illustrative examples from their own experience."

"Difficulty with common technical terms."

"Depends whether they know the answers."

Overall Deficiency

"Overseas students are often as good (many are better) as British students. The main problem is an overseas student whose knowledge of English is so bad that he should not be on the course anyway."

Applicability

"Most of your detailed questions do not apply to mathematics as we teach it."

"This question D6 is difficult to answer as it has little relevance to a subject like mine. Fluid mechanics is taught, like mathematics, authoritatively and students rarely have occasion to ask questions of other students. There is little discussion - the subject is one in which you are right or wrong - there is no grey area."

As well as asking students in Question D6/11 whether they had experienced any other general problems we also asked in Question D7/7 if they had any other specific difficulties,

and how often these occurred. As the general and specific seemed to merge in the replies to these two open ended questions we have amalgamated the answers and an anthology of the replies is recorded below under various categories. The comments are taken from the replies of the overseas students only, few British students making any comments here other than to draw attention to the difficulties they had in understanding overseas students or to their own physical disabilities.

Listening Comprehension

"Not understanding what is being said to me."

Problems with the addressee

"If the person that one is speaking to is not so fluent in the English language, one's own spoken English has the tendency to deteriorate also."

"On informal occasions, when I address strangers, I am not always readily understood."

"Knowing if the other person has really understood what I meant to say. Difficulty depends on the person I am speaking to."

Personal Factors

"Feeling self-conscious when I have to speak."

"No confidence when faced with native speakers."

"Not wishing to speak for fear of being misunderstood."

"Having enough confidence to make a point (not because of my English)."

"Not taking part in discussions may be due to a person being shy or simply not confident enough."

"I have always the feeling that I do not express myself the way I would like to. Also I can't speak very quickly without making mistakes. When I'm tired my ability in speaking the language falls dramatically. Generally my mood and state of mind play a significant role in my efforts to speak the language clearly and correctly."

Group factors

"The other students always laugh on hearing what I have to say, may be due to my spoken English."

"Reluctant to ask questions because of class attitudes."

"People do not allow foreigners to enter the discussion, especially English people, who do not want to communicate with overseas students."

"Problem of entering into discussion easily."

"Knowing whether to enter the discussion or not."

"There is some difficulty in working with other students using English to communicate especially when the other students are all English and when they talk very fast."

Lack of opportunity

"Apart from technical reading we have had no chance to practise daily usage intensively before coming to U.K."

"Not many opportunities to speak in English."

Accent/Pronunciation

"Problem of getting the English to understand my accent, especially the less educated ones. Most experienced lecturers have no such problem especially if they have lived abroad."

"Can't make myself understood due to bad pronunciation and accent."

"Some people cannot understand me unless I speak very slowly."

"Worrying about the correct pronunciation of some words."

"I speak in a clear voice and I pronounce my words quite well, yet very often I find I have to repeat things several times in order to be understood. This could perhaps be due to my accent although it is not very strong, and more American than Dutch."

Speed

"Difficulty in speaking quickly."

"Difficulty in answering when other person speaks very fast or with a strong accent."

"Difficulty when under immediate stress."

Idiom

"Difficulty understanding and replying to idiomatic expressions."

Clarity

"Expressing myself so that they can understand exactly what I mean."

Appropriacy

"Appropriate words do not come spontaneously."

"I get stuck sometimes trying to find the right word and have to look for it in Arabic first then translate it into English."

"Use of appropriate words to describe certain situations or actions."

"I think the main difficulty is that I cannot say what I really want to say."

"Finding the most suitable adjective on the spur of the moment."

Grammar

"My only problems are with tenses."

"Difficulty using the right grammar."

Subject Content

"A lack of confidence in using the 'jargon' of the subject may stop me asking questions."

"Some mathematical terms."

"Difficulty in talking about special topics."

"Sometimes I don't read therefore I am unprepared for the discussion."

"When the topic discussed is too advanced and requires a lot of mathematical treatment."

"Difficulty explaining complicated theory in physics."

Social English

"On informal occasions, I am not always welcome (effectively) to enter a discussion because I cannot keep up with the talking speed, especially with some intellectuals who are used to (and enjoy!) speaking very fast and using fancy words."

"I have problems in expressing feelings and emotions."

"Problem in talking at 'social level' (outside academic subjects)."

"People don't want to chat with overseas students with poor English."

"This is especially so outside the university, e.g. if I want to ask anyone in the street about a particular place or shop I find it difficult to make him or her understand what I say because of my pronunciation and also because the people in the street do not speak grammatical English."

"I am sure I had improved my English practically since I had already passed my 'O' level English but I found most people from my country, they are capable of understanding English and all written work BUT some could not even have conversations with friends in English because they are afraid of their broken English."

3.4.5.4 Constituent Enabling Skills

An argument can be made that any E.A.P. test(s) of oral production should reflect the parameters that have emerged from the research and one should accordingly test oral ability in an integrated manner. However, if a more discrete approach is favoured it may be useful to have before us a checklist of the constituent enabling skills which appear together to contribute to overall ability in spoken English. Again we draw heavily upon Munby's (1978) list of microskills as a major informing source.

1. The use of stress in connected speech.
2. Expressing attitudinal meaning.
3. Expressing conceptual meaning.
4. Expressing information explicitly.
5. Expressing information implicitly.
6. Indicating the main point or important piece of information in a piece of discourse through vocal underlining, end focus and end weight, verbal clues.

7. Using indicators in discourse for introducing an idea, developing an idea, transition to another idea, concluding an idea, emphasising a point, explaining or clarifying points already made, anticipating an objection or a contrary view.
8. Initiating in discourse: starting off the discourse (eliciting, informing, directing, etc.), introducing a new point (using verbal and vocal clues), introducing a new topic (employing appropriate microfunctions such as explanation, hypothesis).
9. Maintaining the discourse: responding (acknowledging, replying, agreeing, disagreeing, etc.), continuing (adding, exemplifying, justifying, evaluation, etc.), adapting, as a result of feedback, especially in mid-utterance (amplifying, omitting, reformulating), turntaking (interrupting, challenging) marking time (stalling, breathing space, etc.).
10. Terminating in discourse: marking boundaries in discourse (verbal and vocal clues), coming out of the discourse (excuse, concession, etc.), concluding of a topic (using appropriate microfunctions such as substantiation, and verbal clues for summing up).
11. Planning and organising information in expository language using rhetorical functions especially description of properties, description of process, description of changes of state, narrative and argumentation.

3.5 SETTING

In Table 3AA below, on the basis of information gathered during the observations, we have summarised details of the academic setting for the various subject groupings and levels.

In section one we have indicated the physical environments in which students of the various disciplines had to operate.

In section two we have given an indication of the spread of hours students spent each week in the various learning situations.

Section three gives an impressionistic account of the type of psychosocial environment the students as a whole operated in.

TABLE 3AA THE ACADEMIC SETTING

1. PHYSICAL SETTING: SPATIAL

1.1 The academic study settings for which English is required:	ENG. U	SCI. U	SCI. P	SCI. A	S.SCI U	S.SCI P
1.1.1 lecture room/theatre	■	■	■	■	■	■
1.1.2 classroom	□	□	□	■	■	■
1.1.3 laboratory	■	■	■	■	□	□
1.1.4 workshop	■	□	□	□	□	□
1.1.5 seminar/tutorial	■	■	■	□	■	■
1.1.6 private study/library	■	■	■	■	■	■

2. PHYSICAL SETTING: TEMPORAL

2.1 Length of course

2.1.1 hours per week student in:

lecture classes	10-18	8-16	4-18	9-14	2-20	0-14
seminars/tutorials	2-7	1-2	2		1-7	4-8
practical classes	5-17	2-12	11-12	11-6		

3. INVENTORY OF PSYCHOSOCIAL ENVIRONMENTS

	1	2	3	4	5	
3.1 intellectual/thinking	■	■	3	4	5	non-intellectual/unthinking
3.2 educationally developed	■	■	3	4	5	educationally undeveloped
3.3 familiar physical	■	2	3	4	5	unfamiliar physical
3.4 familiar human	■	■	■	■	5	unfamiliar human
3.5 quiet	■	■	■	■	5	noisy
3.6 demanding	■	■	■	4	5	undemanding
3.7 hurried	1	■	■	■	5	unhurried
3.8 formal	■	■	■	■	5	informal
3.9 authoritarian	1	■	■	■	5	unauthoritarian/laissez-faire
3.10 entertaining/festive	1	■	■	■	5	serious
3.11 sympathetic	■	■	3	4	5	unsympathetic

Shaded areas indicate occurrence of a feature

3.6 INTERACTION

Table 3BB illustrates the type and number of people with whom students in the various disciplines might interact and the possible social relationships that could occur as a result. The profiles are based on information gathered during the series of visits made in 1979-1980 (v. Appendix 3.2.2, pp.690-693).

TABLE 3BB INTERACTION IN THE ACADEMIC CONTEXT

1. Position	Students
2. Role Set (people with whom they interact in English, by virtue of their "position")	2.1 lecturers, teachers, tutors, supervisors, laboratory/workshop staff 2.2 fellow students 2.3 official contacts on industrial and commercial visits 2.4 writers of books, papers, articles, handouts (where different from 2.1)
3. Role Set Identity	ENG. U SCI. U SCI. P SCI. A S.SCI U S.SCI P
3.1 Number:	
Lecture	25-200 50-180 10-50 10-22 23-160 10-20
Tutorial and Seminar	8-15 5-12 2-6 4-14 4-20
Practical Classes	5-40 60-85 10-22
3.2 Age Group	18+ 18+ 22+ 16+ 18+ 22+
3.3 Sex	Mixed Mixed Mixed Mixed Mixed Mixed mainly mainly mainly sometimes male male male more women
3.4 Nationality	Mainly British except at post-graduate level where there are often more non-native speakers
4. Inventory of Social Relationships that occur as a result of role relationships in 2.	
Social relationships - asymmetrical	4.1 instructor/authority - learner + 4.2 adviser - advisee + 4.3 insider - outsider + 4.4 native - non-native + 4.5 older generation - younger generation + 4.6 man/male - woman/female +
Social relationships - symmetrical	4.7 equal - equal 4.8 insider - insider 4.9 group member - group member 4.10 friend - friend 4.11 acquaintance - acquaintance 4.12 outsider - outsider 4.13 adult - adult 4.14 own generation - own generation 4.15 own sex - own sex

3.7 INSTRUMENTALITY

In Table 3CC we have summarised the medium, mode and channel of communication students have to operate in.

TABLE 30C INSTRUMENTALITY IN THE ACADEMIC CONTEXT

	ENG. U	SCI. U	SCI. P	SCI. A	S.SCI U	S.SCI P
1. Medium						
1.1 spoken : receptive	■	■	■	■	■	■
1.2 spoken : productive	■	■	■	■	■	■
1.3 written : receptive	■	■	■	■	■	■
1.4 written : productive	■	■	■	■	■	■
2. Mode						
2.1 monologue, spoken to be heard	■	■	■	■	■	■
2.2 monologue, spoken to be written (dictation)	■	■	■	■	■	■
2.3 monologue, written to be read	■	■	■	■	■	■
2.4 monologue, written to be spoken (oral reports)	■	■	■	■	■	■
2.5 dialogue, spoken to be heard	■	■	■	■	■	■
3. Channel						
3.1 face-to-face (bilateral)	■	■	■	■	■	■
3.2 face-to-face (unilateral)	■	■	■	■	■	■
3.3 radio (live relay)	■	■	■	■	■	■
3.4 television (live relay)	■	■	■	■	■	■
3.5 tape (audio/video)	■	■	■	■	■	■
3.6 film	■	■	■	■	■	■
3.7 print (unilateral)	■	■	■	■	■	■
3.8 print (bilateral)	■	■	■	■	■	■
4. Non-verbal medium						
4.1 pictorial (unlabelled illustrations, charts and plans, technical drawings)	■	■	■	■	■	■
4.2 mathematical and other scientific, arithmetical, algebraic, geometrical, chemical symbols	■	■	■	■	■	■
4.3 kinesic (facial and body movements or gestures)	■	■	■	■	■	■
4.4 realia (e.g. specimens, scientific equipments)	■	■	■	■	■	■

Shaded areas indicate occurrence of a feature.
 R: Receptive *Low frequency of occurrence
 P: Productive

3.8 CONCLUSION

The data on the General Descriptive Parameters of Communication, Task Dimensions and Dynamic Communicative Characteristics, collected through the procedures described above, provided us with an empirical base for the construction of our Test in English for Academic Purposes (T.E.A.P.).

In Chapter 4 below we examine the feasibility of realising this specification in a proficiency test appropriate for students coming to study in the United Kingdom at a variety of levels and in a variety of disciplines.

C H A P T E R F O U R

TEST DESIGN AND CONSTRUCTION

CHAPTER FOUR CONTENTS

	<u>Page</u>
4. TEST DESIGN AND CONSTRUCTION	
4.1 TYPE OF TEST BATTERY NEEDED	315
4.2 FROM SPECIFICATION TO REALISATION	318
4.2.1 The Power of the Data: A Cautionary Note	318
4.2.2 Handling the Data	319
4.2.3 General Descriptive Parameters	322
4.2.4 Dynamic Communicative Characteristics	324
4.2.5 Task Dimensions	325
4.2.5.1 Size	325
4.2.5.2 Complexity	326
4.2.5.3 Functional/Referential Range	326
4.2.6 A Further Problem in Text Selection	327
4.2.7 The Problem of Format	330
4.3 ASSESSING READING SKILLS IN AN E.A.P. CONTEXT	332
4.3.1 General Descriptive Parameters	332
4.3.2 Formats for Testing Reading Comprehension	333
4.3.2.1 Multiple Choice Techniques	333
4.3.2.2 Short Answer Questions	338
4.3.2.3 Cloze and Gap-Filling	340
4.3.3 Testing the Range of Skills	346
4.3.4 Variety of Format	346
4.3.5 Conclusions	347
4.4 ASSESSING LISTENING SKILLS IN AN E.A.P. CONTEXT	349
4.4.1 General Descriptive Parameters	349
4.4.2 A New Paradigm	351
4.4.3 Formats for Testing Listening Comprehension	353
4.4.3.1 Job Sample Tests: Monologue and Interactive Discourse	353
4.4.3.2 Dictation and Listening Recall	360

	<u>Page</u>
4.4.4 Conclusions	366
4.5 ASSESSING WRITING SKILLS IN AN E.A.P. CONTEXT	367
4.5.1 Why Test Writing Skills?	367
4.5.2 Formats for Testing Written Production in an E.A.P. Context	369
4.5.2.1 Objective Formats	371
4.5.2.2 Integrative Tasks	374
4.5.3 The Assessment of Written Production: A Problem of Reliability	378
4.5.3.1 Impression Versus Analytical Approaches	378
4.5.3.2 Establishing Appropriate Criteria for Assessing Written Production in T.E.A.P.	390
4.5.3.3 Further Considerations in Designing Writing Tasks for Inclusion in a Test Battery	395
4.5.3.3.1 Number of writing tasks	395
4.5.3.3.2 Question choice	396
4.5.3.3.3 Amount of time allowed for each writing task: The ramifications of time limits	397
4.5.3.3.4 Checking the viability of the writing tasks	399
4.6 CONCLUSIONS	401

4. TEST DESIGN AND CONSTRUCTION

4.1 TYPE OF TEST BATTERY NEEDED

In Chapter 3 we described the communicative demands that were placed on students in the various study modes they encountered in the academic context and drew attention to the relative difficulty levels experienced by both native and non-native speakers in coping with these. Having established these frames of reference we now turn to consideration of the form the pre-test version of T.E.A.P. should take.

Our investigations showed that there was a need for assessing the communicative competence of non-native speakers of English in ways which would usefully show up any deficiencies in English before the commencement of their academic courses. This was closely connected to the role to be played by those responsible for running the pre-sessional and in-sessional courses in English for Academic Purposes (E.A.P.). They needed profiles of students' language abilities, in the various study modes so that, where possible, those not meeting the required standards could be allocated to reasonably homogeneous, remedial classes. This chapter is concerned with establishing the nature of the test battery required to construct such a profile of the student's productive and receptive abilities.

A feature of recent research in this field has been the limited attention paid to procedures for testing the productive skills. Most attention has been paid to the testing of listening comprehension (cf. Holes 1972; Templeton 1973; Morrison 1974 and Kelly 1978), reading comprehension (cf. Sim 1974; Engineer 1977; Alderson 1978a; Klein-Braley 1981 and Ng'ombe 1981) and to more traditional tests of discrete levels of vocabulary and grammatical structures (v. Chaplen 1970).

With the exception of Chaplen (1970) and Hawkey (1982) who experimented with essay sub-tests, researchers in the language testing field seem, for the most part, to have neglected the assessment of written production (v. Davies 1978). Given the importance of writing

in particular, this has been a serious omission from the E.A.P. standpoint. Moller (1977, p.32) commented:

"More satisfactory ways of testing the productive skills (speaking and writing) should be developed ... If oral communication in a variety of academic and non-academic situations and the writing of papers and essays present major problems for so many overseas students, the refinement of assessment procedures is most important."

In Section 2.1 we discussed the reasons why we are initially omitting the design of an oral component from T.E.A.P. To a certain extent we were justified in this because of the lesser need for an oral component, shown up by the data collection exercises described in Chapter 3 above and by Johns, A.M. (1981), Lee, M.Y. et al. (1981) and James (forthcoming). There was, of course, no question of our omitting tests of productive skill in order just to safeguard test reliability; to do so would have threatened the validity of our E.A.P. test battery. As we argued in Section 2.1, given the high content and face validity of tests of the productive skills in E.A.P., there was a need to investigate ways of improving their reliability and streamlining their administration rather than avoiding the problems. Given the limited time and resources available we decided to concentrate our efforts on establishing valid and reliable methods for assessing the written production of overseas students.

We argued above that in order to test communicative competence in an academic context there was a need for:

- (a) an analysis of the communicative tasks facing overseas students in their courses of study and an estimate of the difficulty encountered in coping with these performance tasks,
- (b) the application of valid and reliable testing procedures which, as far as possible, took account of these parameters.

The steps taken to carry out the first of these are described in Chapter 3. It is the purpose of this chapter to see how closely we can achieve the second, in which we are concerned with examining how

far we can realise our specification in the form of a reliable, valid, practicable test battery for assessing reading, listening and writing abilities in the various academic study modes under consideration.

4.2 FROM SPECIFICATION TO REALISATION

4.2.1 The Power of the Data: A Cautionary Note

The data obtained from the observations were, in certain important respects, different from those gathered through the questionnaire survey. The data from the observations were recorded in respect of the occurrence of activities within a single teaching period or over a limited number of teaching periods in a course and are thus very much dependant on the nature of the lesson(s) observed. Further variation occurred when we observed the same type of period, in different courses on the same programme and looked at similar programmes in different institutions. These different observations were collapsed in Chapter 3 to show a composite variation in the occurrence of a particular activity within study modes, in each broad discipline area. It is pertinent to remember that had we been able to observe a greater number of periods in the same programme, or across programmes, then in all likelihood a greater variation in the frequency of occurrence would probably have been recorded.

The main value of the data gathered in the observations was that they provided us with the requisite information for framing our categories in the questionnaire. It was thus an essential pilot activity; it also enabled us to check on the data we received back through the questionnaire.

The questionnaire enabled us to ask the students to estimate the frequency with which certain events and activities occurred in respect of the total programme of study they were enrolled on; it also provided us with a general estimate of the amount of difficulty various activities and constraints caused them. The staff were asked to give an overall impression of the frequency of occurrence of various activities with particular reference to the courses they taught, on the programme we had specified. They were also asked to estimate the proportion of overseas and British students on these courses, who had encountered difficulty with particular activities or

under particular constraints. The data from the returns to the questionnaire, in terms of estimates of frequency and difficulty occasioned, provided us with a much broader basis for taking decisions about what to include in our tests.

4.2.2 Handling the Data

We decided to treat differently the data from the questionnaires relating to the difficulty of various activities experienced by overseas and British students on the one hand and those associated with the frequency of occurrence of these activities on the other. Because the difficulties students encountered were likely to vary as a result of differences in individual language ability rather than arising solely from the nature of the particular course they were following, it was appropriate to aggregate these data across the disciplines without losing meaningful information. This view was supported by the data themselves which illustrated that a greater frequency of an activity in a particular discipline area did not necessarily occasion a greater degree of difficulty in performing that activity.

In contrast, the frequency of occurrence of an activity did depend to a very large extent upon the discipline area and even more so on the particular course being followed. The time spent on any one activity differed from course to course.

A complication in interpreting these data arose from the fact that there were different numbers of respondents to the questionnaire in respect of courses, discipline areas and levels (v. Appendix 3.3.2 p.801). For example, there were replies from 441 overseas 'A' level science students as against only 44 from overseas science undergraduates. In our opinion, to combine frequencies of occurrence across discipline areas and levels would mean discarding much valuable information as well as creating false aggregates (v. King 1980). For these reasons we aggregated only the data on frequency of occurrence within each of the discipline areas and felt that any further aggregation would be pointless.

A further problem arose in dealing with the data concerning the various constraints and activities. Although we were able to collect data in respect of the level of difficulty these activities and constraints caused, it was often not possible to collect data on the frequency with which the constraints and the activities occurred.

This problem would seem to arise from the nature of the language skills themselves. One might consider reading and writing as non-serial and, by their permanent nature, they allow for random access, i.e. one can go back to recover, check, etc. any earlier information one wants. Speaking and listening would seem to be serial activities and not usually susceptible to replication. The essentially ephemeral nature of these macro-activities means that additional performance constraints affect their operation. In terms of the type of question that we were able to ask in the questionnaire, therefore, reading and writing differed from listening and speaking. It was easier to ask a more complete set of frequency questions about the former because they are relatively free of the performance constraints that affect the latter grouping.

It also seems that a student might reasonably be expected to say how often he has to write at lengths greater than a paragraph, whereas it is much more difficult to quantify the length of spoken contributions. For these reasons there are far more blanks in the frequency sections of the summary tables (v. Appendix 3.5, pp.821-826), where we deal with the constraints and activities involved in listening in the academic context.

All the information we succeeded in collecting concerning frequency, difficulty and importance was condensed on to single sheets in respect of each of the macro-skills (for the purposes of analysis treated here separately). On the basis of these summary sheets (v. Appendix 3.5, pp.821-826) and the supporting information presented in Chapter 3, we were in a better position to take decisions about the more performance oriented activities we might wish to include in the integrated, communicative components of our test battery.

From the results of our investigations and a survey of the literature we had also drawn up a list of enabling skills for each broad macro-skill area (v. Chapter 3). We subsequently attempted to introduce some hierarchical notion into these lists of enabling skills (v. Tables 4B, p.335, 4D, p.358 and 4F, p.370 below). This ordering was informed by discussions with three groups of post-graduate applied linguistics students, at the Universities of Reading, Lancaster and London and by advice received from the Project Working Party. In the case of reading, we would expect skills 7-14 to subsume skills 1-6 (v. Table 4B, p.335). In listening (v. Table 4D, p.358 below) we think that skills 6-15 are superordinate to skills 1-5 and possession of the former might imply possession of the latter. In writing (v. Table 4F, p.370) skills 8-9 would appear to subsume skills 1-7.

In the pre-testing of T.E.A.P., we wished to include tasks which involved an integration of different macro-skills where for example reading and/or listening activities might act as stimuli for writing tasks. In these integrated tasks we would not be concerned with directly testing the 'discrete' enabling skills that we had been able to identify but rather attempting to simulate the types of communicative activity students might encounter in an academic context.

A problem might arise, however, in determining at what stage the integrated process had broken down. This would seem to demand that we provide separate profiles of reading and listening ability in addition to a profile of writing ability. Certain components of the test battery would need to focus more discretely on the enabling skills that appear to contribute to proficiency in reading and listening. The marks obtained in these more discrete test tasks would be converted into behavioural grades for the reading and listening macro-skills.

We were aware of the problems in advocating one approach to testing in preference to all others (v. Chapter 2). We decided, accordingly, that the only practical solution for an experimental battery would be to trial a variety of formats from various positions along the testing continuum, e.g. 'discrete point' tests, integrative tests

and communicative tests which, a priori, seemed to be the most useful for our purposes, and compare their relative effectiveness. We would investigate how far we could realise, through a variety of test formats, the general descriptive parameters, the dynamic communicative characteristics and task dimensions specified above in Chapter 3.

4.2.3 General Descriptive Parameters

In each Session of the pre-test version of T.E.A.P., it was proposed that there should be a number of performance tasks of the kind identified as important in the E.A.P. situation in which our students would have to operate. In these performance tasks candidates would communicate on their own behalf, interacting as students might in situations which reflected those that they could be exposed to in an academic context. The purposes of the participants in these activities would be similar to those underlying their involvement in the performance events described in Chapter 3, Section 3 above. The activities students were to be involved in would reflect what has to be done to achieve these purposes in an academic context (v. Appendix 3.5, pp.821-826).

With reference to those psychosocial environments earlier identified as important (v. Table 3AA, p.306), the settings for our test tasks might be characterised as intellectual/thinking, educationally developed, quiet, demanding, relatively hurried, formal to fairly informal, fairly serious and sympathetic.

As regards the interactions involved in taking the test, the role set should accord for the most part with those of the specification described in Table 3BB, page 308, e.g. lecturers, tutors, writers of books, articles, except that there would be no interaction with other students in the present trial versions of the test.

The social relationships we intend including would be largely asymmetrical, e.g.:

- | | |
|--------------------|------------------------|
| learner | - instructor/authority |
| advisee | - advisor |
| non-native | - native |
| younger generation | - older generation |

and there would be no symmetrical social relationships, e.g. friend - friend, equal - equal, given the initial absence of an oral test with a social interactive component.

The parameter of instrumentality covering medium, mode, channel, was described in Table 3CC, page 310. As regards media, all are to be covered with the exception at present of spoken productive. All modes would be included in T.E.A.P. except monologue, written to be spoken (oral reports). We shall include tasks which involve monologue, spoken to be heard, monologue, spoken to be written and monologue, written to be read.

As regards channel, the constraints of a test system would start to interfere here. Because of the need for standardised exposure to spoken texts, we would have to rely on tape (audio) in the listening sections rather than face to face channels. This has obvious drawbacks in that it would remove all the exophoric reference and paralinguistic information from the context of situations and make the listening tasks much more artificial and difficult.

We would need to ensure that we include a variety of intelligible accents in the taped exercises so that there is a reasonable cross-section, not unrepresentative of the accents candidates could be exposed to in their courses of study.

As well as more integrated, performance-oriented, job-sample tasks, we would also wish to include in the battery items designed to focus on the range of enabling skills we have identified. Here we would be intent on assessing competence in a more discrete and less direct fashion rather than attempting to simulate target group performance activities and constraints.

It is also proposed that the knowledge of discrete features of the linguistic code, e.g. structural items, should be assessed in the experimental pre-test. We did not attempt to collect data on the linguistic code in our empirical investigations, as it was not directly part of the construct we wished to measure. Nevertheless we were most interested in how our measures would compare with a more

traditional test of knowledge of grammatical structure and, accordingly, a component testing this would be included in the pre-test.

4.2.4 Dynamic Communicative Characteristics

The second set of general, evaluational requirements discussed in Chapter 3 above, was concerned with the dynamics of communication. Given that performance tasks allow of descriptions of who does what, why, where, with whom and how, it should be possible to check their communicative dynamics (v. Chapter 3, Section 1.3. pp.116-120). Because of the current absence of an oral component in our battery, the pre-test version of T.E.A.P. would be limited in terms of these dynamic characteristics (v. Kelly 1978). Nevertheless several of the characteristics identified might be applicable at the test construction stage. We should try to provide a realistic context, in terms of the texts selected and what students are asked to do with them. In those tasks where the students are asked to transfer information from one source to another, there should be an information gap which has to be filled. There should be scope for development of activity by the participants in certain of the tasks set and also allowance should be made for self-monitoring of their own work by the candidates.

We are particularly concerned in the reading and listening activities that participants should be expected to process appropriately sized input and we discuss this more fully below in the section dealing with the task dimension of size of text. Linked to this is the intention that, as far as possible, normal time constraints should be operative and, accordingly, in the listening tests it is intended that the discourse should only be heard once and at normal speed. As regards the written texts students have to handle, we are aware that individuals in real time would approach these idiosyncratically. We feel that we should ensure that the reading tasks we set students can be completed satisfactorily in the time allocated, by their native speaker contemporaries and that both a range of intensive and extensive skills are tested.

4.2.5 Task Dimensions

We consider below the third set of general evaluational requirements discussed in Chapter 3, pp.120-121 above. These are the task dimensions of size, complexity and range. These dimensions are to be used mainly as post facto checks on the test tasks, first priority being given to the communicative relevance of the performance tasks.

4.2.5.1 Size

Engineer (1977), commenting on the reading comprehension components of tests used internally by British universities to assess the English proficiency of first year foreign students, felt that a tacit assumption had been made in many tests, that a few short reading passages represented an adequate sample of the kinds of reading students are required to do at the tertiary level. The ability to understand short passages of about 150 words, single sentences, or even individual grammatical items, was considered to be equivalent to the ability needed to comprehend larger units of continuous prose. Engineer argued for and illustrated the advantages of using longer passages of over 1000 words. She reached the conclusion that a long passage was not only more representative of academic reading in terms of length, but it actually provided more reliable data regarding candidates' reading ability. Employing cloze procedure and multiple-choice items, Engineer showed that the longer the text, the easier it was to discriminate between different levels of reading ability.

We decided that, where the nature of the format allowed the testing of the range of reading skills identified, the lengths of written texts should be about 1500 words. Given that reading to extract specific, assignment-oriented information appeared to be the most important extensive reading task for all students across all levels, the inclusion of formats testing comprehension of longer passages from appropriate text-books, was seen to be a realistic activity. To focus on more intensive reading skills a shorter extract of about 250-300 words would also be included.

With regard to the length of texts for listening comprehension components it was decided that in order to come nearer to replicating the type of spoken text students are exposed to in real life, we should aim to include texts of about 12-15 minutes in length. Whilst realising the impossibility of exposing students to the normal length of spoken discourse they encounter, we felt that these lengths were more realistic than the extremely limited lengths they are exposed to in, for example, the present C.P.E. examination.

4.2.5.2 Complexity

During the observations (v. Chapter 3, Section 2.1.2, pp.125-140) it became clear that the range of complexity of text that students were exposed to even in the same course and the difficulties involved in adjudging complexity, meant that no easy answer was available for the question 'how complex should the texts selected for the test battery be?'

Our tests were to be aimed at a range of levels from G.C.E. Advanced to post-graduate students. The first term of most science and engineering undergraduate courses is spent in ensuring that people with disparate science and engineering Advanced level backgrounds are brought up to a common standard. We decided that if we selected texts at a degree of complexity that students would have to face in the first undergraduate term or at G.C.E. Advanced level, then we could be reasonably sure that they would not be too difficult for post-graduate students in terms of complexity of language and subject content.

Accordingly, we selected texts from appropriate sources at this level and informally checked their suitability with groups of overseas and British students, subject specialists and testing experts.

4.2.5.3 Functional Range/Referential Range

Despite empirical investigation these dimensions, like that of complexity, defied adequate, precise description. For only if we were writing a test for an extremely limited set of students could we precisely specify the nature of suitable texts. It appears, from

our earlier survey (v. pp.132-136) that, across disciplines, most texts students faced were in the middling to high functional and referential range categories. There were, however, sufficient entries in the low category to prevent us from making anything other than a general and none too helpful statement, that most students in the three broad discipline areas have to deal with a wide divergence of texts exhibiting a variety of levels in terms of functional and referential ranges.

We decided, therefore, to pitch texts in the pre-test at a level, in terms of functional and referential range, that one would expect an undergraduate in his/her first term at university to be capable of handling.

Taking into account the rather inconclusive evidence available on task dimensions we selected a number of texts and the Project Working Party and groups of language teachers commented on which they thought were the most suitable. In general it was felt that the reading material could not be subject free and it should not be too dense or contain difficult vocabulary unless a glossary could be provided. In the listening components it was felt there should be a variety of accents, speeds and registers. Preliminary trials were then carried out on a group of 30 G.C.E. second year 'A' level native and non-native speakers and 60 first year undergraduate native speakers to ensure that the various texts were appropriate, as regards our task dimensions, the metalanguage of the rubrics was clear and that sufficient time was available for the completion of the tasks set.

4.2.6 A Further Problem in Text Selection

A problem was raised in Chapter 3, Section 1.1 as to whether the content of proficiency tests should be subject-specific and if so, how to make it subject-specific.

Davies (1965, p.242) commented:

"The EPTB made tentative attempts to include alternative reading and listening tests for scientists and non-scientists. Much detailed work needs to be done in specialist areas, medical, legal, commercial and so on."

The situation has improved only slightly since the inception of E.P.T.B. Analyses of the discourse used in the vast variety of courses under review are still not available. Given this current lack of subject-specific analyses in E.A.P./E.S.T., we were forced to compromise.

One approach would be to use tests of general English structures and vocabulary. Research in this area was carried out by Chaplen (1970) who concluded that non-native speakers following courses in British universities need a firm foundation of everyday English if they are to be able to master the variety of English used in their field of study. Accordingly, the two tests of vocabulary and grammar which he constructed were based on 'common-core' English.

The importance of Chaplen's study was that he provided some evidence for using tests of common core English to test students' English language skills, in situations where the testees come from a wide range of disciplines.

We showed in Chapter 3 that there was a good deal of common ground between students in different disciplines and at different academic levels, in terms of the types of activity faced in the various study modes, the attendant performance constraints and the levels of difficulty encountered. This does not remove the possibility, however, that the subject matter of the text they are presented with in a test may affect performance (v. Alderson et al. 1982).

We were concerned to investigate whether Science and Engineering students perform better on science texts than they do on non-scientific texts and whether the reverse is true for Arts, Social, Administrative and Business Studies (A.S.A.B.S.) students. To this end we decided to construct a version of the test with texts deemed suitable for scientists/engineers (Session IIB) and one with texts more appropriate to students in the humanities and social science

disciplines (Session IIA). In addition, we agreed to design a general academic version which was aimed at all students irrespective of discipline (Session I), in which texts are selected from what might be termed 'science for everyman' sources. We would trial the different sessions on students from the two broad groups so that all three Sessions: Sessions I, IIA and IIB would be attempted by groups of Science and Engineering and A.S.A.B.S. students.

In Session I we aimed to construct a version suitable for students in the whole of the target group. We selected texts from a general science for everyman area. Having looked at texts in a variety of topic areas we decided to focus on the area of health education for the purposes of the pre-test. This seemed a topic area of relevance to both scientist and non-scientist. Thus we selected a reading passage on smoking and health and a text for listening comprehension on ways of preventing illness. We also included a dictation task which would include those types of general utterances all students were likely to hear and be expected to take down in various study modes.

For Session IIA which is aimed at A.S.A.B.S. overseas students, written texts were chosen which related to:

- (a) demographic trends amongst the overseas student population,
- (b) problems experienced by overseas students and
- (c) remedial teaching services overseas students might expect to take advantage of.

A spoken text was selected which related to the last of these.

Session IIB was to be aimed at students studying in science and engineering (Sci./Eng.). For this version written texts were chosen which related to:

- (a) practical laboratory instructions,
- (b) description of an engineering process and
- (c) a general scientific theory relating to the origins of life on earth.

A spoken text was selected which related to the last of these.

In Session I it was argued that students should listen to a short lecture in a similar area, health care and the prevention of illness, to that of the reading passage in that Session.

In both versions of Session II the third reading text would be on the same topic as the taped listening component that followed. Thus, in Session IIA students would hear a discussion of the problems faced by overseas students and how tutors at a pre-session course in a British university try to help. In Session IIB students listen to an interview, with two people, about Sir Fred Hoyle's theory of the origins of life on earth. In both cases they would already have read material connected with this topic in the previous section of the test.

In brief then, the texts in the three sessions relate, as far as is feasible, to the fairly broad, heterogeneous audiences they are aimed at.

Session I: Science for Everyman/General academic
Session IIA: A.S.A.B.S.
Session IIB: Scientists and Engineers.

4.2.7 The Problem of Format

An additional problem in constructing T.E.A.P. would be selecting the general types of test format to be employed, e.g. multiple choice, short answer questions, gap-filling, essay, etc., for assessing students' proficiency in the various skill areas. We were aware from earlier research (cf. Murphy 1978a, 1980; Alderson et al. 1982 and Porter 1983) that test format might affect student performances.

Given the limited state of knowledge concerning the effects of various formats, we felt that the only practical approach open to us would be to safeguard against possible format effect by spreading the base of the test more widely. Thus, in experimental trials of T.E.A.P., we include multiple choice, short answer and gap-filling formats for assessing reading ability and in the listening components

we compare the relative effectiveness of dictation, listening recall and short answer questions. In the writing components we include a summary and a more objectively scored editing task. We felt a variety of techniques to be essential for safeguarding against the possibility of one format having an undue effect on student performance (v. Porter 1983). This variety of measures was intended to present us with a more reliable and valid overall assessment of a student's ability to operate through English in an academic context. We now look more closely at the reading, listening and writing components to be included in the T.E.A.P. pre-test.

4.3 ASSESSING READING SKILLS IN AN E.A.P. CONTEXT

4.3.1 General Descriptive Parameters

According to our survey of the reading activities in which students were involved (v. Chapter 3, Section 4.3 and the summary of the data collected repeated in Table 4A below) it is possible to distinguish two different kinds of complementary reading activities to which students are exposed, namely extensive reading and intensive reading.

As part of their preparation for written work and seminar discussions, students often have to search-read to get information specifically required for assignments. This requires the ability to read quickly and with ease, selecting salient features from paragraphs and longer units of prose. In particular, it involves the skills of surveying, i.e. skimming through a text in order to become familiar with the gist of the content and scanning, which refers to the skills used when skimming through a text in order to locate specific pieces of information. It also requires the ability to separate the essential from the non-essential in a text and presupposes understanding of explicitly stated information.

Some of the reading material which students encounter will require more intensive study to understand all the information contained therein. In such cases they will need to examine the text as a unit in closer detail and understand how the various parts are related to each other.

As mentioned above, a guiding principle employed in the selection of test format was that, given the current state of knowledge, it was best to assess a construct by a variety of test formats, the scores on which would be taken as a composite for reporting purposes. The main proviso was that our test tasks should, as far as possible, reflect realistic discourse processing and cover as wide a range of the enabling skills as we had been able to identify (v. Table 4B, p.335).

We were also concerned that the test formats we employed should have a suitable 'wash back' effect on teaching (v. Porter 1983). Thus, although we wished to include a number of indirect methods of establishing proficiency in reading, e.g. gap-filling, we would also ensure that there was at least one task which directly attempted, as far as possible, to simulate the communicative context described in Chapter 3. We discuss below the test formats we thought would be most suitable for inclusion in the pilot versions of T.E.A.P.

4.3.2 Formats for Testing Reading Comprehension

4.3.2.1 Multiple Choice Techniques

A multiple choice test item is one set out in such a way that the candidate is required to select the answer from a number of given options, only one of which is correct. The marking process is totally objective because the marker is not permitted to exercise judgement when marking the candidate's answer; agreement has already been reached, prior to the test, as to the correct answer for each item. Selecting and setting items are, however, subjective processes and the decision about which is the correct answer is a matter of subjective judgement on the part of the item writer. Similarly, when candidates select one of the options given, they are making a subjective judgement; thus, taking the examination is not an objective process either. Only the marking process is objective and as Pilliner (1968) pointed out, it is this objectivity of marking alone which distinguishes objective tests such as multiple choice from all other forms of testing.

TABLE 4A SUMMARY OF FREQUENCY AND DIFFICULTY DATA RELATING TO READING COMPREHENSION ACTIVITIES

A) DIFFICULTY		B) FREQUENCY									
Col. 1 OS	Col. 2 OS-BR	Eng. U N	Eng. P O	Sci. U N	Sci. P O	Sci. A N	S.Sci. U O	S.Sci. P N	S.Sci. A O	S.Sci. U N	S.Sci. P O
(1) 67.9	(7) 8.4										
Reading texts where the subject matter is very complicated											
(2) 55.3	(2) 21.8	XX	X	XX	X	XX					++
Critical reading to establish and evaluate the author's position on a particular topic											
(3) 49.7	(1) 26.2	X	+	+	X	+	+	X			++
Reading quickly to find out how useful it would be to study a particular text more intensively											
(4) 39.6	(3) 18.3	++		++	++	++	++	+	++	++	++
Search reading to get information specifically required for assignments											
(5) 35.0	(6) 8.5										
Reading carefully to understand all the information in a text:											
		++	++	++	++	++	++	+	++	++	++
duplicated notes											
		++	++	++	++	++	++	++	++	+	+
questions done in class or for homework											
		++	X		+	X	+	XX	XX		
laboratory worksheets											
		++	++		+	++	++	+	+	+	+
examination questions											
		+	++	++	++	++	++	+	++	++	++
textbooks, whole or part											
(6) 25.9	(5) 12.7	X	+		+	+	+	+	++	++	++
Making notes from textbooks											
(7) 25.8	(4) 14.3		+		+	++	+	+	++	++	++
Reading to get the main information from a text											

Key as illustrated for Appendix 3.5, p.822

TABLE 4B
ORDERED LIST OF READING COMPREHENSION ENABLING SKILLS
IN AN E.A.P. CONTEXT

-
1. Reference skills, e.g. using bibliography, index, footnotes

 2. Deducing the meaning and use of unfamiliar lexical items through understanding word formation and contextual clues

 3. Understanding relations within the sentence

 4. Understanding relations between parts of text through cohesion devices especially grammatical cohesion, e.g. reference

 5. Understanding relations between parts of text by recognising indicators in discourse especially for introducing, development, transition and conclusion of ideas

 6. Understanding the communicative function of sentences with and without explicit indicators, e.g. definition, example

 7. Understanding conceptual meaning, e.g. comparison, means, cause, result, purpose

 8. Understanding explicitly stated ideas and information

 9. Understanding ideas and information in a text not explicitly stated

 10. Separating essential from the non-essential in text: distinguishing the main idea from supporting detail, e.g. by differentiating the whole from its parts, fact from opinion, statement from example, a proposition from its argument

 11. Transfer of information from one medium to another (*science/engineering*)

 12. Skimming (a) surveying to obtain the gist
(b) scanning for specifics

 13. Notemaking (a) extracting salient points for summary of a specific idea/topic in the text
(b) selective extraction of relevant points from a text for summary especially involving the co-ordination of related information
(c) reducing a text through rejection of redundant or irrelevant information or items

 14. Critical evaluation (*social science*)
-

In many ways, answering multiple choice items is an unreal task, as in real life one is rarely presented with four alternatives from which to make a choice to signal understanding. Normally, when called ^{upon} to do so, we will communicate understanding of what we have read through speech or writing. The distractors present us with choices which we otherwise might not have thought of. If we take a divergent view of the world, we might argue that in any case there is possibly more than one right answer to some questions, especially at the inferential level. What the test constructor has inferred as the correct answer might not be what other readers infer, or necessarily be explicit in the text. Pre-testing items should help resolve a number of these problems though.

If a candidate gets a multiple choice item wrong because of some flaw in the question, the answer sheet on which he records his answer will not reveal this fact; written answers to traditional questions often show whether the answer was wrong because the candidate had wrong information or because he misunderstood the question. However, pre-testing items should reveal flaws in the question if the numbers are sufficiently large. In addition, we do not know whether a candidate's failure is due to lack of comprehension of the text or lack of comprehension of the question. A candidate might get an item right by eliminating wrong answers, a different skill from being able to choose the right answer in the first place. We have no idea from this format why candidates choose certain answers. They might get the right answer for the wrong reasons or the wrong answer for the right reasons.

The scores gained in a multiple choice test, as in true-false tests, may be suspect because the candidate has guessed all or some of the answers. The format of these tests encourages the candidate to guess and it is sometimes considered necessary to take steps to discourage candidates from doing so. It may also be possible to complete some items without reading the texts and, if this is so, whatever it is that we are testing, it is not understanding of the text.

Multiple choice tests take much longer and are more expensive and difficult to prepare than more open-ended examinations, e.g. compositions. A large number of items have to be written carefully by item writers who have been specially trained and these then have to be pre-tested before use in a formal examination. Each item has to be rigorously edited to ensure that: there is no superfluous information in the stem; the spelling, grammar and punctuation are correct; the language is concise and at an appropriate level for candidates; enough information has been given to answer the question; there is only one unequivocally correct answer; the distractors are wrong but plausible and discriminate at the right level; the responses are homogeneous, of equal length and mutually exclusive and the item is appropriate for the test.

In the preparation of the multiple choice items for the first reading passage in Session II we encountered more problems than we had with any other test format. We had to write more than double the number of items that was eventually used. It was thus extremely time-consuming and demanding to get the requisite number of satisfactory items for each passage, especially for testing the higher order enabling skills. A particular problem lay in devising suitable distractors for these items.

There are, however, numerous advantages in employing a multiple choice format. Firstly, since the candidate does not have to write out the answer, a greater number of items can be answered in a reasonable time than would be possible using other kinds of test.

In multiple choice tests there is almost complete marker reliability. Usually the tests are machine marked, which means that a candidate's mark, unlike those in subjective formats, cannot be affected by the personal judgement or idiosyncrasies of the marker. The marking as well as being more reliable, is much more rapid and often more cost effective than other forms of written test.

Because items can be pre-tested fairly easily, it is usually possible to estimate in advance the difficulty level of each item and that of the test as a whole. Pre-testing also provides information about the

extent to which each item contributes positively towards what the test as a whole is measuring. Ambiguities in the wording of items may also be revealed by analysis of the pre-test data and can then be clarified or removed in the test proper.

The format of a multiple choice test item is such that the intentions of the test compiler are clear and unequivocal; the candidates know what is required of them. In open-ended formats ambiguities in the wording of questions may sometimes lead to the candidates submitting answers to questions different from those which the examiner had intended to ask. Though again pre-testing of items would help eradicate this.

Perhaps the most powerful argument for the initial inclusion of multiple choice as a trial format for the measurement of reading comprehension is that other more open formats, e.g. short answer questions, involve the candidate in writing. So, to demonstrate reading comprehension, the candidate has to deploy the skill of writing.

The multiple choice format was, accordingly, employed for assessing reading comprehension in both versions of Session II (v. Appendix 4.1.2 and 4.1.3, pp.870-874 and 905-909) and enabled us to cover a wide range of the enabling skills outlined in Table 4B, page 335 above.

4.3.2.2 Short Answer Questions

These are questions which require the candidate to write down specific answers in spaces provided on the question paper. They possess some of the advantages and disadvantages of multiple choice questions. The main difference is that they do allow the candidate some freedom of expression. Since the candidate's response is expected to be brief, a large number of such questions may be set in this format, thus allowing a wide coverage.

If the number of acceptable answers to a question is limited it is possible to give fairly precise instructions to the examiners who mark them. In those cases where there is more debate over the acceptability of an answer, e.g. in questions on the higher order

inferencing skills, there is a possibility that the variability of answers might lead to marker unreliability. In those cases it might be difficult to arbitrate, for how do we judge whether or not a student has understood a text when he arrives at an answer different from the one we have thought of. It is less of a problem to reconcile than was the case with the multiple choice format but why students give certain answers is still a problem. It raises the question of what it means to understand a text. Maybe a student can satisfy his own criterion, but in a test situation he has to satisfy the tester's definition of this.

Students in tertiary level courses are often required to read texts, extract relevant information and write it down. Accordingly, 'integrated' tests of reading and writing such as short answer questions on a reading passage would appear to possess high face and construct validity. Further the skills we discuss below all seem more testable by short answer questions than multiple choice, the latter involving more spotting and matching of material in the text with the options.

At tertiary level, simple recognition of facts, e.g. through direct reference questions (Davies 1977b) is too simple an ability to test as all it requires of the testees is the recovery of information directly from the text as an almost automatic procedure. There is clearly a need to introduce more complex activities such as inference, recognition of a sequence and comparisons. These require the relating of items in the text with other items which may be some distance away in the text. This is better done through short answer questions where the answer has to be sought rather than being one of those provided. Davies (1977b, p.169) described the difference as follows:

"Whereas the direct reference question takes the sentence as its limit (the sentence being the largest unit of grammatical description) the inference question is directed towards the discovery of the relationship between sentences and the manner in which they combine in communication."

Ryan (1979, p.165) pointed out that:

"Part of the art of reading effectively at an advanced level is to be able to read different kinds of material at appropriate rates and with different amounts of attention."

Given the likelihood that speed develops as a function of general comprehension (v. Davies et al. 1974) it would seem a valid exercise at this level to test whether students taking T.E.A.P. can read some of the texts under reasonable speed conditions and comprehend at the same time. Whilst concentrating on comprehension it seems acceptable to discriminate between the various levels of proficiency of different candidates by asking them to read a passage under a certain amount of time pressure given that, within the time limit set, the vast majority of the native speakers at this level are able to cope effectively with all the tasks set.

Tests under speeded conditions are particularly relevant for assessing the skill of skimming, noted to be one of the most frequent reading activities for all our students. Skimming involves surveying to grasp the essential points of a reading passage and scanning to locate specific pieces of information. Heaton (1975a) noted that, in scanning tests, it is more helpful to set simple open-ended questions rather than multiple choice items; otherwise students will find it necessary to keep in mind four or five options for each item while they are reading.

The difficulties of constructing suitable multiple choice items, especially in finding suitable distractors for those testing the higher level skills, encouraged us to include a short answer format for testing reading comprehension in our experimental pre-test. The short answer format was adopted with reading passages in Session I, Part One, Task 2 and in both versions of Session II, Part One, Task 3 (v. Appendix 4.1, pp.834, 877 and 912).

4.3.2.3 Cloze and Gap-Filling

In a comparison of cloze and multiple choice, Engineer (1977) concluded that the two techniques were measuring different aspects of

the reading activity - namely that a timed cloze measured the process of reading, i.e. the reader's ability to understand the text while he is actually reading it; multiple choice, on the other hand, measures the product of reading, namely the reader's ability to interpret the abstracted information for its meaning value.

There is a good deal of supportive evidence in the literature for using the cloze format. Klein-Braley (1981, p.229) commented:

"Up to now, in the main, the results of research with cloze tests have been extremely encouraging. They have shown high validity, high reliability, objectivity, discrimination and so on."

She quoted Brown (1979, p.13):

"As demonstrated in this and other studies, it can be a valid and reliable test of overall second language proficiency ..."

Alderson (1978a, p.2) described how:

"The last decade, in particular, has seen a growing use of the cloze procedure with non-native speakers of English to measure not only their reading comprehension abilities but also their general linguistic proficiency in English as a Foreign Language."

and added (p.39):

"The general consensus of studies into and with cloze procedure for the last twenty years has been that it is a reliable and valid measure of readability and reading comprehension, for native speakers of English ... As a measure of the comprehension of text, cloze has been shown to correlate well with other types of test on the same text and also with standardised tests of reading comprehension."

He pointed out that though this evidence is not available for non-native speakers (p.63):

"... it does seem cloze procedure is a potentially interesting measure of language proficiency for non-native speakers."

The term 'cloze' was first introduced by Taylor (1953) who took it from the gestalt concept of 'closure' which refers to the tendency of individuals to complete a pattern once they have grasped its overall significance. Taylor (p.416) described it as follows:

"A cloze unit may be defined as: Any single occurrence of a successful attempt to reproduce accurately a part deleted from a 'message' (any language product), by deciding, from the context that remains, what the missing part should be."

The reader comprehends the mutilated sentence as a whole and completes the pattern. Alderson (p.8) pointed out that:

"... the cloze procedure becomes a measure of the similarity between the patterns that the decoder is anticipating and those that the encoder had used."

Taylor first applied the procedure to gauging the readability of a text but it has since come to be highly regarded as a measure of testing reading comprehension and even as a measure of overall language proficiency. For Bormuth (1963, p.134):

"... cloze tests are valid and uniform measures of reading comprehension ability."

and for Heaton (1975a, p.122):

"... cloze tests measure the reader's ability to decode interrupted or mutilated messages by making the most acceptable substitutions from all the contextual clues available."

Enginer (1977) found that a cloze test given under timed conditions provided valid and reliable indices of students' proficiency if two conditions are met: first, that the textual material used is of the appropriate level of difficulty for the population and second, that it contains a sufficient number of deleted items.

Despite the arguments adduced in favour of cloze procedure, a number of doubts have been expressed, largely concerning its validity as a testing device. Rankin (1974), reviewing the use of cloze over the previous twenty years, suggested that cloze was less a measure of comprehension than a measure of redundancy and concluded that cloze is a better measure of readability than of reading comprehension.

Alderson (1978a, p.392) discovered that:

"... cloze procedure is not a unitary procedure, since there is a marked lack of comparability among the tests it may be used to produce. The fact emerges very

clearly from this study that different cloze tests, produced by variations in certain of the variables, give unpredictably different measures, particularly of proficiency in English as a foreign language but, also, probably of other abilities and of readability."

This confirmed Darnell's (1968) earlier finding that traditional random deletion cloze is only remotely associated with reading comprehension and much more closely associated with core proficiency (grammar and vocabulary). Alderson concluded (p.377):

"... it appears that, for non-native speakers of English, the cloze test is not notably an integrative test, nor a test of reading comprehension and high order skills, nor of the ability to handle texts rather than sentences but that it is more a sentence-bound test of low order linguistic skills closely related to core proficiency tests of English as a foreign language."

Rankin (1974) had suggested that a rational, as against a random, deletion was more likely to measure comprehension than linguistic redundancy. Alderson (1978a, p.397) expressed a similar view:

"... perhaps the test constructor should use a rational cloze, selecting items for deletion based upon what is known about language, about difficulty in text and about the way language works in a particular text."

and (p.399):

"One of the major implications of this study, however, is that the emphasis on random selection be downgraded and that the rational deletion of items be given more consideration and be subject to further research."

Klein-Braley (1981, p.244), after a rigorous investigation of the cloze procedure, likewise suggested that a:

"... promising approach to the production of better cloze tests would be to make full use of the information produced by item analysis to delete unsatisfactory items and improve test performance in the usual fashion. A further, probably better, suggestion is that the test constructor should be reintroduced into cloze testing. In such cases the test constructor would be someone who uses linguistic reasoning to decide on deletions. If this were done, it would be easier to state exactly what each cloze test was intended to measure."

It seems that because we are not sure what random cloze is actually measuring we would be better served by rational deletions of passages. We attempt in Session II, Part 1, Task 2, selectively to delete items which would collectively test as wide a range of the enabling skills as possible. Thus, strictly speaking, ours is to be a rational deletion, gap-filling test, as against a 'random deletion' cloze. For the purpose of this experiment we have only excised single words, whereas in future studies it would be interesting to see if we could better test what we wanted to, by removing larger elements from a text.

An interesting finding from experiments conducted with M.A. students in Applied Linguistics at Lancaster and London and with groups of lecturers in Sri Lanka and Poland, was that the selective deletion, gap-filling technique restricts you to sampling a much more limited range of the enabling skills than do the short answer and multiple choice formats.

It transpired, from our limited investigation and the construction of our own tests (v. Appendix 4.1, p.827 et seq.) that whereas short answer and multiple choice questions allow the sampling of skills 1-14 in our list of reading enabling skills (v. Table 4B, p.335 above), gap-filling is much more restrictive. A common finding with the groups who tried to delete items on a passage to test the range of enabling skills was that it was only possible to test skills 2-8. This would seem to accord with Alderson's (1978a, p.99) findings that:

"... cloze is essentially sentence bound ... Clearly the fact that cloze procedure deletes words rather than phrases or clauses must limit its ability to test comprehension of more than the immediate environment, since individual words do not usually carry textual cohesion and discourse coherence (with the obvious exception of cohesive devices like anaphora, lexical repetition and logical connectors)."

and (p.394):

"... cloze is related at least as much to supposedly discrete point tests as to integrative tests and, in particular, that it relates more to traditional tests of core proficiency than to tests like the dictation. It is a better test of ability to deal with syntax and lexis at sentence level than of reading comprehension in

general, the ability to handle metalanguages, or of inferential or deductive abilities; in short, of what have here been termed higher order abilities."

In our selection of texts and scoring procedure we relied on Alderson (1978a) as our major informing source. With regard to the level of difficulty of text to be employed, Alderson (1978a, p.345) noted:

"Easy texts seemed to be a less adequate test of this core proficiency than were more difficult texts, but no evidence was found to support a hypothesis that, in contrast to difficult texts, easy texts permit the measurement of reading comprehension or global skills. It would appear that easy texts also measure low-order skills, but that they do not measure them as well as more difficult texts."

As regards the scoring procedure to be adopted, he concluded (p.395):

"Although no scoring procedure measured any different ability, the semantically-acceptable procedure appeared to be superior to any other, including the exact word method, because it correlated best with criterion measures of proficiency, improved the differentiation achieved by the cloze between native and non-native speakers of English, reduced the effects of the variables of text and deletion rate on the prediction of proficiency in English as a foreign language and also reduced the differences in mean scores of different deletion rates. It resulted in improved score distributions on both medium and difficult texts, improved reliability figures, improved item facility and discrimination statistics, and a reduced incidence of extreme scores."

Klein-Braley (1981, p.243) supported this:

"As Oller reported in 1972, and as Enkvist and Kohonen found in 1978, acceptable scoring is definitely the more reliable procedure for students who have not received explicit training in completing cloze tests."

We tried out each gap-filling test on thirty non-native and eighty native tertiary level students prior to the inclusion of these formats in our pre-test battery as Session II, Part 1, Task 2 (v. Appendix 4.1, pp.875 and 910). Items that caused undue difficulty to native speakers were replaced.

4.3.3 Testing the Range of Skills

As well as employing a variety of test formats, we aimed to cover as many of the enabling skills in each of the reading sub-tests in Session II, Part 1 and in the short answer format used in Session I, Part 1, Task 2 as was feasible. In those cases, in the short answer and multiple choice formats, where this was not possible, we would attempt to include at least one item on each of the higher order skills 7-14. In the gap-filling tasks we would try to achieve a balanced coverage of the lower order skills.

In Appendix 4.3.1, page 948, we indicate opposite each item in the reading sub-tests what the Project Working Party and other experts in the field considered to be the major focus of that item. We were aware that though an item might be seen to be dependant on a particular enabling skill for successful completion, other skills might be contributing to getting the right answer. We realised that the skills we wished to sample were not necessarily discrete.

The enabling skills would also be indirectly involved, though less identifiable, in the more integrated elements Session I, Part 1, Task 1 and Session II, Part 3, Task 1, where candidates have to extract relevant information from text and reconstruct it in an extended writing task.

4.3.4 Variety of Format

We have discussed above the three different formats we are intending to employ for assessing reading comprehension, namely multiple choice, gap-filling and short answer questions. In Chapter 5 below we discuss how these different formats behaved in the pre-test administration of T.E.A.P. and compare their reliability and validity in relation to both internal and external criteria.

We had hoped to investigate format effect further by taking the three reading passages from a version of Session II and constructing a multiple choice, a gap-filling and a short answer test on each of

these. The aim would have been to see if altering the format used to test comprehension of a passage affected the performance of candidates.

Practical reasons, in particular the limited availability of students, meant that we were unable to incorporate this particular experiment into our T.E.A.P. pre-testing programme. A further pragmatic reason against such an exercise was that, as a new version of the test was to be prepared each year we would not be able to trial annually every passage we were contemplating using with a variety of different formats. More particularly, given the fact that we would have to trial the three formats, on the same passage, on the same students, there would be a danger of contamination, and because of the need for varying the order in which the tests were taken, with different groups, the population sampling would be even more complex. This was already a serious problem for our main experimental trials and we were unwilling to spread the limited opportunistic sample at our disposal more thinly.

4.3.5 Conclusions

In integrated formats, where reading and/or listening tasks feed into writing tasks there may be a problem in establishing where the process has broken down. We decided that we would need to assess reading separately as a study mode, as well as combining it with listening/writing activities in order to see if any resultant problems in coping with the integrated task were due to faulty comprehension of the written text.

To allow for possible format effect in this endeavour it seemed sensible to employ a variety of test techniques, which would contribute to an overall profile of reading ability. Both the multiple choice format and short answer questions would seem to allow the testing of the range of enabling skills outlined in Table 4B, page 335 above and also afford economy of marker time. A case is put forward for the use of long texts with these formats on the grounds that these are more representative of academic reading in terms of length and can also provide more reliable data about a candidate's reading ability (v. Engineer 1977).

As regards cloze procedure, we have taken account of the reservations noted by Alderson (1978a) and have constructed a gap-filling task with rational deletions of a fairly difficult text and employing an any-acceptable answer marking scheme. This gap-filling format would enable us to sample the lower order, sentence-bound reading skills.

4.4 ASSESSING LISTENING SKILLS IN AN E.A.P. CONTEXT

4.4.1 General Descriptive Parameters

In this section we consider the problems involved in selecting appropriate testing procedures for assessing students' proficiency in comprehending the different types of spoken English they are likely to encounter in their coursework. Details of the listening comprehension activities and attendant performance constraints relevant to our test population are included again as Table 4C, page 350 below.

Briefly they can be looked at under two headings.

1. Understanding non-interactive discourse in lectures (plus some note taking).
2. Coping with a dialogue situation in coursework in informal lectures, seminars and tutorials involving questions and discussion; what might be termed a reciprocal listening situation, since the various participants have to listen to each other in order to carry on a meaningful interchange.

TABLE 4C
SUMMARY OF THE DIFFICULTY AND FREQUENCY DATA
RELATING TO LISTENING COMPREHENSION ACTIVITIES

	A) DIFFICULTY				B) FREQUENCY (Based on highest returns (N and O) for all classes)									
	Col. 1 OS	Col. 2 OS-BR	Col. 3 Staff OS	Col. 4 Staff OS-BR	Eng. U N	Eng. P O	Sci. U N	Sci. P O	Sci. A N	S.Sci. U O	S.Sci. P N	O		
Understanding teachers and other students when they talk very fast	(1) 55.0% (2) 35.7%													
Understanding when their accents or pronunciation are different from what one is used to	(2) 52.7% (3) 27.6%													
Writing down quickly and clearly all the notes one wants to	(3) 41.5% (10) 8.9%													
Understanding when more than one person is speaking as in group discussion	(4) 41.1% (4) 26.4%													
Understanding informal language	(5) 38.6% (1) 36.9% (3) 52.5% (1) 47.5%													
Thinking of and using suitable abbreviations	(6) 33.9% (7) 21.7%													
Understanding spoken description or narrative	(7) 31.0% (5) 26.3% (2) 53.2% (2) 41.9%													
Recognising individual words in what is being said	(8) 30.8% (5) 26.3%													
Understanding when people speak quietly	(9) 29.3% (13) 2.3%													
Recognising what is important and worth noting	(10) 28.0% (12) 5.8%													
Understanding completely what the speaker is saying and linking this to what he has said earlier	(11) 23.6% (9) 12.9%													
Understanding spoken instructions	(12) 21.1% (8) 19.0% (4) 44.1% (3) 35.3%													
Organising the notes one takes down so that one can understand them when one reads them later	(13) 18.2% (15) -0.8%													
Understanding the subject matter of the talk	(14) 18.1% (14) 0.6% (1) 66.2% (5) 11.4%													
Recognising where sentences end and begin	(15) 9.9% (11) 7.8%													
Making notes	- - - (5) 40.2% (4) 17.6%													

Key as illustrated for Appendix 3.5, p.822 below.

4.4.2 A New Paradigm

The rationale behind the construction of many of the earlier listening comprehension tests was described by Valette (1967, p.49):

"The main object of a listening test is to evaluate the student's comprehension. His degree of comprehension will depend on his ability to discriminate phonemes, to recognize stress and intonation patterns, and to retain what he has heard."

It was thought that, if a learner was tested in phoneme discrimination, stress and intonation, the sum of the 'discrete' sub-tests would be equivalent to his proficiency in listening comprehension.

An example of a test of this type is the E.L.B.A. test battery constructed by Ingram (1964) which placed the emphasis on 'discrete' listening items such as sound recognition, intonation and stress, using short items rather than continuous passages of discourse or dialogue. As Ryan (1979) pointed out, even the section described as listening comprehension seemed more a test of appropriate-response mechanisms than a test of comprehension of continuous speech in an authentic context.

A noticeable trend in recent years has been the attempt to differentiate between tests of auditory discrimination and contextualised tests of listening comprehension. Templeton (1973) outlined how research began to focus on these integrative tests of listening comprehension in preference to discrete point tests of phoneme discrimination, intonation and word and sentence stress.

Since 1969 the J.M.B. no longer tests individual aural skills in isolation, but instead tests listening comprehension in an integrated context of lecturettes or dialogues (v. McEldowney 1976). This paradigm shift can also be observed in the 1977 version of E.P.T.B. (v. Davies 1978) which substituted for the earlier analytical, phoneme discrimination, stress and intonation tasks, an overall listening comprehension sub-test containing, for example, an integrated test of listening comprehension based on a lecture with simulated note-taking.

Davies (1978, pp.146-148) illustrated how similar changes had occurred between the listening tasks described in the first and second editions of Valette's book on testing (cf. Valette 1967, 1977):

"... we can characterise the difference between Valette (1967) and Valette (1977) as a move from linguistics to sociolinguistics, from structuralism to functionalism, from taxonomy and breaking down into skills, into discrete parts, to integration and building up into wholes."

In the second edition of Valette (1977) Davies noted (p.147):

"... a move from a concentration on sound, the production of speech, the phonology, to meaning and communication."

A strong argument against auditory discrimination as a test of proficiency in listening comprehension was that the ability to distinguish between phonemes, however important, did not necessarily imply an ability to understand verbal messages. Furthermore, as Ryan (1979) pointed out, occasional confusion over selected pairs of phonemes does not matter too greatly, because in real life situations the listener has contextual clues to facilitate understanding. For Valette (1977, p.102):

"The key concern of the evaluator is to determine whether the students have received the message that was intended and not on whether they made certain sound discriminations or identified specific structural signals."

Morrison (1974), after assessing the listening comprehension needs of science students at the University of Newcastle-upon-Tyne, concluded that at the E.S.P./E.S.T. level, performance needs to be considered at a level beyond phonology and grammatical structure, thus taking into account the communicative context of spoken discourse. Chaplen (1970, p.19) had earlier concluded:

"Whatever the contribution of the elements of oral/aural communication - intonation, stress and phonemic discrimination - to a test of oral/aural communication, their importance appears to be minimal at any level of proficiency beyond a very elementary stage."

Holes (1972) developed test instruments which focused on the ability to handle academic lectures, a communicative task regarded by departments as both crucial and difficult for their overseas students. He approached test design from a 'job-sampling' viewpoint and attempted to assess the more global, less 'pure' ability of students to interpret 'message content' as well as more narrowly linguistic competence. He used the Davies Test as part of his concurrent validation procedures and an interim academic success/failure rating for predictive validity purposes. Though the predictive validity correlations of tests versus subject examination results were non-conclusive, Holes concluded (p.134):

"The value of the tests lay rather in what they revealed about the kinds of difficulty which overseas students experience in lectures."

In line with the paradigm shift described above we decided to provide ongoing and sequential texts as stimuli for a number of our listening tests, though in terms of the tasks, items and scoring we would wish in certain components of the test to focus on discrete items. As with our tests of reading comprehension, a balance of integrative and 'discrete point' was felt to be the most satisfactory approach for maximising reliability and validity.

4.4.3 Formats for Testing Listening Comprehension

4.4.3.1 Job Sample Tests: Monologue and Interactive Discourse

The investigation of the listening comprehension tasks expected of overseas students at tertiary level, described in Chapter 3, showed that the main activities faced by students were understanding the discourse of lectures (with some note-taking), coping with discussion and understanding verbal instructions. We felt that in an E.A.P./E.S.T. proficiency test at the tertiary level, for validity reasons, it would be necessary to test students' ability to comprehend 'integrated' English, based as far as possible on a job sample (v. Heaton 1975a).

At tertiary level what is required is a test which will indicate whether a student is likely to experience difficulty in understanding lectures whilst simultaneously listening to them and taking selective notes that will later be meaningful to him. It would also be necessary to test ability to comprehend interactive discourse as might occur in seminar/tutorial activities. The commonest technique for assessing integrated skills of this kind is to test comprehension of a short lecturette or discussion.

The main advantage of the lecturette or discussion as stimuli is that they present candidates with the elements of formal discourse; students can be tested on their understanding of a range of features of a lexical, structural, rhetorical and conceptual nature. They have high validity in terms of the General Descriptive Parameters established above - activities, setting, instrumentality, dialect and enabling skills.

In real life the student comes to a lecture or a seminar with a set of expectations and a context provided by the position of that study mode in a course and its relationship to activities in other concurrent study modes. We were aware of the need to provide a context for listening activities and we tried to do this by setting questions on a reading text, thematically related to the subject of the lecture/discussion prior to the listening component of the test. In addition, we would provide a framework of statements from, and questions on, the spoken discourse which would be made available before the passage was heard, in an attempt to make the listening purposive, i.e. task-based. This framework was constructed on the basis of what members of the Project Working Party considered to be credit-worthy points of the listening passage. It was intended in some measure, to redress the lack of exophoric reference and paralinguistic clues inherent in a taped format, which would normally aid the listener in the real life processing of discourse.

In Session I, Part 3, Task 1 and Session II, Part 2 (v. pp. 842, 882 and 915) candidates are given the written framework first; after reading it through carefully for several minutes, they then listen to the tape and are permitted to take notes to help them subsequently to answer a set of short answer questions (v. Heaton 1975a). The

candidates are required to answer the questions in the ten minutes allowed after the tape is finished.

The lecture framework provided in Session I is more comprehensive than that for the seminar discussion in Session II because in lectures, especially in science and engineering disciplines, this type of outline is normally made available to the student along with a great deal of supportive, exophoric reference. The majority of students, however, have to take notes at one time or another with the A.S.A.B.S. groups having to perform this task more frequently than their counterparts in science and engineering. Note-taking is therefore included in the integrated test activity to complete the linguistic skills, though it is not assessed (v. Heaton 1975a). In the more interactive study modes, the discourse is less formally structured and, accordingly, the framework employed in Session II consists only of a set of questions. Whereas a lecture would normally be sequenced, the elements in a seminar would be more disparate. Contextualisation in the latter would be concerned with what the candidate could predict of what was to be discussed. It was felt that candidates should see the open-ended questions before hearing the discussion. In this way they would be alerted to general issues which would ordinarily be predictable in the material to be listened to, so that notes could be made as necessary.

Chaplen (1970) described a similar approach in his aural battery for the July 1969 Test in English (Overseas), consisting of three lecturettes. Each lecturette lasted about five minutes. Students were given two minutes to read five open-ended questions before listening to each passage and were permitted to make any notes they wished while listening to it. They were given ten minutes after each passage finished in which to write their answers. This, it was felt, would place them in the normal lecture situation of listening for specific information. Chaplen (1970, p.175) concluded:

"Although the aural test is recommended for use in identifying non-native speakers enrolled on university courses who would probably benefit from an intensive audio/lingual English course, this recommendation is not made unreservedly."

An alternative not taken up in the T.E.A.P. trials was a multiple choice method for testing listening comprehension. In Paper 4 of the present Cambridge Certificate of Proficiency in English, for example, candidates are required to answer fifteen multiple choice questions testing understanding of three very short passages. This has the advantage of reducing the memory load but its drawback is that it is an artificial process and it is difficult to set up task based listening when four options have to be kept in mind (v. Heaton 1975a). As in the case of multiple choice reading comprehension, preparation of the questions tends to be both difficult and time consuming and there is a serious problem of finding sufficient valid distractors for each item. Given that, for the purposes of validity, we wished to combine listening to a lecture with a simulated note-taking activity (in Session I, Part 3, Task 1) and to assess the ability to understand the general drift of a discussion (in Session II, Part 2) a short answer format rather than a multiple choice format better suited our purposes.

When we came to writing items for the listening element in T031 (v. Appendix 4.1, pp.842-845) we began as in the reading sub-tests, with the intention of testing the range of identified enabling skills (v. Table 4D, p.359 below). The serial nature of extended spoken discourse and the greater processing problems associated with understanding spoken English meant that we could not include items which focussed on the more specific lower order skills. It is extremely difficult for students to backtrack and focus on very specific features of discourse while listening to and attempting to understand non-interactive, uninterrupted monologue. To preserve the integrated nature of the test, therefore, we would have to focus our questions on the higher order processing skills. Thus, with reference to Table 4D below, we would only be able to include questions on the range of enabling skills 6-15.

In the Session II version (v. Appendix 4.1, pp.881 and 916) where we are concerned with interactive discourse and the candidate has to listen to a number of different interlocutors, the questions were to be aimed more at understanding of gist and opinion rather than detailed fact. We felt that students in seminar/tutorial situations

concentrate on establishing a general impression of what participants are saying rather than trying to assimilate everything. We felt that these were more like the type of questions students would ask of themselves in those situations and covered the type of information that students would be expected to retain from such a discussion.

In many ways, the setting up of these listening tests presented us with the greatest difficulties experienced at the test construction stage. We relied on taped broadcasts from the radio for Session I, Part 3 and Session IIB, Part 2 (v. Appendix 4.2, pp.931-944 for full texts). We were looking for broadcasts which would tie in topically with preceding reading passages, in an attempt to provide a context for the listening, as would occur naturally in an academic environment. The idea was that students would first be asked to answer a series of items on the reading passage and subsequently, listen to the tape on a similar thematic area. Lectures in academic courses have their own context by virtue of their occurrence in a sequence and are usually linked to additional background reading in a course of study. We encountered great difficulty in matching up suitable reading texts with taped broadcasts, apart from the numerous inadequacies of the taped broadcasts themselves, e.g. in terms of level of formality, complexity of language, subject specificity, intelligibility of accents, structural organisation and density of information. In the end we managed to find radio broadcasts (v. pp. 931-934 and 940-944) which were deemed adequate for our purpose, though we remained unhappy about the degree to which they conformed with our frames of reference.

For future versions of the test we would recommend strongly that lectures and discussions are commissioned on the particular topic areas we are interested in and that these are recorded while being delivered live to invited audiences. Because of the lack of any suitable pre-recorded material for Part 2 in Session IIA, we conducted a semi-scripted interview with the organisers of the pre-session course at the University of Reading and were able to come far closer to producing the spontaneous exchange of views, agreement and disagreement that was required for the attempt to simulate a seminar-like discussion required for Session IIA of the test battery (v. Appendix 4.2, pp.935-939 for full text).

TABLE 4D ORDERED LIST OF LISTENING COMPREHENSION
ENABLING SKILLS IN AN E.A.P. CONTEXT

-
1. Deducing the meaning and use of unfamiliar lexical items through understanding word formation and contextual clues

 2. Understanding relations within the sentence, the syntactic and morphological forms characteristic of spoken language, especially elements of sentence structure, modification structure and negation

 3. Understanding relations between parts of a text through cohesion devices especially grammatical cohesion devices such as reference

 4. Understanding relations between parts of text by recognising indicators in discourse especially for introducing, transition and conclusion of ideas and for anticipation of objection or contrary view

 5. Understanding the communicative function/value of sentences and utterances with and without explicit indicators, e.g. definition, example

 6. Understanding conceptual meaning, e.g. comparison, degree, cause, result, purpose

 7. Skills concerned with understanding and meaning, especially the ability to recognise the speaker's attitude towards the listener and topic of utterance, as conveyed mainly by intonation

 8. Identifying the main point or important information in a piece of discourse especially through vocal underlining or verbal cues

 9. Distinguishing the main idea from supporting detail, e.g. by differentiating the whole from its parts, fact from opinion, statement from example, a proposition from its argument

 10. Understanding explicitly stated ideas and information

 11. Understanding ideas and information in the text not explicitly stated, e.g. through making inferences

 12. Interpreting text by going outside it, relating information in the text to information not contained in the text, e.g. through picking up exophoric reference

 13. Transcoding information in speech to diagrammatic display, e.g. through completing a diagram/table/chart

 14. Skimming (a) listening to obtain the gist
(b) listening for specifics

 15. Note-taking (a) extracting salient points to summarise the whole text, specific idea or topic, the underlying idea or point
(b) selective extraction of relevant key points from a text especially involving the co-ordination of related information and the tabulation of information for comparison and contrast.
(c) reducing the text through rejecting redundant or irrelevant information or items especially determiners, repetition, compression of examples, use of abbreviations, use of symbols denoting relationships between states, processes, etc.
-

Because of the urgent need to get the trial versions of the test ready in order not to miss the new intake of students at the start of the 1982/3 academic session, we were forced to make do, for Session I, Part 3 and Session IIB, Part 2, with the best of the tapes that had been recorded from the radio.

All these tapes were recorded professionally by a sound engineer at the University of Reading. Although they were deemed equal in quality to recordings used in G.C.E. 'O' level examinations, misgivings were expressed later about their audibility by a small number of overseas candidates during the pre-tests. There is obviously a danger in any large scale aural examination that different schools will not have comparable equipment and, therefore, sound quality and audibility will not be constant across Centres. Even the very room the examinations are held in can markedly affect performance conditions in listening tests. After trying out the tapes and the tests on appropriate groups of native speakers in the summer of 1982 we were satisfied from the feedback obtained that they were intelligible and comprehensible to native speakers at the same level and in the same disciplines as our non-native speaker target group.

We would, however, be extremely cautious in claims as to the degree to which they fully represent the listening study modes students are exposed to in their academic lives. Though they represent a variety of accents, speeds, levels of formality and are more appropriate in terms of length than existing tests, we are still concerned about the degree to which they simulate the real world. The lecture in Session I, Part 3, Task 1 is scripted monologue, 'reading style'. The pre-test does not therefore take account of the more spontaneous 'conversational style' of lecture delivery that was found to be common in our observations, or other forms of lecture discourse, e.g. 'rhetorical style', the lecturer as performer, to which students might be exposed (cf. Morrison 1974; Dudley-Evans et al. 1981 and Skehan 1983). The 'conversational style' which occurs frequently in lectures is much more repetitive, with a lower density of linguistic information than a prepared monologue to be read out aloud. According to Morrison (1974) informal lectures are more difficult

for overseas students to understand than formal lectures.

Another and perhaps more serious problem is that the visual element, the wealth of normal exophoric reference and paralinguistic information, is not available to the candidate and perhaps, therefore, the listening task is made that much more difficult for the student. A normal lecture is not similar to a disembodied voice coming from a tape recorder. Until there is greater accessibility to video equipment in this country the artificiality of a straight audio listening task will remain a problem. Even video is likely to have its own practical difficulties though, e.g. the number of screens required so that all viewers are treated equally or the incompatibility of various systems especially abroad. Whatever the instrumentality, in a test situation the student is in any case denied the natural context provided by the contiguous study modes of an academic course.

As we mentioned above, an attempt was made to compensate partially for these inadequacies by giving the candidates a printed structural framework to aid note-taking, which they read in advance of hearing the tape.

4.4.3.2 Dictation and Listening Recall

Because of the unsuitability of the multiple choice format for assessing performance in the lecture study mode we felt it would be advisable to improve the overall reliability of our listening test by including an alternative, more discrete format with a reasonable number of items. A dictation or a listening recall test seemed likely to provide this discreteness as well as being valid in content terms. Further, being shorter as regards size of text, students would be less likely to suffer from the wandering of attention that occurs in listening to the lengthier lecture-like text. A format of this type would also enable us to cover the lower order enabling skills 1-5, in Table 4D, page 358 above.

In our discussion of communicative testing theory in Chapter 2 we pointed out the importance of the notion that the subjects should be

assessed in situations as close as possible to those in which they would be required to use the language. For dictation, this would involve material which incorporated oral messages typical of those the students might encounter in the target situation.

The needs analysis described in Chapter 3 showed that dictation of short texts (for example, definitions, details of assignments, references) took place, even at post-graduate level. As a major aim in the experimental version of T.E.A.P. was to simulate in certain components the tasks students might perform in their studies, dictation would seem to be a candidate for inclusion. We accordingly constructed a short dictation containing text with language features that were relevant to our target population's likely use of the language. As an alternative it was thought that a listening recall exercise (cf. Furneaux 1982; Henning 1982 and Beretta 1983 for a full description of this procedure), though having less face validity than dictation for our target population, might be a more economical and efficient way of tapping the same competence. We, therefore, decided to carry out a small scale piece of research in co-operation with the University of Reading to ascertain whether this was the case or not and to establish which system of marking both types of test was the most reliable (v. Furneaux 1982).

Of the two types of test involved, dictation was the more familiar and the more researched, though in disfavour until fairly recently, due to uncertainty as to what was being measured, and doubts regarding its accuracy as a testing device. Anderson (1953, p.43) had written:

"Some teachers argue that dictation is a test of auditory comprehension, but surely this is a very indirect and inadequate test of such an important skill."

Lado (1961, p.34) commented on similar lines:

"... on critical inspection it appears to measure very little of language. Since the order of words is given by the examiner as he reads the material, it does not test word order. Since the words are given by the examiner, it does not test vocabulary. It hardly tests aural perception of the examiner's pronunciation, because the words can, in many cases, be identified by context

if the student does not hear the sounds correctly ... Spelling and a few matters of inflection and punctuation can be tested through dictation but the complex apparatus of dictation is not required to test these matters."

Rivers (1968) considered that it could not be used as a valid test of listening comprehension alone and was best used only as a teaching exercise and Harris (1969, p.5) found that, although dictation was undoubtedly a useful device when used in moderation with low-intermediate level learners of a foreign language, it was "generally both uneconomical and imprecise" as a testing technique. Even as late as 1975, Heaton (1975a, p.186) commented:

"... as a testing device it measures too many different language features to be effective in providing a means of assessing any one particular skill."

These criticisms stemmed from a viewpoint, heavily influenced by structural linguistics, that favoured testing the more discrete elements of language skills. The proponents of dictation considered its very 'integrative' nature to be an advantage which reflected more faithfully how people process language in real life contexts. Its critics felt it axiomatic that each test item should be testing a single identifiable element. In so far as dictation tested discrete areas, it was felt that it could be done more reliably and validly by other techniques.

The new interest in dictation reflects the paradigm shift in testing values and objectives referred to above. Whereas in 1967 Valette had cautiously observed that foreign language specialists were not in agreement on the effectiveness of dictation as an examination for more advanced students, significantly, ten years later she was able to state that dictation was a precise measure of overall proficiency and an excellent method of grouping incoming students according to ability levels.

An important factor in the return of dictation to popularity as a testing device was the research carried out by Oller, which formed part of a wider interest in integrative testing (v. Chapter 2). Oller (1971) rejected current criticisms of dictation and argued that it was an adequate test of listening comprehension because it

tested a broad range of integrative skills. He claimed that a dynamic process of analysis by synthesis was involved and that dictation tested not only a student's ability to discriminate phonological units but also his ability to make decisions about word boundaries; in this way the testee discovered sequences of words and phrases that made sense and from these he reconstructed a message (v. Oller 1971).

Oller rejected Lado's (1961) criticisms that in a dictation the order of the words and the words themselves are given, arguing that the candidate listens only to a sequence of sounds which he then processes into words and orders into sequence on the basis of his interpretation of what has been said (v. Morrow 1977). The identification of words from context as well as from perceived sounds is seen by Oller as a positive advantage of dictation in that this ability is crucial in the functioning of language. The success with which the candidate reconstructs the message will depend on the degree to which his internalised 'expectancy grammar' replicates that of the native speaker. Fluent native speakers nearly always score 100% on a well administered dictation while learners make errors of omission, insertion, word order, inversion, etc., indicating that their internalised grammars are, to some extent, inaccurate and incomplete; they do not fully understand what they hear and what they re-encode is correspondingly different from the original. The majority of the group of sixty native speaker undergraduates on whom we tried the pre-test in the summer of 1982 scored 100% in the dictation.

According to Oller (1979), research showed that dictation test results were powerful predictors of language ability as measured by other kinds of language tests (cf. Oller et al. 1975 and Valette 1977). Other research (Oller et al. 1971 and Oller 1972b) produced correlations of 0.8 and better between cloze and dictation. Oller et al. (1975, p.33) explained this correlation:

"... we may reasonably conclude that they are tapping an underlying competence ... The two tests cross-validate each other."

Alderson (1978a, p.190) was more cautious in his interpretation:

"... it is reasonably clear that dictation is closely related to various measures of E.F.L. ability, one of which is the cloze. Why this should be is perhaps less clear."

There is, as yet, no real understanding of what such underlying competence consists of.

Alderson (1978a) concluded that the evidence concerning dictation was inconclusive. He pointed out (p.365):

"The reason it correlates more with some sub-tests than with others does not appear to be due to the claimed fact that it is an integrative test, but because it is essentially a test of low level linguistic skills. Hence the dictation correlates best with those cloze tests, texts and scoring methods which themselves best allow the measurement of these skills ..."

In contrast to dictation, very little evidence is available relating to listening recall tests (cf. Oller 1979; Henning 1982 and Beretta 1983), where the student is given a printed copy of a passage from which certain content words are normally omitted (v. Appendix 4.4, pp.951-954). They have to fill in the blanks, having heard a tape recording of the complete passage twice. They just listen the first time and then attempt to fill in the blanks on the second hearing. This involves many of the linguistic factors outlined above for dictation and this is reflected in the other names which have been given to the test: spot dictation, partial dictation and combined cloze and dictation.

Valette (1977) and Alderson (1978a) commented on the amount of 'dead material' in a dictation scored by the error count method. Valette (p.243) argued that spot dictations have two advantages over conventional dictations in that:

"... they can be administered more rapidly and scored more objectively ... Second, they permit the teacher to test only the problem areas; students do not waste time writing words and phrases that they already handle relatively accurately."

An experiment was carried out with Furneaux (1982) at the University of Reading to determine the relative merits of a listening recall as against a dictation sub-test. The dictation sub-test was administered to groups of overseas students and then was marked according to three schemes:

- (a) a mark was given for each of fifteen segments of the dictation, recorded by the student, which was semantically acceptable and from which the meaning was easily recoverable;
- (b) a mark was given for each of fifteen partial segments, exactly re-encoded as the original. These parts were considered to be the points of the passage central to its message, which a student, on hearing in a lecture or seminar, would note down;
- (c) one mark was deducted for each error.

A fuller description and discussion of these mark schemes can be found in Furneaux (1982). The reliability figures were comparable, but methods (a) and (b) were found to have greater face validity than (c), emphasising as they do the fact that students have to retrieve the vital units of information in the passage, not every word, as they are required to do in a lecture or seminar. These methods were also much quicker than method (c). Marking scheme (a) was seen by Furneaux to be preferable to (b) on validity grounds, both achieving similar reliability figures.

Furneaux (1982) compared student performance on the T.E.A.P. dictation and on the listening recall tests we had constructed (v. Appendix 4.4, pp.951-954). She concluded that the dictation was the more valid of the two formats and that the reliability of these listening recall tests relative to the dictation, did not suggest that the validity considerations be over-ridden. Nor was there any time saving in marking the listening recall in comparison with dictation marking schemes (a) and (b). The two formats were also found to correlate quite highly and this was confirmed later in a similar study by Beretta (1983), who found a correlation of 0.87 between the two formats.

With little to choose between the two formats the greater validity of the dictation in the end encouraged us to include it in our pre-test battery and it forms Part 2 of Session I (v. Appendix 4.1, pp.838-840). The dictation was recorded on tape at normal conversational speed. This would be played to subjects through an extension loud speaker for maximum clarity. Students would hear the dictation once only, 20 seconds pauses between each segment having been found to be ample time for twenty colleagues and sixty native speaker undergraduates to record the information successfully. It was to be scored on a semantically acceptable basis (v. Appendix 5.4.1, p.999).

4.4.4 Conclusions

We felt there was a need to assess listening ability in a more discrete fashion because of the problems foreseen in establishing where the process had broken down in the more integrated parts of the battery, where reading and/or listening feed into writing tasks (v. Appendix 4.1, pp.832, 846, 887 and 921).

In the components of the test battery which focus on listening skills, it is intended that candidates should be exposed to both monologue and interactive discourse of the types they might encounter in lectures, seminars and tutorials in the academic context. They will hear a tape recording of either a lecturette or a short discussion once only. A written framework of the discourse would be provided in the test booklet to help them follow what is being said. The candidates would have to make notes in spaces provided while they are listening to the tape. After the tape is finished they would have time to go through the notes they had made and use them to complete short answer questions.

There would be an additional test of aural comprehension where candidates would hear a taped dictation once only. During pauses in the recording they would have to write down in spaces provided in the answer booklet what the speaker had said.

4.5 ASSESSING WRITING SKILLS IN AN E.A.P. CONTEXT

4.5.1 Why Test Writing Skills?

In the past there have been mixed feelings over including extended writing tasks in a test battery. Many British examining boards, however, use free response or essay type questions in their examinations. In these the candidate has to decide how to set about answering the question and what to include in his answer. The candidates are not usually guided in any way as to how they are expected to answer the question. Candidates approach the questions in different ways and the examiners have to assess the relative merits of these different approaches. Murphy (1979, p.14) stated the case for using this type of technique:

"The principal merit of the essay-type question is its suitability for testing skills, such as the ability to develop an extended argument in a logical way, which cannot be tested in other ways ..."

but added the caution (p.15):

"One major disadvantage of the essay-type answer to the free response question is that it can be difficult to mark in a precise way ... Although reliability of marking is not an overriding consideration in the construction of examinations, it should, as far as is possible, be enhanced. Thus, where essay-type questions have to be used, they should, wherever possible, be combined with other more reliable forms of assessment."

In some examinations writing tests are not included because they are time-consuming both for the examinees and for the markers and they are subject to scorer unreliability. Davies (1965) omitted a test of written production from his Proficiency Test Battery because of the serious practical problems involved. He argued (p.62) that:

"The demands of a reliable test of written expression cannot easily be met within the needs of a second language proficiency test."

Perhaps the most surprising feature concerning tests of extended writing, considering the manifold problems associated with them, is their survival. Coffman (1971, p.271) commented on the continued

prevalence of this technique in the United States:

"In view of ... the extensive criticisms aimed at the traditional methods by experts in the field of measurement and evaluation, one might have expected a dramatic reduction in the uses of essay examinations ... Even in the United States, where the growth of objective testing methods has been vigorous and widespread, essay examinations continue to flourish."

Gipps and Ewen (1974, p.121) discovered a similar situation in Britain:

"... written composition is still widely viewed as one of the most valid means of assessing writing ability despite the inevitable complications which arise in the development of scoring procedures."

The reasons for this continued interest are various. The traditional prestige of the essay and its high status as an examination technique in this country and abroad may partially explain the widespread reluctance to discard it. There is obviously also a strong case for testing extended writing on the grounds of the perceived content validity of job sample tasks. It tests important skills which no other form of assessment can sample adequately. Even Hartog et al. (1936, p.18) concede that:

"... the traditional 'essay' examination should be preserved, because it tests, though at present with considerable uncertainty, skills which cannot be tested by 'new-type' tests, e.g. the power to present a complex series of facts or arguments."

To omit writing tests in situations where writing tasks are an important feature of students' coursework would seriously lower the validity of an E.A.P. testing programme. McEldowney (1976, p.13) summed up the J.M.B.'s rationale for including a test of writing in its test of overseas students.

"Most candidates for the test are hoping to enter scientific or technological courses. In some cases continuous writing may not play a very prominent part in their studies. All will clearly need it to some extent. Some will need it more than others. Though listening and reading skills will probably be of greater importance to most, it seems that a test of continuous writing should be included ..."

The needs analysis described above in Chapter 3 (v. Section 3.4.4, pp.226-268) showed that writing skills were an important element in coursework and examinations for many of our target population and we therefore needed to include a writing component in our pre-test battery.

We examine below different methods of testing and assessing written production in an E.A.P. context to see how we might come closer to constructing valid, practical and reliable measures of this construct.

4.5.2 Formats for Testing Written Production in an E.A.P. Context

In Chapter 3 (v. pp.226-271) we described the manifold writing tasks students are faced with, e.g. writing reports, coursework essays, examination essays and additionally, in the case of post-graduates, dissertations, theses, etc. It would have been impractical, given this wide variety of text types, to get teachers and students to comment on the relative frequency of each, so we had to be satisfied with rather coarse information on the size of text students were expected to produce in coursework and examinations. The summary in Table 4E below focuses on the frequency with which students have to produce written texts of varying lengths, in the various disciplines, in respect of both coursework and examinations.

TABLE 4E

THE FREQUENCY WITH WHICH STUDENTS HAVE TO PRODUCE
WRITTEN TEXTS OF VARYING LENGTHS, BROKEN DOWN
ACCORDING TO DISCIPLINE AND ACADEMIC LEVEL

	Eng. U		Eng. P		Sci. U		Sci. P		Sci. A		S.Sci. U		S.Sci. P	
	N	O	N	O	N	O	N	O	N	O	N	O	N	O
less than a paragraph in	(coursework)													
	(examinations)													
about a paragraph in	(coursework)													
	(examinations)													
more than a paragraph in	(coursework)													
	(examinations)													

Key as illustrated in Appendix 3.5, p.822 below

In addition to establishing the general nature of the writing tasks we also sought to describe the constituent enabling skills that students might need to cope with the writing tasks encountered in an E.A.P. context. These are listed in Table 4F below.

TABLE 4F
ORDERED LIST OF WRITING ENABLING SKILLS
IN AN E.A.P. CONTEXT

-
1. Manipulating the script of the language: handwriting, spelling, punctuation

 2. Expressing relations within the sentence

 3. Expressing relations between parts of a text through cohesion devices, e.g. reference especially grammatical cohesion

 4. Using indicators in discourse, e.g.
 - (a) introducing an idea
 - (b) developing an idea
 - (c) transition to another idea
 - (d) concluding an idea
 - (e) emphasising a point, indicating the main or important information
 - (f) explaining or clarifying a point already made
 - (g) anticipating an objection or contrary view

 5. Expressing the communicative function of sentences
 - (a) using explicit indicators
 - (b) without explicit indicators

 6. Expressing conceptual meaning

 7. Expressing information
 - (a) explicitly
 - (b) implicitly

 8. Planning and organising information in expository language
 - (a) narrative
 - (b) straight description of phenomena and ideas
 - (c) description of process and change of state
 - (d) argument
-

4.5.2.1 Objective Formats

We needed to establish which testing techniques could be used to assess control of English on all levels in the functional areas we had identified. It seemed we could adopt two different approaches. Firstly, we could attempt to measure control of the writing skill by setting extended writing tasks of various types, requiring more subjective assessment. Secondly, we could divide writing into discrete levels, e.g. grammar, vocabulary, spelling and punctuation and test these elements separately by the use of objective tests. Although the emphasis in the present study is upon integrative techniques for assessing writing ability, it seemed desirable to consider both approaches empirically.

Both the productive and receptive skills can be broken down into levels of grammar and lexis according to a discrete point framework. McEldowney (1974, p.8), commenting on the syllabus of the J.M.B. Test in English (Overseas), stated:

"To be able to operate these four skills (listening, reading, speaking and writing) in the various function areas it is necessary to be able to manipulate items from three levels of language. That is, to communicate, it is necessary to have an adequate vocabulary, to know basic items of English grammar and to be able to handle English sounds, stress and intonation."

The J.M.B. Test in English (Overseas), as well as including tasks testing written production, also has tasks which test knowledge of 'basic productive vocabulary' and 'minimum grammatical items'.

The problems which face the constructors of vocabulary tests are manifold, particularly in an E.A.P. context. Chaplen (1970), who constructed the sub-tests for the vocabulary sections of the J.M.B. Test in English in 1969, noted two main problems:

1. The selection of lexical items for testing.
2. Methods used to test the lexical items.

If the testees are studying a variety of different subjects in a university, there is a serious problem of selection. As Ryan (1979, p.191) pointed out:

"The more generalised the subject matter to be tested, the more difficult it is to draw up criteria for the selection of vocabulary items. In specialized courses containing an agreed register required of trainees at different levels, the problem of selection is more straightforward but still exacting."

A further difficulty is the relative weighting that should be given to items selected from students' reading material and the vocabulary items they will be expected to use in report writing and written assignments. Do we test active or passive vocabulary? On top of this how do we judge the frequency levels of the lexical items intended for use in the test?

We felt that a test of knowledge of grammatical structures was a more fruitful area for an experimental comparison between a more indirect, 'discrete point' test and more direct, integrative writing measures. In Session I of the pre-test we decided to include a component consisting of sixty multiple choice items designed to test the candidate's knowledge of grammatical structure (v. Appendix 4.1, p. 849).

A quantitative survey of the occurrence of various structural items in the receptive and productive written material our test population have to cope with in the discipline areas they are studying in was obviously beyond the scope of this investigation. We were, therefore, forced to employ more pragmatic, subjective methods in taking decisions concerning which structural items to include.

It seemed sensible to examine the content of existing tests at an equivalent level to determine what experts in the field had regarded as suitable items for inclusion. We accordingly surveyed those sections of the following tests which assessed knowledge of grammatical structure: E.P.T.B. Form A (v. Davies 1965); Chaplen Test (cf. Chaplen 1970 and James 1980b); E.L.B.A. (cf. Ingram 1964, 1973 and Howatt et al. 1979); Structure-Tests English Language (S.T.E.L.) (v. Best et al. 1976); Middlesex Polytechnic Test in English Language Performance (T.E.L.P.) (cf. Reid 1976 and Riddle 1978); Nelson English Language Tests (v. Fowler et al. 1976); E.P.T.B. Form D (v. Davies et al. 1977); University of Birmingham Assessment and

Diagnostic Test (cf. Johns 1979, 1980, 1981 and Johns et al. 1977b); University of Leeds, English Language Test for Overseas Students (cf. Heaton et al. 1974, 1975 and Heaton 1980); University of Southampton, Pre-Sessional English Language Test (v. Blue 1979).

An analysis of these sources together with reference to appropriate text-books (cf. Quirk et al. 1972; Alexander et al. 1975; Archer et al. 1976; Swan 1976 and Ward 1976) indicated that the areas listed below would be suitable for inclusion in a test of grammatical structure for our intended population. Suggestions were also made by colleagues in the language teaching field on the basis of errors made by target level students in their coursework and from those areas on which it had been found necessary to concentrate, in remedial in-session and pre-session courses.

TABLE 4G
STRUCTURAL AREAS FOR GRAMMAR TEST

1. ± Continuous
2. ± Perfect ± past
3. Probability and obligation
4. Conditional
5. Monotransitivity
6. Ditransitivity
7. Complementation
8. Determiners
9. Number agreement
10. Intensifiers and downgraders
11. Logical connectives
12. Prepositional usage

A feeling was expressed by some colleagues and A.R.E.L.S. teachers that a multiple choice test of grammar might not be the most appropriate format for assessing linguistic competence. So, as well as the multiple choice format included as Part 4 of Session I of the pre-test, we decided to include an editing task in Session II, Part 3, Task 2 as a further measure of competence. In the editing task the student is given a text containing a number of errors of grammar,

spelling and punctuation of the type noted as common by remedial teachers at this level and is asked to re-write the passage making all the necessary corrections (v. Appendix 4.1, pp.890 and 924). As well as being intended as another more objective measurement of competence, it was hoped that this task would also have a good wash-back effect in that students might be taught and encouraged to edit their written work more carefully.

4.5.2.2 Integrative Tasks

With regard to a more integrative approach, it was decided that we should incorporate items which tested a candidate's ability to perform certain of the functional tasks required of him in academic writing. For a scientist this might be, for example, the ability to describe a process or change of state or to summarise an argument (v. list of enabling skills in Table 4F, p.370 above).

A problem arises, however, in terms of the specificity of the text candidates are expected to produce. We were to have considerable problems in selecting examples of various text types, such as description of process or instructions, as we found that the only suitable ideas for texts were so subject specific that there would be too many problems for non-specialists in the subject and the tests would be therefore invalid. The alternative was to choose deliberately obscure texts which in theory favour nobody, to get at underlying abilities, e.g. the ability to classify or to describe. These often involve the testee's imagination and willingness to play a particular game the examiner has in mind, in order to produce various types of writing.

We were also aware of the validity problems if science and engineering students were expected to write essays using a wide range of non-scientific English vocabulary and requiring qualities of literary style and imagination. Though a foreign student may be unable to write a composition on why a cat might make a suitable pet for an old lady, he might more easily be able to describe a process or compare two different objects in his specific area of study.

Wall (1982) carried out an illuminating investigation of the kinds of writing task engineering students were required to perform as part of their coursework and compared these with the types of essay they were set in the Michigan Battery used for assessing students' language proficiency on entry to the university. She (p.166) summarised the differences as follows:

"The main difference seems to be that in the engineering tasks there is much prior input and the task itself is explicitly outlined, whereas in the composition the writer has only a suggestion to respond to and must not only create the content of his writing but a context, audience and purpose as well. The criteria for marking would also seem different."

Disturbingly she concluded:

"Two analyses were carried out on data received for overseas engineering post-graduates. The first was a correlation study between the Michigan Battery total and part scores and the students' first term G.P.A., in which no significant relationship between tests and the criterion for academic success could be found."

We would agree that free, uncontrolled writing is an invalid test of the writing ability required for academic work. Control is necessary in determining the media, the audience, the purpose and the situation (v. Wall 1982). When the task is determined more precisely in this manner, it is also easier to compare performances of different students and to obtain a greater degree of reliability in scoring. If the writing task is uncontrolled, testees may also be able to cover up weaknesses by avoiding problems.

Dunlop (1969) experimented with methods of giving factual information for short written reports by means of tapes and written guidelines. Students were asked to write a short report after reading the guidelines and listening to the factual information presented on the tape. Although the subject matter was not of a scientific nature, the same principle could be applied to the writing of controlled reports in an E.A.P./E.S.T. context. Dunlop (1969) concluded that it was possible to steer the writing of English into more factual channels by using integrated techniques of this nature. Certainly, his testing methods simulated to some extent an authentic

situation in which a student received his factual data for a report or essay from verbal or visual stimuli. Similar approaches are employed by the J.M.B. in their Test in English (Overseas).

In the light of our earlier empirical investigation we feel that a controlled writing task would possess face, content and construct validity if it presents the candidate with a body of spoken and/or written or non-verbal information from which he is asked to extract the main points and then re-combine these in a written form. Positive responses for this integrated, thematic approach were received from the Project Working Party and the members of A.R.E.L.S. consulted. There was a strong feeling that in an academic context there is necessarily some input for any writing task that has to be carried out.

There are various types of stimuli that can be used in controlled writing tasks. Stimuli can be written, spoken or non-verbal, e.g. in the form of a graph, plan or drawing which the student is asked to interpret in writing. The advantage of non-verbal stimuli is that if they present information in a clear and precise way, the candidate does not have to spend a long period of time decoding a written text. The task is most effective when the candidate is asked to comment on particular trends shown in the graph, or to compare and contrast one set of figures with another set. Different stimuli can be used to elicit written performance of a number of different functions such as description of a process, comparison and contrast, writing a set of instructions or argumentation.

Problems have arisen however when, because of the difficulty of constructing knowledge-fair tests of this type, the J.M.B. Test in English (Overseas) has often had to resort to extremely specialised areas such as bookbinding and mediaeval helmets for its visual stimuli. Often candidates are unable to cope with the mental challenge of taking this sort of test and give up rather than jump through the intellectual hoop necessary to get into the writing task. Problems are always likely to occur when the complexity of the stimulus obstructs the desired result, i.e. you need to understand a very complex set of instructions and/or visual stimuli to produce a

description of a process or a classification of a set of objects.

Because of the problems discussed above we only included a limited number of items of this type in our battery, e.g. Session I, Part 1, Task 2, where students have to interpret a graph and in Session IIA, Part 1, Task 3, where students have to extract information from a chart (v. Appendix 4.1, pp.837 and 880).

We viewed summary as potentially the most valid test of a student's writing ability in terms of the tasks he has to cope with in the academic situation. The writing of reports and essays at tertiary level requires the ability to select relevant facts from a mass of data and to re-combine these in an acceptable form. Summary of the main points of a text in this fashion involves not only reading and/or listening comprehension, but also the ability to write a controlled composition containing the essential ideas of a piece of writing and omitting non-essentials.

The main difficulty with this component is marking the product reliably and consistently. To evaluate students' responses reliably one needs to formulate the main points contained in the extract, construct an adequate mark scheme and effectively standardise suitable markers to the scheme. Some subjectivity inevitably remains and it is easy to underestimate the difficulty of marking a summary of this type reliably.

We tried out each of the writing tasks in T.E.A.P. on small samples of native and non-native students prior to the main pre-tests, to try and ensure: that candidates were interpreting the tasks as intended and responding appropriately; that there was sufficient information available in the stimulus material for the candidates to extract; that the kind of writing produced revealed skills that we were trying to measure and that there was sufficient text produced to which our assessment criteria could be applied.

The topics in T.E.A.P., unlike G.C.E. 'O' level English language examinations, are not designed to elicit personal experience or test 'creativity', but rather to reflect the process of information

transfer from written and/or spoken sources into a reconstituted written version that might be produced by a student for his subject tutors. The vast number of the text types this can cover are described in Chapter 3, pp.233-242. In all three Sessions the subject material for the extended writing task is provided for the candidate in a prior reading and/or listening passage.

In this way it is hoped to get closer to the way academic writing tasks are carried out and at the same time guard against the advantage of prior subject knowledge or the disadvantage of having nothing to say on a topic.

Because of the time factor we were forced to select. We could not focus on all the skills described in Table 4F, page 370 above, although it may be possible to vary the functions tested in future versions of the test. We decided that the task which most students would have to cope with across a range of disciplines would be the selective extraction of relevant information from a written corpus of information and verbal input and subsequent reformulation of data from either or both sources in a piece of extended writing. Having made this decision on the grounds of content validity, we next had to face up to the problems of how these written tasks could be assessed reliably.

4.5.3 The Assessment of Written Production: A Problem of Reliability

4.5.3.1 Impression Versus Analytical Approaches

We have discussed how, by controlling the writing tasks in our battery, we might improve the validity and reliability. We concluded that there was a need for 'controlled' writing sub-tests in which the register, context and scope of the writing task were determined for the candidate. This would facilitate marking and allow more reliable comparison across candidates. In this section we examine how the application of impressionistic and analytic approaches to marking might also aid us in our attempt to improve the marking reliability and validity of our writing sub-tests.

Brooks (1980, p.6) defined these approaches as follows:

" 'Analytical' marking refers to a method whereby each separate criterion in the mark scheme is awarded a separate mark and the total mark is arrived at by the addition of these marks. 'General Impression' marking describes a procedure whereby a mark is awarded on the basis of an examiner's overall impression of an essay. Although specific criteria may be borne in mind, these are not assessed separately."

The impression method of marking usually entails two or more markers giving a single mark based on their total impression of the composition as a whole (cf. Wiseman 1949 and Ingram 1970). Each paper is scored using an agreed scale and a testee's score is the average of the combined marks. The notion of impression marking specifically excludes any attempt to separate the discrete features of a composition for scoring purposes. According to Francis (1977), in its purest form, impression marking usually requires each marker to read a sample of scripts, perhaps 10%-25%, to establish a standard in his mind and thereafter to read all scripts quickly and allocate each script to a grade or mark range.

Brooks (1980, p.40), in a comprehensive study, summed up the current state of knowledge concerning these two systems:

"General impression and analytical marking have become established as the two principal methods of evaluating English composition. The continued status and popularity of both of these methods is largely due to the fact that research has been unable to demonstrate conclusively the superiority of either method. The two methods have been found to be of roughly equal merit when judged by the criterion of reliability."

We will examine briefly below the evidence on the relative merits of these differing approaches.

Hartog et al. (1936) were among the first to investigate the relative effectiveness of analytical and general impression marking for assessing English composition. They were intent on finding out which method produced the superior results in terms of ability to reduce marker error. They found (p.123) that variation between markers was, to some extent, reduced by the analytic method:

"... there are greater discrepancies between marks awarded by impression than between marks awarded by details ... it appears that these discrepancies are entirely due to greater differences in the standards of marking of different examiners when they mark by impression."

Their investigation demonstrated that a large number of examiners were consistently biased in terms of either leniency or severity in their marking (v. Ryan 1979). This however, could have been corrected by efficient standardisation of examiners prior to the marking exercise. It is the evidence they produced on discrepancies in rank order placements which is by far the greater threat to reliability, since disagreements of this kind are not susceptible to correction in the same way as differences deriving from bias (v. Ryan op. cit.). In both cases though, a detailed mark scheme might prove useful.

Like Hartog et al. (1936), Cast (1939) found the analytical method slightly superior in a single marker system. His criticisms of the impression method were that, though it discriminated more widely among individual candidates, it judged them on more superficial characteristics than the analytic method. However, although the analytical method was considered the more suitable, Cast conceded that the results did not provide definitive evidence of the superior reliability of analytical marking. Cast was unable to demonstrate conclusively the absolute superiority of either method and, therefore, he refused to advocate the exclusive use of either method.

Cast pointed to some very important characteristics inherent in the two systems. An important feature of the analytical method to which he drew attention (pp.263-264) was:

"... on averaging their marks for all the questions, the range inevitably shrinks ... This 'regression' is the inevitable consequence of all forms of summation of incompletely correlated figures."

In comparison, Cast noted (p.263) that impression marking discriminated more widely among the individual candidates and that the range of marks awarded by different examiners to one and the same script tended to be unusually wide.

Cast (p.264) noted other important facets of each method such as the tendency of impression marking:

"... to seize on a few salient or superficial points - errors of spelling, grammar or fact, perhaps - and weight those out of all proportion to the rest: on the other hand, the analytic methods, by dealing with numerous isolated and possibly inessential points, may overlook certain general qualities that characterize the essays as a whole."

Francis (1977) similarly pointed out that a great danger of impression marking a piece of writing is that impression of the quality as a whole will be influenced by just one or two aspects of the work. He argues that the prejudices and biases of the marker may play a greater part in determining the mark than in the analytical scheme.

Steel et al. (1936) attempted to construct an analytical marking scheme capable of providing a more reliable system of marking. They based their marking procedure on a distinction between two separate aspects of composition writing: subject matter and expression/style. In their view the former cannot be measured objectively and any attempt to incorporate subject matter into mark schemes is likely to lead to poor reliability in essay marking. Consequently, the scheme they proposed paid exclusive attention to details of style and expression. They analysed expression into three components:

- (A) Vocabulary and idioms.
- (B) Sentence structure.
- (C) Sentence linkings.

Their marking procedure worked as follows. Columns A, B and C are ruled in the margin of each essay. For example, for A, opposite each word/phrase which is incorrect, misleads or baffles, a negative sign (-) is recorded. A positive sign (+) is given for good examples of A, B or C. A scheme was devised for converting these scores into percentages.

Morrison et al. (1941) evaluated the reliability of a group of examiners using the Steel-Talman (ST) method of marking with a group of examiners using their own preferred method (impression or analytical or a combination of both). Neither group was standardised. The results were to some extent equivocal. Analysis of

variance showed that the Steel-Talman method was significantly superior to the individual method as regards reducing variation between examiners and reducing random errors on the marking.

Morrison et al. (1941, p.113) pointed out, however, that:

"... the fact that the variability of ST marks is not appreciably smaller casts considerable doubts upon the alleged objectivity of the method."

In addition the method was seen to be slower and more tedious than conventional methods. Brooks (1980, p.17) concluded that:

"... the real success or failure of this method hinges upon its claimed objectivity. It was devised with the specific aim of removing the subjective elements of essay assessment and concentrating exclusively on those which could be objectively measured. Its failure to achieve this provides perhaps the greatest flaw in this method. The apparent tightness and objectivity of the method was belied in practice. There was, in fact, considerable room for divergences in interpreting the method."

Wiseman (1949) investigated the possibilities of improving assessment by summing the multiple marks of four independent, unstandardised markers, using a rapid impression method. He found that multiple marking by impression method improved reliability and was much quicker than comparable analytic procedures. He (p.205) estimated that if the average inter-correlation of a group of four impression markers was as low as 0.6 with each other:

"... the estimate of the probable correlation of averaged marks with 'true' marks is 0.92. This is very much higher than we could expect from one analytic marker."

Wiseman (p.208) argued that:

"The efficiency of markers should be judged primarily by their self consistency."

He pointed out (p.204) that the consistency coefficient obtained by a pure mark, re-mark correlation, using the same marking method on both occasions:

"... is the one single measure which is quite clearly a true consistency, and one which is closest allied to the normal concept of test reliability."

By using a system of multiple marking based on this principle of self consistency, he was able to achieve very high levels of reliability.

Though some doubt has been expressed in the past (v. Edgeworth 1888) about the expediency of having more than one marker, more recently Britton (1963), Britton et al. (1966), Head (1966), Lucas (1971) and Wood et al. (1976) all found that multiple marking improved the reliability of marking English essays.

Britton et al. (1966), in an experiment designed to devise a more reliable marking apparatus for use by examining boards, compared experimental multiple marking with the official single marking carried out by a G.C.E. examining board. They found (p.21):

"The figures clearly indicate that in this case marking by individual examiners with very careful briefing and elaborate arrangements for moderation was in fact significantly less reliable than a multiple mark ..."

When the official marking and multiple marking were correlated with external criteria of coursework produced by candidates throughout the year, multiple marking was found to correspond more closely.

Lucas (1971) found that despite using somewhat inconsistent markers, (mean mark/re-mark correlation only 0.65) multiple marking by impression increased the reliability of the mark awarded significantly. The greatest increase in reliability occurred in the change from one to two markers.

Head (1966) conducted an experiment to discover whether the added impression marks of two examiners would be more reliable than those of individual examiners. He found (p.71):

"The raising of the coefficient from 0.64 for single marks correlations to 0.84 for paired marks correlations shows clearly that the added marks were more reliable."

Wood et al. (1976) using '0' level English Language essay and summary questions found that impression marking by pairs of markers was more reliable than a single marking. Wood et al. suggested though that there is no more to be gained in reliability from a

single analytic marking than from a single impression marking. The real improvement is in double marking.

The work of Wood et al. (1974) and Coffman et al. (1968) highlighted a further problem of the instability of examiner marking behaviour. They produced evidence that marking behaviour does not remain stable during the whole marking period, when a large number of scripts are involved (v. Edgeworth 1888). They argued for subjecting each script to more than one judgement, which might help to neutralize the effects of inconsistent marking behaviour over a protracted period of assessment.

As regards the advantages of impression as against analytic marking though, there is evidence which indicates that multiple impression marking is not necessarily superior to multiple analytic marking. Penfold (1956) compared impression marking with analytic marking and found the latter much more effective in reducing inter-marker variance than the impression scheme. Francis (1977) cited the work of Morrison (1968, 1969 and 1970) who investigated the effectiveness of using different marking methods to improve marker consistency and the overall reliability of the marking of G.C.E. 'O' level English. In the first experiment (1968) Morrison found that using impression marking did not produce more reliable marks than the standard analytical marking procedures employed by the Examining Board at the time. Morrison's findings (1968) were confirmed by a similar study he conducted a year later (Morrison 1969). Morrison (1968) concluded that neither the analytic nor the impression marking schemes, as they stood at that time, constituted the final answer to the marking of English composition, but he did emphasise that (p.18):

"... each is improved by multiple marking."

In the studies we have examined there seems to be an undisputed belief that work marked independently by two different markers, with their marks being averaged, is a more reliable estimate than if it were marked by a single marker. This general viewpoint needs qualifying though, for it is dependent on the markers being equally consistent in their own individual assessments for the duration of

the marking period. If this is not the case the reliability of the more consistent marker on his own might in fact be superior to the combined reliability estimate for two markers who exhibit unequal consistencies.

Pilliner (1969) also investigated the possibility that an increase in reliability obtained through multiple marking might have no meaning in regard to real differences in the qualities of an essay. He concluded (p.315):

"There is some substance to this criticism if each marker is highly self-consistent and if at the same time each agrees poorly with every other. Under these conditions, the mark re-mark reliability of the total (or average) of them all will be both positive and high. But this correspondence will be largely independent of the real differences which presumably exist in the merits of the essays. Instead, it will express the obstinacy with which each marker maintains his own judgments, a reflection of enduring personal idiosyncrasies, an agreement to disagree.

If on the other hand, there is a fair measure of agreement among individual markers about the scripts' merits, the aggregated marks from a team of markers will be a valid expression of the team's consensus of opinion, the reliability of which will increase as the size of the team increases."

These important provisos must be borne in mind in considering the potential value of a double marking system. With an adequate marking scheme and sufficient standardisation of examiners however, a high standard of inter-marker and intra-marker reliability should be feasible and the advantages of a double as against a single marker system would obtain.

Logistical considerations (time, money, computing, personnel) affecting multiple marking have, however, led to a widespread reluctance especially amongst examining boards, to adopt it in large scale marking operations (v. Penfold 1956). A serious problem with multiple marking is that examiners sometimes find it difficult to avoid annotating a script to help them form their impression. If this script is to be re-marked then either the second examiner approaches it in a dissimilar state to the first, the marks have to be tediously removed, or multiple copies of the script need to be

made. In addition, practical difficulties in getting results out in a reasonable period after the conduct of an examination and the cost effectiveness of the procedure have led the Associated Examining Board to employ single markers for all its examinations and this situation is not likely to change initially for those markers employed for T.E.A.P. It is, however, proposed to set up a marking experiment during the first administration of T.E.A.P. in May 1984 to examine the relative merits of single as against double marking. In the meantime it is hoped that by having different examiners mark Sessions I and Sessions II according to a detailed analytical scheme, we will, to some extent, alleviate some of the problems associated with the single marking of essays in each Session. We hope the use of different examiners for the two Sessions will help counteract examiner/candidate bias.

Jacobs et al. (1981) offered a different perspective on the various approaches to composition evaluation. They made a primary distinction between holistic scoring and frequency-count marking as against the rather overlapping division into impression and analytic marking used by the body of researchers referred to above. It was based on a classification by Cooper (1977). Jacobs et al. (p.29) described the division as follows:

" 'Holistic' in Cooper's terms means 'any procedure which stops short of enumerating linguistic, rhetorical, or informational features of a piece of writing'."

In holistic evaluations, markers base their judgements on their impression of the whole composition; in frequency-count marking (v. Steel et al. 1936), markers total or enumerate certain elements in the composition such as: cohesive devices, misspelled words, misplaced commas, or sentence errors. Jacobs et al. argue that the latter method is highly objective and, therefore, also highly reliable. Not so certain is its validity because a composition evaluated by a frequency-count method has been judged not for its communicative effect, but for its number or kinds of elements.

Holistic evaluation would seem to be far more subjective as it still depends on the impressions formed by the markers. Jacobs et al. (p.29) point out though:

"In spite of (or perhaps because of) this subjectivity, holistic evaluations have been shown capable of producing highly reliable assessments. Most of the studies cited ... were, in fact, based on holistic evaluation of one type or another and all of those studies obtained reader reliabilities in the mid-to-high eighties or nineties. Intuitively it would seem that composition scores based on holistic responses from readers who attend to the writer's message must be more valid than those based on frequency-count methods, which at best pay only lip service to the writer's meaning and ideas. As Cooper (1977) puts it, 'holistic evaluation by a human respondent gets us closer to what is essential in communication than frequency-counts do'."

Holistic evaluation appeared to suit our purposes better as we are primarily concerned with evaluating the communicative effectiveness of candidates' writing.

We preferred an analytic, holistic marking scheme to an impressionistic one, favouring an explicit, rather than implicit, list of features or qualities to guide our judgements.

We felt strongly that too little attention had been paid in the past to the actual criteria to be applied, implicitly or explicitly, to samples of written production. Even in the analytic schemes referred to in the studies above, we feel there is too much room for idiosyncratic interpretation of what constitutes the criterion that is being applied to a script. The application of clear, appropriate criteria is felt to be important.

Chaplen (1970) had suggested that more reliable results might be obtained from the impression method of marking if the scale employed was one in which each grade was equated with a distinct level of achievement which was closely described. This was the approach adopted by the British Council in the E.L.T.S. testing system. It may be described as an impression based banding system. We include below an example of such a banded mark scheme taken from Carroll, B.J. (1980).

TABLE 4H
ACADEMIC WRITING SCALE

Band

- 9 Expert writer. Writes with authority, accuracy and style. Has a mastery of appropriate and concise English.
- 8 Very good writer. Clear and logical presentation with accurate language forms and good style. Just the occasional slip or infelicity reveals he is not a native writer. Often approaching bi-lingual competence.
- 7 Good writer. Can develop a thesis systematically with well-structured main and subordinate themes and relevant supporting detail. Generally accurate and appropriate language, layout and style. Responds to tone or purpose of writing task. Mainly distinguished from Band 8 performer in fluency, accuracy and appropriateness.
- 6 Competent writer. Uses a wide range of skills to convey thesis - presenting it in quite a well-structured fashion, arranging main and supporting themes and details logically. Use of lexis and grammatical patterns reasonably accurate. Slight limitation of style and mastery of appropriate idiom in an otherwise intelligible presentation.
- 5 Modest writer. Conveys basic information competently, but logical structure of presentation will lack clarity. Work will show several slips and formal errors. Use of style and conveyance of tone is present but not consistent. Essay may well lack interest but the basic message gets through.
- 4 Marginal writer. Presentation has coherent appearance and several factual statements can be sequentially made. Work lacks logical structure and use of discourse markers. Often makes lexical and grammatical errors. Uses basic punctuation conventions. Uses restricted range of skills. Will backtrack and may still repeat. Basic theme is conveyed but imperfectly.
- 3 Extremely limited writer. Produces a string of sentences rather than an essay. Some theme but not logically presented. Use of simple sentence structure and restricted lexis with errors and inappropriacies abounding. Main merit is the conveyance of straightforward information.
- 2 Intermittent writer. No working facility; perhaps sporadic uses.
- 1/0 Non-writer. Not able to write.

(Carroll, B.J. 1980, p.136)

Carroll's approach is fine in conception as it allows a more detailed description to be presented to institutions. The problem is that, as with Chaplen's (1970) band system, it fails in practice because it does not cater for learners whose performance levels vary in terms of different criteria. A candidate may be a band 7 in terms of 'fluency', but a band 5 in terms of 'accuracy'. This leaves aside other trenchant criticisms we might have, such as the vagueness of such descriptions as 'authoritative writing', 'good style', 'fluency', etc.

This problem of collapsing criteria is avoided by a more 'analytic' mark scheme, whereby a level is recorded in respect of each criterion and to a certain extent the most integrative of our measures is brought back somewhat to a discrete point position. This method has the added advantage in that it would lend itself more readily to full profile reporting and could perform a certain diagnostic role in delineating students' strengths and weaknesses in written production.

Additionally, as far as the Associated Examining Board is concerned, an analytic mark scheme is a far more useful tool for the training and standardisation of new examiners. Francis (1977) pointed out that, by employing an analytic scheme, examining bodies can better train and standardise new markers to the criteria of assessment. A measure of agreement about what each criterion means can be established and subsequently markers can be standardised to what constitutes a different level within each of these criteria. Analytic schemes have been found to be particularly useful with markers who are relatively inexperienced. The data reported by Adams (1981) and Murphy (1982) are consistent with this view.

Analytic mark schemes are devised in an attempt to make the assessment more objective, insofar as they attempt to force examiners to be more explicit about their impressions. Although one of these criteria may take account of the relevance and adequacy of the actual content of the essays, they are normally concerned with describing the qualities which an essay is expected to exhibit. Brooks (1980) pointed out that the qualities assessed by analytical mark schemes in

the past were often extremely elusive. She cites as examples the qualities 'gusto' and 'shapeliness of rhythm' outlined in 'Schools Council Working Paper 49 - Monitoring Grade Standards in English', as being particularly nebulous and inaccessible to assessment. Thus, although analytic schemes may facilitate agreement amongst examiners as to the precise range of qualities that are to be evaluated in an essay, the actual amount of subjectivity involved in the assessment in many schemes may be reduced very little because of lack of explicitness, with regard to the applicable criteria, or through the use of vague criteria.

4.5.3.2 Establishing Appropriate Criteria for Assessing Written Production in T.E.A.P.

The failings of analytic mark schemes in the past have been in the choice and delineation of appropriate criteria for a given situation. We feel that the assessment of samples of written performance should be based on appropriate, behaviourally described, analytic criteria, graded according to different levels of performance. The criteria need to be comprehensive and based on empirical job sample evidence. Our data came from the survey we had carried out amongst A.R.E.L.S. schools (v. Appendix 3.1, pp.666-669) and more particularly the returns to that part of the questionnaire (v. Table 4I below) which had requested from academic staff an estimation of the relative importance of the different criteria (v. Chapter 3, pp.242-258 for discussion of this) they employed in assessing the written work of their students. We gathered empirical evidence from 560 lecturers to help us decide upon those criteria which could be used for assessing the types of written information transfer exercises that occur in an academic context. As a result of our investigations and on the advice of the Working Party, the criteria of relevance and adequacy, compositional organisation, cohesion, referential adequacy, grammatical accuracy and spelling and punctuation were seen as being most suitable for assessing the pre-test writing tasks.

TABLE 4I
THE PERCENTAGE OF ACADEMIC STAFF WHO CONSIDERED EACH
CRITERION IMPORTANT IN THE ASSESSMENT OF WRITTEN
PRODUCTION RANKED IN ORDER OF PREFERENCE

The subject matter	(1)	91.8
Expressing what you want to say clearly	(2)	90.9
Arranging and developing written work	(3)	82.1
Using appropriate vocabulary	(4)	69.6
Tidiness	(5)	62.8
Writing grammatically correct sentences	(6)	46.9
Handwriting	(7)	44.6
Using appropriate grammatical structures	(8)	43.3
Spelling	(9)	42.3
Using a wide and varied range of vocabulary	(10)	41.3
Punctuation	(11)	39.3
Using a variety of grammatical structures	(12)	22.2

From the returns to the staff questionnaire it appeared we needed evaluation procedures that would help us to assess students, particularly in relation to their communicative effectiveness and in such a way that we could present a profile containing a coarse diagnosis of candidates' strengths and weaknesses.

To apply these 'valid' criteria reliably we attempted to construct an analytic marking scheme in which each of the criteria is subdivided into four behavioural levels on a scale 0-3 (v. Table 4J below). A level 3 corresponds to our base line of minimal competence. At this level we feel that a student is likely to have very few problems in coping with the writing tasks demanded of him by his course in respect of this criterion. At a level 2 a limited number of problems arise in relation to the criterion and remedial help would be advisable. A level 1 would indicate that a lot of help is necessary with respect to this particular criterion. A level 0 indicates almost total incompetence in respect of the criterion in question.

TABLE 4J
T.E.A.P. ATTRIBUTE WRITING SCALES

1. Relevance and Adequacy of Content
 0. The answer bears almost no relation to the task set. Totally inadequate answer.
 1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
 2. For the most part answers the task set, though there may be some gaps or redundant information.
 3. Relevant and adequate answer to the task set.
2. Compositional Organisation
 0. No apparent organisation of content.
 1. Very little organisation of content. Underlying structure not sufficiently apparent.
 2. Some organisational skills in evidence, but not adequately controlled.
 3. Overall shape and internal pattern clear. Organisational skills adequately controlled.
3. Cohesion
 0. Cohesion almost totally absent. Writing so fragmentary that comprehension of the intended communication is virtually impossible.
 1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.
 2. For the most part satisfactory cohesion though occasional deficiencies may mean that certain parts of the communication are not always effective.
 3. Satisfactory use of cohesion resulting in effective communication.
4. Adequacy of Vocabulary for Purpose
 0. Vocabulary inadequate even for the most basic parts of the intended communication.
 1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
 2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.
 3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.
5. Grammar
 0. Almost all grammatical patterns inaccurate.
 1. Frequent grammatical inaccuracies.
 2. Some grammatical inaccuracies.
 3. Almost no grammatical inaccuracies.
6. Mechanical Accuracy I (Punctuation)
 0. Ignorance of conventions of punctuation.
 1. Low standard of accuracy in punctuation.
 2. Some inaccuracies in punctuation.
 3. Almost no inaccuracies in punctuation.
7. Mechanical Accuracy II (Spelling)
 0. Almost all spelling inaccurate.
 1. Low standard of accuracy in spelling.
 2. Some inaccuracies in spelling.
 3. Almost no inaccuracies in spelling.

The set of criteria and behavioural descriptions of the levels within each of them (v. Table 4J above) are not seen as irrevocable, but they represent the outcome of a long process of practical trialling and revision. In all, the behavioural descriptions of the levels within the criteria went through five major revisions.

Versions were trialled in experimental marking sessions with the M.A. in Applied Linguistics groups at the Universities of Exeter and London, with thirty Advanced Level E.F.L. teachers from A.R.E.L.S. schools at a staff course run by the A.E.B. in 1982, with four groups of secondary school English language teachers and one group of university lecturers in Sri Lanka in 1982. In addition, professional advice was given by colleagues in the field and by the Working Party attached to the project.

The first problem in earlier versions of these assessment criteria was that in some of the criteria we were trying to assess two things, namely communicative effectiveness and degree of accuracy. As a result we found great difficulty in attempting to apply the criteria reliably. Gradually we were able to refine our scheme so that the first four criteria (1 to 4) related to communicative effectiveness and the latter three to accuracy (v. Table 4J above). It may well be that the latter three criteria contribute to communicative effectiveness or lack of it, but attempts to incorporate some indication of this into these criteria proved impracticable.

Secondly, we were gradually able to establish distinctions between each of the four levels and to achieve roughly equivalent level distinctions across the criteria.

Thirdly, great problems were experienced in the trial assessments in gaining agreement as to what was meant by certain of the descriptions of levels within the criteria. We gradually eliminated most sources of confusion, as far as we were able, and this inevitably resulted in a much simplified scale for these descriptions of level, particularly in the accuracy criteria 5 to 7.

In terms of the criteria for assessing the productive mode of writing it seems that we are able to be far more explicit than is possible in the receptive skills. We have been able to develop criteria of assessment, write behavioural descriptions of four levels within each of these criteria and then apply these to samples of students' writing. Because of the private, internalised nature of the reading and listening processes it does not seem possible to devise such explicit criteria by which candidates' proficiency in the receptive skills can be judged. Whereas we are able to assess candidates' productive ability directly, we can only take indirect measurements of what we label as listening or reading ability. We also have to make the assumption that the sum of the listening or reading skills being measured is equivalent to the whole of what might be described as proficiency in listening or reading.

Though we have attempted to specify what each item is testing in these receptive areas, the candidates' responses can only be judged right or wrong. It does not seem possible to establish levels of attainment on individual receptive skill items, however explicit we can be about what an individual item is testing. Whereas in writing or speaking tests we are presented with something more tangible to make qualitative judgements about, it is more difficult to see at what stage the process might have broken down in items testing the receptive skills or to employ anything other than a dichotomous rating scale. We cannot normally say how near a candidate came to getting an item right in assessment of receptive skills and the more discrete the item the more this must be the case.

In the final version of T.E.A.P. we intend to present a profile of proficiency in the macro-skills of reading, listening and writing. The differences in the way we are able to assess these will affect the manner in which we are able to decide how overall grades are to be awarded in each macro-skill. We return to this problem in Section 5.7 below.

4.5.3.3 Further Considerations in Designing Writing Tasks for Inclusion in a Test Battery

4.5.3.3.1 Number of writing tasks

In our discussion of the writing components in T.E.A.P., we have so far concentrated on how we might achieve marker reliability. There are, however, other factors contributing to the reliability of a test which merit attention. Firstly, the number of samples of a student's work that are taken can help control the variation in performance that might occur from task to task.

Both reliability and validity have been found to be increased by sampling more than one composition from each candidate. Finlayson (1951, p.132) found:

"... the performance of a child in one essay is not representative of his ability to write essays in general ..."

The research of Vernon et al. (1954, p.69) also threw:

"... very grave doubt on the common practice ... of trying to assess English ability in general from a single essay marked by a single examiner ..."

Ebel (1972) showed that the more samples there were of a student's writing in a test, the more reliable the result. Ebel outlined how a test score comprised two elements: the true score and the error measurement. He showed (pp.250-251):

"... the contribution (i.e. variability) of the true component in the total score is proportional to the number of elements (items) composing it ... increasing test length increases the true score variance more rapidly than it increases the error variance."

In other words, reliability of a test score tends to increase as the number of items in the test is increased (v. Willmott et al. 1975).

Murphy (1978b, p.200) also found that an important factor in determining the varying reliability of eight G.C.E. examinations under review was:

"... the number of marks for individual parts contributing to the final examination marks. This effect of increasing reliability, by having more parts of an examination is well demonstrated by the case of English 'A' level ... This observation is consistent with the established principle that combinations of unreliable measurements are more reliable than the individual measurements themselves ..."

Jacobs et al. (1981, p.15) recommended that:

"In general, it is advisable to obtain at least two, if not more, compositions from each student. This helps ensure that the test provides a representative sampling of a writer's ability, by reducing to some extent the effects of variation in an individual's performance from topic to topic or from one test period to another. For large-scale proficiency testing which will be a basis for admission, placement or exemption, maximum writer (or score) reliability is essential; otherwise, the test scores may be of only limited value for making such decisions ... Our experience and that of others (Diederich 1974; Cooper 1977; Harris 1969 and Heaton 1975a) suggest that two carefully formulated writing tasks are probably sufficient for most testing situations."

Obviously the more samples of students' writing that we can take the better, provided each sample gives us a reasonable estimate of the same ability. It was decided that we would include two writing tasks in each Session of T.E.A.P. so that, overall, we would have four measures of the student's proficiency in this area.

4.5.3.3.2 Question choice

As regards selection of topic(s) it is necessary to ensure that students are able to write something on the topic(s) they are presented with. Whether this means allowing a choice of topics is an important decision that has to be made, for it too could affect the reliability of the test.

Jacobs et al. (1981, p.1) advised:

"For large-scale evaluations, it is generally advisable for all students to write on the same topics (Godshalk et al. 1966; Harris 1969; Diederich 1974 and Heaton 1975a) because allowing a choice of topics introduces too much uncontrolled variance into the test - i.e. are *observed* differences in scores due to *real* differences

in writing proficiency or to the different topics? There is no completely reliable basis for comparison of scores on a test unless all of the students have performed the same writing task(s); moreover, reader consistency or reliability in evaluating the test may be reduced if all of the papers read at a single scoring session are not on the same topic (Coffman 1971 and Diederich 1974)."

Heaton (1975a) suggested that offering a choice means, in addition, that some students may waste time trying to select a topic from several given alternatives. Where tests are to be conducted under timed conditions, forcing all students to write on the same topic might also be an advantage for indecisive candidates.

Jacobs et al. (1981, p.17) concluded:

"In view of the problems associated with offering a choice of topics, the best alternative, unless skill in choosing a topic is among the test objectives, would seem to be to require all students to write on the same topic, but to provide them more than one opportunity to write."

By basing the writing tasks in T.E.A.P. on written and/or spoken text supplied to the candidates, we sought to ensure that in terms of subject knowledge all would start equally, at least in terms of the information available to them. All would be required to write on the same topic, but they would write on a variety of topics over the two Sessions.

4.5.3.3.3 Amount of time allowed for each writing task: The ramifications of time limits

Jacobs et al. (1981, p.17) pointed to the need to give due consideration to the purpose of the writing test:

"Is the test a direct outgrowth of certain learning activities, including perhaps, advance preparation for the test composition (reading certain books or conducting research on an assigned topic, practising with a similar topic or the same mode in class and so forth), or is it an impromptu test, which focuses almost entirely on the composing *product*, rather than the composing *process*?"

About the only real life parallel of the closely-timed test is that students may encounter examination essays in their academic courses. If we were to replicate reality more closely the test tasks would not be timed at all and students would be allowed maximum opportunity and access to resources for demonstrating their abilities with regard to this construct. Considerations, such as time constraints, reliability and test security requirements make longer, process-oriented tests impractical for most testing of this kind. In T.E.A.P. a battery of tests rather than a single composition test is involved, so the time available for any one component is necessarily restricted.

Jacobs et al. (1981, p.17) pointed to some of the ramifications of this distinction:

"... a closely-timed impromptu test can hardly begin to tap the writer's resources in the whole composing process, other than to require that all of the process skills be compressed into a speeded time frame, with the result resembling only vaguely what writers usually do in processing written discourse. It is important to remember this serious limitation of a timed, impromptu test."

As regards an appropriate time for completion of product-oriented writing tasks in an actual examination setting, Jacobs et al. (1981, p.18) argued:

"A composition test given in conjunction with a battery of other measures must of necessity be limited in time if the total test time is to be practical and not introduce too much variance due to fatigue in the examinees ... We have used this 30-minute time limit for composition tests given as part of the Michigan Test and believe this time allowance probably provides most students enough time to produce an adequate sample of their writing ability."

They found in their research (p.19) that:

"... with a 30-minute composition test ... students at all but the most basic level of proficiency can generally write about a page or more."

Accordingly, we decided to allow thirty minutes for each extended writing task in the pre-test battery.

4.5.3.3.4 Checking the viability of the writing tasks

Following the advice given in Harris (1969) and Coffman (1971), we tried out the writing tasks on sixty native speaker undergraduates working at the Associated Examining Board in the summer of 1982 and on about forty non-native speakers attending summer courses at Padworth College, Berkshire. Candidates appeared to understand what was being required of them and the tasks seemed to be at an appropriate level of difficulty for first year undergraduate students. The candidates produced sufficient written text, in the time allotted, for us to be able to make judgements in terms of our specified criteria.

Model answers to each question were prepared by the writer and several members of the Working Party to satisfy ourselves that it was feasible to compose a 'good' response to the topics within the time limits allowed.

The nature of the writing tasks, their number, the time allowed and the method of assessment, outlined in detail above, were subsequently implemented in the first pre-test administration of T.E.A.P. (v. Appendix 4.1, pp.832, 846, 886 and 920).

The markers for the pre-test administration of T.E.A.P. were to be selected from examiners with a variety of backgrounds, modern languages, 'O' level English and E.F.L. in order to establish the transmissibility of our assessment criteria. We also had hoped to get a subject tutor from a science or engineering discipline involved, but this was not to prove possible.

The three markers would be trained until they achieved close agreement in their assessment of the same specimen papers. The markers would be monitored throughout the marking period to check that they were applying the criteria consistently.

As a follow up study thirty essays from each Session would be randomly selected from four Centres and the markers involved in assessing the pre-tests would later re-mark all ninety scripts. In

this way, we hoped to establish the degrees of both inter- and intra-marker reliability pertaining in the application of the analytic marking scheme described above.

4.6 CONCLUSIONS

In this Chapter we have described the development of the pre-test versions of T.E.A.P. (v. Appendix 4.1, pp.827-930). This was essentially an experimental battery. The design was eclectic, embracing formats from along the testing spectrum; 'discrete point'; integrative and integrated (communicative). It was our intention to examine the relative merits of different test formats of varying degrees of directness, for assessing candidates' proficiency in performing reading, listening and writing tasks that they might encounter in tertiary level education. By providing two versions of Session II we also hoped to determine whether student performance was affected by the subject content of the text they had to deal with.

As far as possible, in the design of the components of our battery, we attempted to comply with the frames of reference described in Chapter 3, viz. General Descriptive Parameters, Dynamic Communicative Characteristics and Task Dimensions. We have pointed to T.E.A.P.'s deficiencies in respect of these in the foregoing discussion.

In the summer of 1982 we informally tried out the pre-test versions of T.E.A.P. on various groups of non-native speakers and on over sixty undergraduate native speakers employed at the A.E.B. for the summer vacation. Their results and comments in interviews afterwards indicated that the native speaker group encountered few problems in coping with any of the test tasks.

In Chapter 5 below we examine the results of the autumn 1982 main administration of the experimental pre-test versions of T.E.A.P. in terms of their statistical validity and reliability.

C H A P T E R F I V E

T.E.A.P. PRE-TEST: PROCEDURES, REACTIONS AND ANALYSIS

CHAPTER FIVE CONTENTS

	<u>Page</u>
5. T.E.A.P. PRE-TEST: PROCEDURES, REACTIONS AND ANALYSIS	406
5.1 ADMINISTRATIVE PROCEDURES AND REACTIONS	406
5.1.1 Introduction	406
5.1.2 Sampling	410
5.1.2.1 Selection of Sample	410
5.1.2.2 Obtaining the Sample	410
5.1.2.3 Breakdown of Sample	410
5.1.3 Preparation of Materials and Administration	413
5.1.4 Associated Examining Board's Examination Procedures	414
5.1.4.1 Marking Schemes	414
5.1.4.2 The Moderation of Question Papers and Marking Schemes	415
5.1.4.3 The Standardisation of Marking	417
5.1.5 Validity Criteria - Administration	418
5.1.6 Reactions to T.E.A.P.	420
5.1.6.1 General Issues Arising	420
5.1.6.2 Tutor and Candidate Responses to Specific Test Tasks	433
5.1.6.2.1 TO	433
5.1.6.2.2 TA	437
5.1.6.2.3 TB	440
5.1.6.3 Conclusions	443
5.1.7 Handling the Data	445
5.2 DATA ANALYSIS AT THE ITEM LEVEL	446
5.2.1 Introduction	446
5.2.2 Reading	448
5.2.2.1 Conclusions	452
5.2.3 Listening	453
5.2.4 Linguistic Competence	455
5.2.5 Writing	456
5.2.5.1 Introduction	456
5.2.5.2 Summary Tasks: T011, T032, TA31 and TB31	459

	<u>Page</u>
5.2.5.3 Editing Tasks: TA32 and TB32	465
5.2.5.4 Inter- and Intra-Marker Reliability	467
5.2.5.5 Conclusions	473
5.3 DATA ANALYSIS AT THE TASK LEVEL (INTERNAL)	474
5.3.1 Introduction	474
5.3.2 Approaching the Test Data	477
5.3.3 Relationship Between Tasks	481
5.3.3.1 Correlational Data	481
5.3.3.1.1 Reading	484
5.3.3.1.2 Writing	485
5.3.3.1.3 Listening	485
5.3.3.1.4 Linguistic competence	486
5.3.4 Factor Analysis Data	486
5.3.4.1 Conclusions	494
5.4 DATA ANALYSIS AT THE TASK LEVEL (EXTERNAL)	495
5.4.1 Introduction	495
5.4.2 Tutors' Estimates (Global and Communicative) of Students' Language Proficiency	495
5.4.2.1 Background	495
5.4.2.2 Reading	499
5.4.2.3 Listening	499
5.4.2.4 Writing	499
5.4.3 Students' Self-Assessment of Language Proficiency	500
5.4.4 Students' and Tutors' Estimates of Language Proficiency Compared with Aggregated Task Scores	502
5.4.5 Relationships Within and Between Tutors' and Students' Assessments of Language Proficiency	504
5.5 COMPILING THE FINAL FORM OF T.E.A.P.	506
5.5.1 Introduction	506
5.5.2 Decisions on Which Tasks to Include	506
5.5.2.1 Reading	506
5.5.2.2 Listening	507
5.5.2.3 Writing	507

	<u>Page</u>
5.5.3 Data Analysis at the Construct Level	508
5.5.4 Reliability of the Constructs	512
5.5.5 The Concurrent and Predictive Validity of the Constructs	515
5.5.6 Decision on Which Constructs to Include	520
5.5.7 Conclusions on the Validity and Reliability of the Constructs	521
5.5.8 Analysis of Performance in the Two Versions of the Test	522
5.5.9 External Validation of Sessions and Combinations of Sessions	528
5.6 INFLUENCE OF BACKGROUND VARIABLES ON ENGLISH PROFICIENCY	531
5.7 SETTING GRADE BOUNDARIES	538
5.7.1 Introduction	538
5.7.2 Judgements about Groups of Candidates	539
5.7.3 Judgements of Test Content	542
5.7.4 Maintaining Grade Boundaries	543
5.7.5 Future Grade Awarding	544

5. T.E.A.P. PRE-TEST: PROCEDURES, REACTIONS AND ANALYSIS

5.1 ADMINISTRATIVE PROCEDURES AND REACTIONS

5.1.1 Introduction

In Chapter 4 we described our attempts to realise in the pre-test version of T.E.A.P. the specification arising out of the earlier, empirical data collection procedures. In this experimental battery, we incorporated a variety of test formats to establish the best methods for assessing a student's performance level on those tasks and under those constraints that the research indicated to be important to overseas students following academic courses through the medium of English. These formats necessarily varied in their directness of fit with the activities and performance constraints our target population would have to cope with in an academic context. In the integrated tasks we were able to incorporate many of the features of the general descriptive parameters outlined in Chapter 3, but we were much more restricted in our attempts to simulate the dynamic communicative characteristics and task dimensions. The less direct, integrative measures such as cloze and dictation and the more discrete formats such as the multiple choice test of grammar bore proportionately less resemblance to the descriptive parameters, communicative characteristics and task dimensions we had been able to establish.

In the pre-test version of the test battery there were two Sessions. The first Session, Session I, was to be taken by all candidates. Spoken and written texts, which the Project Working Party thought would be accessible to candidates from all disciplines, were selected from the area of general science. Two versions of the second Session, Session II, were also prepared. Session IIA was intended for students in the fields of arts, social, administrative and business studies (A.S.A.B.S.) and the texts in this version were selected from written and spoken sources in these discipline areas. Whilst it was hoped that Session IIA might prove suitable for

students in science and engineering disciplines (Sci./Eng.), we also prepared a second version, Session IIB, containing texts specifically selected from these discipline areas. The test tasks in Sessions IIA and IIB were similar (v. Appendix 4, pp.868-930), but differed in terms of the texts selected.

In Session I and in both versions of Session II candidates were required to demonstrate their proficiency in reading, listening and writing. To counteract the possible effect of test format on student performance a variety of test formats were used in Session I and Session II to test a candidate's proficiency in the range of enabling skills required to operate successfully in the various study modes. In addition to testing specifically the constituent enabling skills underlying abilities in reading and listening, we also included a more integrated task in each of the Sessions in which reading and/or listening activities provide the stimulus material for a writing task.

We list below details of the two Sessions of the pre-test version of T.E.A.P. In the left hand margin we include the task codes for the various components of the test. In the subsequent text, we refer to the various tasks by these codes.

SESSION I

<u>CODE</u>	<u>TO</u>
	<p style="text-align: center;">Part One</p> <p>is a test of candidates' ability to read in English and to write in English about what they have read. They have <u>2 tasks</u> to do in 75 minutes.</p>
T011	Task One - They have to write a <u>summary</u> of parts of a passage. To help them to do this, they should make brief notes while reading the passage.
T012	Task Two - They have to write short answers to a number of questions on the same passage.
	<p style="text-align: center;">Part Two</p> <p>is a test of candidates' ability to understand spoken English. They have <u>1 task</u> to do in approximately 10 minutes.</p>
T021	They hear a short tape recording once only. During pauses in the recording, they have to write down, in the space provided in the answer booklet, what the speaker has said.
	<p style="text-align: center;">Part Three</p> <p>is another test of candidates' ability to understand spoken English. They have to make notes and use them to answer a number of questions. They have <u>2 tasks</u> to do in approximately 50 minutes.</p>
T031	Task One - They hear a tape recording of a short lecture once only. A written outline of the main points of the lecture is printed in the answer booklet to help them to follow what the speaker is saying. This lecture outline consists of three important statements from the passage, each followed by questions. While listening to the lecture they have to make notes in the spaces provided as, after the lecture, they have time to go through these notes and use them to write answers.
T032	Task Two - They have to write a <u>summary</u> of parts of the lecture, using the lecture outline and their notes and answers.
	<p style="text-align: center;">Part Four</p> <p>is a test of candidates' knowledge of English grammar. It consists of 60 multiple choice items. They have 30 minutes to complete this final task of Session I.</p>
T041	

SESSION II

<u>CODE</u>	
TA/TB	(A and B)
	<p style="text-align: center;">Part One</p> <p>is a test of candidates' ability to read in English. There are 3 different reading passages. They have <u>3 tasks</u> to do in 80 minutes.</p>
TA/B 11	Task One - They have to answer multiple choice questions on the first reading passage.
TA/B 12A TA/B 12B	Task Two - They have to A) find words missing from a second passage and B) write these words in boxes provided.
TA/B 13	Task Three - They have to write short answers to a number of questions on a third passage.
	<p style="text-align: center;">Part Two</p> <p>is a test of candidates' ability to understand spoken English by making notes and using them to answer questions. They have <u>1 task</u> to do in 30 minutes.</p>
TA/B 21	They hear a tape recording of a short interview once only. A written outline of the interview is printed in the answer booklet to help them to follow what the speakers are saying. The outline consists of a number of questions. They have to make notes in the spaces provided while they are listening to the interview. After this interview, they have time to go through the notes they have made and use them to write answers.
	<p style="text-align: center;">Part Three</p> <p>is a test of candidates' ability to write in English, in complete sentences, and organise their work so that what they write is clear and answers the questions they are asked. They have <u>2 tasks</u> to do in 65 minutes.</p>
TA/B 31	Task One - They have to write a summary using: (a) notes made on the third reading passage in Part One and (b) relevant information from Part Two.
TA/B 32	Task Two - They have to re-write a short passage which contains a number of errors, making all the necessary corrections.

5.1.2 Sampling

5.1.2.1 Selection of Sample

The conditions stipulated for the non-native speaker (N.N.S.) group in the pre-test were that for each of the individual Sessions and combinations of Sessions, we should try and include representatives from our three academic levels (v. Table 5SS, p.971), from our three broad discipline areas (v. Table 5NN, p.962) and, if possible, from the major language groups (v. Table 5RR, pp.968-970). For the native speaker (N.S.) group, included as a control (e.g. as a check on item difficulty), we attempted to sample across academic levels and disciplines as widely as was possible (v. Table 5NN, p.962).

5.1.2.2 Obtaining the Sample

In Spring 1982 we wrote to a majority of the departments where we knew, from our 1979 enquiries (v. Appendix 1.2, pp.612-632) and personal contacts, that there had been reasonable numbers of overseas students, and asked them if they would be willing to co-operate in the final pre-test stage of the project. We tried to keep a balance in our requests with regard to academic level and discipline area, though practical difficulties such as availability of students, suitability of dates, geographical distances between institutions, etc., affected our ability to achieve this in test administration. From July 1982 to November 1982 we pre-tested all groups who offered their co-operation in response to our requests. The raw numbers of those tested in the various Centres are listed in Appendix 5.1, pages 955-956.

5.1.2.3 Breakdown of Sample

In Appendix 5.2, pages 957-981, we have included background details of the N.N.S. candidates who took either a single Session or a combination of Sessions of the pre-test version of T.E.A.P. A very small number of candidates who took various parts of the pre-test did not fill out a background details questionnaire (v. Appendix 5.2.1, pp.959-960) and, therefore, the n's reported in

connection with these details are slightly less than some of the n's listed in other sections of this chapter.

We feel that these data give a reasonable impression of the salient characteristics of the N.N.S. candidates involved in the pre-test. Data are provided on the students taking the following single Sessions or combinations of Sessions: TO, TA, TB, TO+TA, TO+TB and TA+TB.

The composition of the N.N.S. group specified in Appendix 5.2.2, Tables 5NN to 5BBB are categorised according to: Academic Level and Broad Discipline Area (Table 5NN, p.962), Age (Table 500, p.963), Sex (Table 5PP, p.964), Nationality (Table 5QQ, pp.965-967), First Language (Table 5RR, pp.968-970), Academic Level (Table 5SS, p.971), Specific Discipline Area (Table 5TT, pp.972-973), Broad Discipline Area (Table 5UU, p.974), Length of Time they Have Been in Britain (Table 5VV, p.975), Amount of Time they Spend Outside of Class with English Speaking People (Table 5WW, p.976), Length of Time Studying in English Language Classes (a) in Own Country (Table 5XX, p.977) (b) in Britain (Table 5YY, p.978), Previous Exposure to English as a Medium of Study (Table 5ZZ, p.979) and Previous Experience of Reading in Their Subject Area in English (Table 5BBB, p.980). The following comments should be borne in mind when looking at these tables.

On advice from specialists in the Project Working Party we grouped the N.N.S. candidates into twelve main language families (v. Table 5RR, pp.968-970). Though we appreciate that the languages grouped under Hamitic and other African languages would normally be considered of different families, we consider that there is a greater cultural identity amongst their speakers than is the case amongst other groupings. They encounter similar problems because of their common background of English as a Second Language in the African context.

All twelve language family groups were represented; the Greek, the Semitic and the Germanic perhaps slightly too much. The Slavic, Turkic and Japanese groups are poorly represented, but this reflects

the situation in the population at large (v. British Council 1982). Overall, in terms of language background, the total sample is, it is suggested, a reasonably representative sample of the overseas student population in Britain.

As each Session took approximately three hours to administer, the numbers of students available for completing two Sessions of the test were necessarily limited. It also proved difficult to find overseas undergraduates to take the pre-test as their numbers are smaller and they are spread thinly over courses and institutions. The numbers in the engineering sub-groups were also fairly small and, for the reasons discussed in Section 5.3.2 below, they were to be combined with the science groups at the various levels to form a composite Science-Engineering sub-group (Sci./Eng.).

Accurate random sampling is difficult to achieve in educational research (v. Anastasi 1982 and Houston 1982). The sample population taking our pre-test was necessarily opportunistic as we were largely dependent on the willingness of institutions and students to give freely of their time and efforts. A purposive attempt was made, however, to try and get roughly equivalent numbers of 'A' level and post-graduate, non-native speakers for each Session and combinations of Sessions together with reasonable numbers of their native speaker counterparts. Balancing the numbers of the arts, social, administrative and business studies (A.S.A.B.S.) group with the science and engineering groups was more difficult. Our attempts were often frustrated by the fact that the numbers of students expected were often greater than those who presented themselves on the day to sit the test. This led to unequal numbers taking each Session and combinations of Sessions. Added to this was the fact that receiving institutions had very limited information on the N.N.S. intake as a whole before arrival, and accurate random sampling was for all these reasons, impossible.

Given that we drew our N.N.S. sample from three of the main pre-session courses in Britain: Reading, Lancaster and Southampton, which specifically cater for N.N.S. students with language problems and from the majority of the new intake of N.N.S. 'A' level students at seven colleges and of the post-graduates and undergraduates at

the Universities of Bath, Exeter and Reading, we might reasonably claim that they represent the type of student our test is aimed at. They included students from fifteen institutions (v. Table 5MM, p.956) from ninety different countries (v. Table 5QQ, pp.965-967) and twelve major language groups containing fifty four different languages (v. Table 5RR, pp.968-970). Their level and subject areas covered a large part of the gamut of the courses the N.N.S. population follow in tertiary education in Britain (v. Table 5TT, pp.972-973). Overall, we consider that it is a not heavily biased sample of the overseas student population in Britain in 1982-83.

5.1.3 Preparation of Materials and Administration

The earlier trials in the summer of 1982 (v. Chapter 4) and the moderation procedures described in Section 5.1.4 below, helped shape the form of the T.E.A.P. answer and source booklets (v. Appendix 4.1, pp.827-931). It was considered that the best arrangement was for the questions to which candidates would write their answers to be in an answer booklet and the longer written texts, which act as stimuli, in a separate source booklet. The earlier trials had also helped us to finalise decisions on the amount of time needed for each of the sub-tests and clarify certain of the accompanying rubrics.

A set of instructions for invigilators was prepared for each of the Sessions (v. Appendix 5.3, pp.982-989) to guide those conducting the few Sessions the writer was unable to be present at. The officers of the Board conducting these administrations were asked to follow the instructions carefully and record any irregularities. Explicit instructions were provided as to timing, playing of the tape recorder and the verbal instructions to be given to candidates.

Candidates were given a copy of the test outline (repeated on the inside cover of the test answer booklet) in the week prior to the administration of the test. We requested that where possible, teachers should go through these details so as to familiarise candidates with the nature of the test tasks prior to the day of the

test. We realise, of course, that this only partially offset the problems engendered by a novel and perhaps over-long and complex test system.

The tests were conducted at the Centres listed in Appendix 5.1, page 956. On completion of the test all scripts were taken back to the A.E.B. offices at Aldershot where they were coded.

Three examiners were appointed from different language examining backgrounds to mark the scripts. This was done in order to assess the transmissibility of the marking scheme which we discuss below. One examiner was selected from the Board's Modern Language Panel, one from the English Panel and an examiner from outside who had had substantial previous experience of marking E.F.L. examinations. All three were inducted into the Board's examination procedures described in detail in the next section.

5.1.4 Associated Examining Board's Examination Procedures

5.1.4.1 Marking Schemes

Murphy (1979, p.19) outlined the nature of the marking scheme demanded by one G.C.E. board:

"A marking scheme is a comprehensive document indicating the explicit criteria against which candidates' answers will be judged: it enables the examiner to relate particular marks to answers of specified quality."

The exact form of the marking scheme will depend upon the type of assessment being used. Parts of the marking scheme for T.E.A.P. were straightforward statements of the correct answer (e.g. in the case of objective questions); others were built around a model answer and comprehensive descriptions of levels of performance to aid the examiners in making necessarily subjective judgements about the worth of candidates' answers. Whatever the form of the marking scheme, its purposes are always the same. Murphy (1979, p.14) described these purposes:

- (1) To assist the Chief Examiner and those who will moderate the paper to check the content validity of the questions being set.
- (2) To help the moderators to check that the demands made in the examination are appropriate and in accordance with stated aims and objectives.
- (3) To allow the moderators to check that the answers expected from the candidates are in accordance with the questions as set.
- (4) To ensure that, where there is more than one examiner, each examiner marks in exactly the same way, awarding equal marks for equal levels of performance.
- (5) To ensure that each examiner marks consistently throughout the marking period.

We have included as Appendix 5.4, pages 990-1038, copies of the mark schemes provided to examiners for the assessment of T.E.A.P. Sessions TO, TA and TB.

5.1.4.2 The Moderation of Question Papers and Marking Schemes

Each of the T.E.A.P. question papers and marking schemes were subjected to the standard process of moderation employed by the Board for all its examinations. A sub-committee was appointed to carry out this moderation, comprising a small group of five language testing specialists. The question papers and marking schemes were moderated together. In this way, we sought to establish that the questions could be answered on the basis of the information supplied and that the expected answers were correct. Using Murphy as our informing source we drew up a set of questions, which were applied to the question papers at the T.E.A.P. moderation meeting.

- (a) Has the paper been set at an appropriate level of difficulty?
(i.e. Are the questions too easy or too difficult for a language proficiency examination at this level?)
- (b) Will the paper discriminate adequately between the performance of candidates of different levels of attainment?
- (c) Does the paper (when considered along with the other papers) test the full range of appropriate skills and abilities, as defined by the objectives of the examination?

- (d) Are the questions unambiguous, giving a clear indication of what the examiner is asking, so that no candidate may take the question to mean something quite different?
- (e) Is there an excessive overlap in enabling skills or communicative tasks being assessed in each Session?
- (f) Can the tasks be satisfactorily answered in the time allowed?

Another duty of the T.E.A.P. Moderating Committee was to consider the format and layout of question papers. This was important because a badly laid out question paper could be the cause of considerable problems for both candidates and examiners.

After a number of earlier try-outs we gradually got nearer to making the instructions as clear and as concise as possible. Bold type and capitals were used, where necessary, to stress particular parts of the rubric.

Because of the difficulties in reading facing many of the candidates unused to the Roman script, great care was exercised in selecting the type-face for the pre-test version of T.E.A.P. After consultation with experts in the field and trials on students at Padworth College, it was decided to employ 'letter gothic' as this proved the clearest, most easily read type-face for the overseas students consulted.

At the same time as the question paper was being moderated, the Moderating Committee considered the appropriateness of the marking scheme.

Again we used Murphy (1979) as our major informing source in the construction of the following set of questions to which the moderators addressed themselves:

- (a) Does it anticipate responses of a kind that candidates are likely to make?
- (b) Are the marks allocated for each part of a question commensurate with the demands made on the candidates?
- (c) Does the marking scheme indicate clearly the marks to be allocated for different parts of a question?

- (d) Does the marking scheme allow for possible alternative answers?
- (e) Has the marking scheme reduced to the minimum possible, the amount of computational work which the examiner has to undertake in order to arrive at the correct mark for a script?
- (f) Does the marking scheme, by specifying performance criteria, reduce as far as possible, the element of subjective judgement that the examiner has to exercise in evaluating candidates' answers?
- (g) Are the abilities being rewarded those which the questions are designed to assess?
- (h) Are the solutions to the problems correct?
- (i) Can the marking scheme be easily interpreted by a number of different examiners in a way which will ensure that all mark to the same standard?

5.1.4.3 The Standardisation of Marking

Even if examiners are provided with an ideal marking scheme, there might always be some who do not mark in exactly the way required. The purpose of the standardisation procedures was to bring the three T.E.A.P. examiners into line, so that candidates' marks were affected as little as possible by the particular examiner who happened to mark their scripts.

The following standardisation procedures were applied. Firstly, the three examiners, one for each Session, attended a standardisation meeting prior to receiving their allocation of scripts. At this meeting, we explained to the examiners how the marking criteria were to be applied. This was done with reference to the marking scheme. If any of the examiners had queries regarding the marking scheme they had an opportunity to raise these problems at this meeting. At the standardisation meeting the examiners were required to mark a set of photocopied scripts and, when this had been done, their marks were compared, to ensure that they were all applying the same marking standards. This provided an opportunity for examiners to do some marking in a situation where they could compare their marking standards with each other and with us and resolve any problems occurring.

After the standardisation meeting, the examiners began their official marking of scripts. During the marking period they had to send to us at least two sample batches of scripts. This enabled us to check that the scripts were being marked according to the marking scheme. Once we had re-marked the first sample batches of scripts, these scripts were returned to the examiners who had first marked them. Any inconsistencies in the examiners' marking at this stage were pointed out. The same procedure was employed for the second batch of scripts.

5.1.5 Validity Criteria - Administration

In Chapter 2 we discussed the importance of validating T.E.A.P. against external criteria, e.g. tutors' estimates of students' language proficiency, students' self-assessments, students' academic progress on their courses and preferably, if possible, against established English language tests. Though the major focus of our enquiry was the a priori validation of T.E.A.P. we recognised the importance of such external validation. We accordingly enlisted the aid of these institutions where pre-tests were being conducted, in establishing these external criteria.

At the time we administered T.E.A.P., the students' English language tutors were asked to provide us with two estimates of their students' proficiency in English. The first of these (v. Appendix 5.5.2, pp.1050-1051) was an analytical communicative assessment (T.C.) in which tutors were asked to give an estimate on a four point scale of proficiency in specific activities research had shown to be important for students operating in an academic context. Given that T.E.A.P. was in part felt to be a 'communicative' assessment, it was considered important that this should be mirrored in the external assessments and the tutors' questionnaire should therefore be an estimate of the same thing as the test itself.

The second estimate tutors were asked to provide was an overall, global rating (T.G.) on a four point scale for each of the four macro-skill areas (v. Appendix 5.5.3, p.1053). The latter was a more traditional impressionistic assessment of the type teachers

were accustomed to making. We were interested in comparing the two types (T.C. and T.G.) of rating to see which might be the most useful for future validation exercises.

In addition, students were asked to provide a communicative self-assessment of their own proficiency according to the same scale and criteria as their tutors' communicative assessments (v. Appendix 5.5.1, pp.1047-1048). These self-assessments were filled in prior to the taking of the test.

At the end of the second term of the academic session, April 1983, we approached the institutions which had co-operated in the 1982 administration of T.E.A.P. and asked them if subject tutors could fill out a pro forma for estimating the academic progress of those students who had sat T.E.A.P. (v. Appendix 5.5.4, p.1055). Tutors were asked to assess how far students had been able to cope with the language demands made on them in their courses of study in terms of the broad macro-skills, reading, listening, writing and speaking. We felt tutors might be unable or more likely unwilling to fill in a more explicit set of questions on specific activities. We also asked about the likelihood of a student failing his or her course and how far failure might be attributed to deficiencies in English.

We are aware of the problems involved in using a criterion which comes a long time after application of the test (v. Davies 1965). During the interval between the administration of T.E.A.P. in October-November and the collection of subject tutor estimates in April, some language learning had probably taken place. In addition, as we saw in Chapter 1, there are a large number of variables involved in academic success which will tend to obscure any posited relationship between language skill and academic success.

Davies (1965, p.152) pointed out that:

"... the English of a native English student is hardly a prognostic of his academic performance, and the more an overseas student approaches the native English in English proficiency the less useful is a test of his English proficiency as a prognostic."

Finally we asked institutions to let us have the results of any internal language tests administered to students who had sat T.E.A.P. together with any results of established English language tests such as E.L.B.A., E.L.T.S., J.M.B., F.C.E. and C.P.E. that students might have taken.

In conclusion, we would emphasise that these validation procedures are not without their practical problems (v. Davies 1965). The criteria in the various assessments may be applied in different ways from individual to individual; the teachers and subject tutors may not be in a position to make reliable estimates of a student's proficiency in certain macro-skills and lastly it may prove difficult to obtain any data at all from certain institutions.

5.1.6 Reactions to T.E.A.P.

5.1.6.1 General Issues Arising

After each test administration candidates were asked to fill in a follow up questionnaire relating to the Session they had just taken (v. Appendix 5.6.1, pp.1057-1065). Given that the students had already spent about three hours sitting the test, we can only express our gratitude for their extra effort in supplying us with this information. The results of these data collection exercises will be discussed in this section.

We first propose to examine general issues which arise out of the data before looking in detail at student and invigilator responses to specific tasks within the Sessions TO, TA and TB. In those cases where we quote directly from student questionnaires below they are to be considered as non-native speaker (N.N.S.) reactions unless otherwise stated.

The number of follow up questionnaires returned were as follows:

	<u>Student</u>		<u>Invigilators</u>
	N.S.	N.N.S.	
TO	115	294	15
TA	75	420	10
TB	123	308	11

With few exceptions the students, after completing a Session, filled in a questionnaire. We would emphasise the need for caution concerning the quality of these data though, as given the arduous nature and length of the test, it is likely that some candidates may not have appreciated this extra imposition on their time. Nevertheless we feel that together with the comments of the invigilators and our own observations, they provide us with useful feedback on the pre-test, which can be taken account of in the design of the final version of T.E.A.P.

The extent to which students are motivated when taking an experimental trial version of a test is difficult to judge. Having personally administered the majority of the test Sessions, it seemed that the N.N.S. group had approached the test positively, regarding it as a potentially useful exercise. Though they were aware that it was not an examination with a qualification dependent upon it, for the most part they appreciated that the results were intended to be used to help them rather than against them. On the whole, participating institutions were extremely helpful in getting across to the N.N.S. group the purpose of the test. Against this it must be said that the pre-test in its size and duration was a daunting prospect especially for the less proficient. The very length of the test, some three hours, may have affected the motivation and performance of some candidates adversely.

The motivation of the native speaker (N.S.) group taking the test was more questionable. It proved to be easy for a majority of this group. Many attempted to get through the test as quickly as possible and most finished well before the allotted time expired. When a Session took place in the afternoon there was an obvious drop off in interest as the afternoon wore on and people wanted, or had, to get away.

In Question 8 of the questionnaire (v. Appendix 5.6, p.1059) we requested further comment on the test. The students' comments listed below are not to be taken as wholly representative of the views of the population, but rather serve to illustrate some of the general points which emerged concerning the tests. We examine below the comments that were made concerning: the length of the test; the attitude to the subject content of the test; the familiarity of students with the type of test and reaction to the general idea of the test. We have grouped the comments according to the Sessions they relate to:

Length:

TO

"It is necessary to give more time because in a normal situation of learning we are not under such pressure of time."

"Session I and II should have both been held on consecutive mornings; one gets extremely tired after a whole morning of examination, and when followed by more, concentration goes completely down the drain - hence bad performance. P.S. Writing gets bad because hands are tired."

"It is a good idea to have this exam. Perhaps it is too long."

"I think that for a test it is enough one day 4 hours like TOEFL or ELTS. Both Session I and Session II - that is too much."

TA

"For a better result out of this experimental test I suggest that the questions be cut and made more concise. By doing this people wouldn't be very tired at the end. Generally though, it wasn't a bad experiment."

"Test too long. 6 hours is too long in one day. Perhaps 3 hour sessions on separate days would be better and would probably produce better results too."

"Concentration and performance may fall off with a test this long."

"It should not be taken in one sitting. The efficiency of a student decreases as a lot of time in the examination room passes."

"It is very hard to concentrate for such a long time. A test in two parts with a break would be more pleasant."

"Overall, too long and potentially boring. Could reduce time allocated for each task and could combine tasks to shorten test. Personally, I think the actual idea behind the test is very valid."

TB

"A bit too long in duration. Too many different sections, I'm sure it could be condensed."

A member of staff made an important point on the issue:

"The whole test (Sessions TA and TB) seemed to put students under too great a pressure for too long. It might be more acceptable to have certain, shortish sections where the time element is crucial and other sections where students have ample time to complete the tasks. But 6 hours of constant pressure seemed a bit punishing."

The length of each Session of the test was a problem we were aware of but, given the desire to trial a variety of formats, we felt it was unavoidable in the pre-test version of T.E.A.P. It was hoped that the analysis of pre-test results would enable us to cut down on the number of tasks within each Session as well as reduce the size of some of the tasks themselves in the final battery.

Attitude to Subject Content:

We were also interested in establishing attitudes to the subject content of the test. In Question 3 of the follow-up questionnaire we sought to ascertain what the reaction to the subject content of the Sessions had been. The results are recorded below.

TO

Most of the staff thought this Session would be equally difficult for all types of students. A few thought it would favour those in the humanities or social sciences and one tutor thought it would favour those studying science, engineering or mathematics. Several invigilators commented that having to process a dense text in a new subject was not the same as taking a test in their own subject.

The replies of the students to this question are listed below.

With which of these statements would you agree?	N.S.	N.N.S.
A. The test would be easier for those studying in the Humanities or Social Sciences.	43	67
B. The test would be easier for those studying Science, Engineering or Mathematics.	3	43
C. The test would be equally difficult for the groups of students referred to in both A and B.	66	162

TA

Of the limited number of staff (10) completing the follow-up questionnaire, all thought it would be equally difficult for both groups of students with the exception of one who thought it favoured those studying in the humanities or social sciences.

The student replies are listed below.

With which of these statements would you agree?	N.S.	N.N.S.
A. The test would be easier for those studying in the Humanities or Social Sciences.	24	118
B. The test would be easier for those studying Science, Engineering or Mathematics.	1	25
C. The test would be equally difficult for the groups of students referred to in both A and B.	43	207

TB

All the staff consulted, thought the test would be easier for those studying science, engineering or mathematics.

The student replies are listed below.

With which of these statements would you agree?	N.S.	N.N.S.
A. The test would be easier for those studying in the Humanities or Social Sciences.	33	40
B. The test would be easier for those studying Science, Engineering or Mathematics.	36	122
C. The test would be equally difficult for the groups of students referred to in both A and B.	45	121

We also collated those replies to Question 8 which had a bearing on students' reaction to the subject matter of the tests and a selection of these can be found below. Very few comments were made on the subject content of T0 and TA.

T0

"Use more general topics for the passages."

"The subjects chosen could have been more interesting."

"The subject matter was a bit too interesting - I was tempted to stop and read it all rather than extract information."

TA

"The subject is so boring that you are not interested in it at all. I would have enjoyed it more if it had been a good story or a tale."

"TA was better than T0 because subject material more general."

"Need for variety in topics."

"I found the texts and the topics in general rather dull, one text concerning these problems would have been sufficient."

"I suggest not to use the same subject all over."

TB

The majority of replies to Question 8 following the TB Session were concerned with the subject specificity of the texts. They presented, in many cases, too great a challenge for the non-science/engineering students as the following comments illustrate.

"Even while choosing an area from science, the pattern should follow a more general trend than, for example, what we have here. If the subject is totally alien, comprehension becomes more difficult because the full import of the technical words is not understood. In my opinion, it would be best to choose a neutral area, such as current events which will not be biased either in favour of or against the discipline of any candidate."

"The English was okay to understand but then it had too much of a one subject like say to do with science, which put me off a bit but anyway I have just tried to answer the few I know. If it had been based on my areas of interest I would have enjoyed the exams."

"Test shows a definite science bias. Not entirely suitable for candidates involved in the Arts."

"The subject matter of the passages was completely unknown to me. I could do better if the passages were on general subjects."

"I don't know anything about engines or outer-space."

"The section on the diesel engine probably favours men."

"When the topics are not related to the subject known to the students it creates inferiority. Therefore the subject of the topics should be either more general or related to the subject known to the student."

"Contains very uninteresting reading passages and it is too scientific. Contains scientific words and calculations one is not used to."

"Shorter more relevant (to our subjects) tests would have been better. Would help if there was more variety in the subject matter."

"The subject under discussion in various parts of the test should be relevant to the people taking the test, and more time allowed on certain texts."

"Choose more carefully the texts. The texts used in this test were technical and it is not fair to evaluate with them people from other disciplines. I am a social anthropologist myself and I am absolutely sure that if other text had been used, my performance would have been much better!! It is necessary to choose more general subjects."

"I am not familiar at all with engineering science terminology. I have never studied in this country physics, chemistry or astronomy. Although I can understand perfectly well any newspaper and all books, magazines and other periodicals, in the area of Chartered Surveying, I experienced some difficulty in understanding certain parts of the reading passages."

"In my opinion the content of the test was too specialised. It would have been better if it covered different vocabulary areas. This test might have been easy for people studying biology or engineering etc. because it covered this field of vocabulary. But this test was not appropriate for people who study for instance literature or linguistics. It is quite difficult to cope with these specific texts (as given in the test) if one is not a native speaker and does not have any knowledge about the particular subject. I think each student should be given a test according to the subject he studies."

It is interesting that most of the adverse comments on subject content come from A.S.A.B.S. students taking Session TB, the Session designed for science and engineering students. No adverse comments

on subject content were made by the Sci/Eng students taking Session TA, the Session designed primarily for A.S.A.B.S. students. If there were to be only one version of Session II, on this evidence, it would seem that TA is preferable.

Familiarity with the Test. Practice Effect:

The question of practice effect also has an important bearing on test performance. We were not able rigorously to take account of this in the pre-test administration. We did, however, reverse the order in which the combinations of Sessions were taken at different Centres, but the numbers are too small to make any comparisons meaningful. It is clear from our discussions of students' replies below and our observations, that students generally felt more confident about their performance in whichever Session they had taken second. It also seemed that they were less pressured for time in whichever Session they took second. It does seem likely that students might perform better, or at least feel they do so, when they get more used to the format of the test.

The novelty of certain of the tasks should not be a problem once the test has been running for a few years. One A.R.E.L.S. principal commented:

"Familiarity with rubrics, procedures and format of the papers was the only serious problem for my students. Need to have practice tests available."

Invigilators expressed the need to put the students at their ease at the start of the Session when confronted by such a mass of papers. The invigilators themselves found difficulty with the large amount of paper and one recommended the Source and Answer Booklets should be in different colours. The wealth of instructions was seen as a difficulty by many and it was felt that this often increased time pressure in completion of the initial tasks in a Session. This was a particular problem in TA and TB when the complexities of filling out a multiple choice answer sheet had to be grasped. Several invigilators suggested that the repetition of instructions for some tasks had only served to confuse some candidates.

Invigilators made the following related comments:

"The considerable paperwork makes initial distribution confusing at times, however once this first hurdle is passed, the presentation seems fairly clear."

"The sheer bulk of paper appears to throw students ... once they embark on the separate sections it's OK but I think the overall impression possibly slows them down."

"It's often difficult to distinguish between the difficulty students have in coping with the format of the test and the questions. Is the ability to understand the test itself, part of the test?"

In the follow-up questionnaire we asked students whether they had any difficulty in adapting to the test. The results are summarised in Table 5A below.

TABLE 5A					
REPLIES TO QUESTION 7 OF THE FOLLOW-UP QUESTIONNAIRE					
In many ways this test was different from language tests you may have taken in the past. Did you have difficulty in adapting to it?					
	Lot of difficulty adapting	Some difficulty adapting	No opinion either way	Easy to adapt to	Very easy to adapt to
TO	$\frac{1}{45}$	$\frac{23}{138}$	$\frac{20}{38}$	$\frac{54}{47}$	$\frac{14}{10}$
TA	$\frac{1}{43}$	$\frac{10}{179}$	$\frac{15}{38}$	$\frac{24}{67}$	$\frac{13}{10}$
TB	$\frac{2}{42}$	$\frac{23}{128}$	$\frac{32}{39}$	$\frac{40}{39}$	$\frac{11}{6}$

Key: $\frac{\text{N.S. replies}}{\text{N.N.S. replies}}$

Most of the N.N.S. group appear to have had some difficulty in adapting to the T.E.A.P. formats though this was less of a problem for the N.S. group.

We have also collated the replies to Question 8 which have a bearing on this issue and include a representative selection below. The main problems seem to have arisen due to the bulky nature of the experimental pre-test. The amount of paper and the numerous rubrics necessary to set up realistic tasks will obviously be a problem until the test becomes established.

TO

"I was not accustomed to this type of examination before, both in G.B. and in my country. So, it became difficult to understand easily and, most of the time went in reading instructions. At the same time prescribed time was not enough."

TA

"The second test was easier to adapt to as I took a similar test yesterday. Yesterday I had no idea what was going on. I had some today!"

"Think the test was more easy than the first one or maybe we have got familiar with the test."

"These types of tests often examine the ability of a person to adapt to new styles of questions. I spent a lot of time sorting out the system."

"Much of my time was spent on searching and organising papers because the style of setting questions was very unfamiliar. The paper was not too bad for those who are used to the style."

"Both Sessions were confusing in the number of separate sheets of paper and the number of sub-sections. Straight numbering or lettering of each question would have been much easier."

"TA11 was very disjointed with questions, answers, instructions and text in different places. The test programme could be better arranged so that it is not so awkward to look for different sections for different tasks. Some explanations to how a task should be done are very confusing especially in the multiple choice tasks."

"There are too many parts in this examination and there are many tasks in each part. This is quite confusing."

"Layout of the paper not good as it involved too many pieces of paper all over the desk, very good test of my ability though."

"Personally, I found myself less able to think properly and answer the questions on the afternoon examination held on first day. In the morning examination next day, I found myself more able to do the tasks. In the first day exam. we wasted a lot of time in reading the questions to know exactly what was wanted from us to do, whereas in next day exam., we spent shorter time."

"Because I have the experience of yesterday, this exam. was more easy because instead of reading first all the questions, then the passage and then answer the questions, I began to answer the questions as I was reading. This result in more time and of course better result."

TB

"Some confusion at beginning of the test to actually decide what we had to do."

"Too much paper to cope with i.e. question sheets, questions, booklet etc. You end up trying to find bits and pieces."

"The system was very complicating. An answer book was provided and the source booklet was separate. They should give one task at a time so nobody could get confused. I myself read a whole section and when I read the question I had to read, it did not relate. By this time I was disheartened."

We recognise that the bulky nature of the T.E.A.P. pre-tests caused some difficulty, but would point out that it was always the intention that the final version of the test would be reduced on the basis of data arising out of the administration of these necessarily lengthier, experimental versions of the battery.

It was also felt that once the final version of the test was in operation and test booklets and back up materials available, students would be able to familiarise themselves much better with the overall organisation of the test prior to taking it.

A number of students made much shorter comments which are not included in the selections made above. We have attempted to quantify them in Table 5B below.

TABLE 5B
 QUANTIFICATION OF SHORT RESPONSES TO QUESTION 8 OF THE FOLLOW-UP QUESTIONNAIRE

	n	<u>Test too difficult</u>	<u>Test too simple</u>	<u>Test too long</u>	<u>Not enough time</u>	<u>Bad Test</u>	<u>Good Test</u>	<u>Tape too fast</u>	<u>Tape not clear</u>	<u>Play tape twice</u>
TO	N.S. (115)	-	13	9	-	5	9	1	1	-
	N.N.S. (294)	38	-	10	34	-	16	28	32	7
TA	N.S. (75)	-	4	5	-	-	6	1	1	-
	N.N.S. (420)	38	-	21	47	-	32	5	15	3
TB	N.S. (123)	-	9	2	-	-	7	1	2	-
	N.N.S. (308)	29	-	9	15	-	16	11	35	3

The main N.N.S. criticisms appear to be concerned with the difficulty of the test though the reverse is true for the N.S. group. Criticisms were also voiced by the N.N.S. group about the length of the test as a whole and that time was too short for completion of some tasks. The only solution to this dilemma would be to cut down on the number and/or size of the tasks which would serve to reduce the length of the test as a whole and increase the time available for completion of items within tasks. The remaining criticisms relate to the listening components of the test and we will discuss these in our examination of reactions to specific tasks below.

We would emphasise again that the pre-test was always seen by us as a rather bulky, experimental version of T.E.A.P. and the aim was to reduce its size once we had determined the best methods for testing what we wanted to test.

Reaction to the Idea of the Test:

In general there was a favourable reaction to the conception underlying T.E.A.P. The E.F.L. teachers attending the A.R.E.L.S. staff course on T.E.A.P. at Aldershot in June 1982 had expressed general approval for the aims and objectives of the examination (v. Appendix 5.7, pp.1072-1075) and their realisation in the pre-tests.

In general, teaching staff who assisted in the administration of T.E.A.P. were in favour of the tests because of their educational implications, the most important of which was the good 'wash back' effect it would have on teaching. This was encouraging as we had purposefully included in the test a number of activities our experience had shown to be useful exercises in teaching on pre- and in-session E.A.P. courses.

The following comments, which reflect on the idea of the test, were made by N.N.S. students in response to Question 8.

"I think it is a good thing that this is done at the beginning of the course to help the tutors become aware of our own shortcomings in order to help show where our weakness lie."

"It was good practice to make one realise that one's English is not as good as one thinks and that there is much more learning required."

"The tests were interesting and I enjoyed doing them."

"I think the combination of reading a text, listening and writing was a good idea."

"It tests one's ability to listen carefully and understand what is being spoken and personally that is the basis of any lecture a student attends. If he can grasp the fundamentals of a lecture, I think the student is well on the way to passing an examination. I think the test is marvellous in its structure and should be given to all overseas students (mind you, some of your home students could benefit from it too!!)."

"This test is very useful to measure English abilities of overseas students. I've had TOEFL, ELTS, Cambridge but, this is the most suitable test. However, it is also very important that the topics used in the test should be selected carefully."

"It has pointed some points about studying techniques out to me - very useful."

5.1.6.2 Tutor and Candidate Responses to Specific Test Tasks

In this section we deal with staff and student comments made in relation to individual tasks within each Session. For an outline of the nature of these tasks see pages 408-409 above.

The returns to Questions 1, 2, 4, 5 and 6 of the questionnaire which deal with specific components of the test are summarised in Appendix 5.6.2, pages 1066-1071. Reference will be made to the replies to these questions when we deal with each of the individual test tasks.

5.1.6.2.1 TO

T011 (Summary)

On the whole this was seen as a good test by both N.S. and N.N.S. students and by the staff. Most of the N.S. group thought there was sufficient time for its completion though about half of the N.N.S. group did not. In general, the amount of time allowed for each task was adequate for the N.S. group in all of this Session (v. Table 5EEE, p.1069). Staff thought that more time should have been allowed for

Part One (T011 and T012) though some felt that problems with time were connected to the difficulties students had in familiarising themselves with the general framework of the test.

T012 (Short Answer Reading Comprehension Questions)

This was also seen as a good test by the majority of the N.S. and N.N.S. groups and staff, though a sizeable number of the latter student group had difficulties with the time allowed. Staff commented on the need for passages containing argument and opinion as well as those which tested purely factual reading comprehension. Several students said they would have preferred to do T012 before T011. One student wrote that he deliberately ignored certain questions, e.g. 6 and 7 which he thought would take too much time.

T021 (Dictation)

The reaction to this was favourable from both the N.S. and the N.N.S. groups and less than half of the latter group were concerned about the time allowed. Student comments on this sub-test were concerned with the length of the phrases in one case and several candidates thought the tape should be played twice for checking purposes rather than through any difficulties engendered by the speed of the exercise. One student made the valid point that it was not mentioned on the tape that this was a dictation so he had taken it to be a note-taking exercise. A tutor stated that it was not clear whether students had to write down every word they heard or not. The question was also raised as to whether they should know the topic of the dictation beforehand.

T031 (Short Answer Listening Comprehension Questions)

This proved to be the task which the N.N.S. students gave the lowest rating to in Question 1 of the follow-up questionnaire (v. Table 5CCC, p.1067) and the one with which the majority had difficulties in terms

of the time allocated. The N.S. group commented that this was one of the best tasks and very few considered it a bad test. The majority of staff thought this was a good E.A.P. test, though a few thought it bad. A few would also have liked more time for the task.

This was the task which aroused the most concern and alarm amongst the N.N.S. group according to the teachers involved in the invigilation. The following points were made by staff:

"Students are encouraged (instruction p.13) to write notes on the parts of the lecture relevant to the questions on pp.14-16. Then in Task 2 they have to use these notes to summarise the whole task. A little unfair. For example, there was a large section on diet and food companies between Questions 3.2 and 3.3 on p.16, they could have written notes on and summarised."

"While listening for details or specifics it was almost impossible to think of gist. Therefore does not the following writing task become an exercise in paraphrase using the outline provided."

"The lecture should be delivered more slowly to compensate for the lack of visual signals."

"It starts too abruptly, not enough of a gap between 'The lecture will start now' and the start of the lecture. Instructions to look at p.13 should be included on the tape as otherwise some students are still reading the general introduction to the task when the lecture starts."

"The content of the lecture was too dense for students to be able to follow the lecture outline at the same time and not get lost and/or confused. Furthermore, it is artificially difficult to set written comprehension questions to be read and answered at the same time as students are listening to a lecture. A more realistic/helpful form of outline might be the blackboard work that could accompany such a lecture. Otherwise it seemed a pretty good test!"

"My class failed to study the outline and questions very thoroughly in the time allowed for this and were therefore distracted by trying to read the questions at the same time as listening to the lecture."

"The outline was difficult to follow. It seemed artificial and too rigid. Could the outline be more in the form of what would go on an O.H.P. or blackboard?"

"Weaker students just gave up on Tasks 3.1 and 3.2. Difficult to see what can be done about this."

"In the listening task T031 weak students were helped to find their place when following the lecture outline by the brighter ones turning over the pages at the correct place."

"Difference in delivery speed between taped instructions (very slow and comforting) and test parts (a sudden shock). Should this difference be less?"

The following comments were made by the students on Task T031.

"Too difficult, needs double playing of tape."

"The test was very difficult for all the foreign students. It shouldn't be like this. The worst part was the listening comprehension. There should be pauses during the tape is playing and there should be a second playing."

"Part III, Task I proved to be difficult for me who is not 'attuned' (unfamiliar with) to a British lecture. The time lapses between the statements and time provided for taking down notes seem to be short and inconsistent. For instance in statements related to Questions 2, 3 and 4 - the statements came one after the other giving us insufficient time to take down notes."

"Outline given in the answer sheet causes confusion due to a complex structure of question sentences. It would be better if questions are designed in simple sentences and talk should be slower than usual with enough pauses in between. Hope you will consider it. Otherwise it is a very good test."

"In general terms, I think that the test is well designed. However, there are two points that could be improved: 1. The length of time: if have had more time (just a few minutes more) we could have been more relaxed. 2. The use of an outline to follow the lecture is always very disturbing, at least, for me."

"I understood the lecture but I was confused with what was written in the source booklet."

"I found it very difficult to listen to the text and to write in the space provided for the answers simultaneously. I could not concentrate on the text."

"The quality of the tape was terribly boring. The speaker appeared to have as much interest in what he was saying as I did listening to him."

"This is better than Session II of test, this allows for good understanding of content by foreign students."

"Part 3 is much more difficult than Part 2 so it should be repeated. But, I think this test is very interesting because it reminds me that I am weak in listening and getting the little information from it."

"The test was very interesting, we have done such a kind but not enough. I really feel after this test that I need far more practice in listening and note-taking even in my own subject."

"I think that this type of text it is important to the student know his level of English and what he have to improve and study and how many hours he needs to watch television to improve the listening that is my problem."

T032 (Summary)

Most staff thought this a good test as did the N.S. students. The majority of the N.N.S. group, on the whole, thought it a bad test.

Some staff thought it should be made clearer that if students have problems with the listening they can make use of the outline in the writing task. Others were less sure that they could do much with the outline if they had not understood the lecture.

Several students commented that they would have preferred a writing task separate from task one (T031) and one said it would be better to write one's own opinion on a topic.

T041 (Multiple Choice Questions on Knowledge of Grammar)

The N.N.S. group liked this part of the test most of all and as Question 2 shows they thought that was the part they had done best in. The N.S. group on the other hand liked this test the least. Staff were divided and many had no opinions either way.

5.1.6.2.2 TA

TA11 (Multiple Choice Reading Comprehension Questions)

Staff invigilating thought this a good to very good E.A.P. test and that time was sufficient for most candidates. In general this was considered a good test by both the N.S. and N.N.S. groups with about a third of the latter group having problems with time. Staff thought the subject content might have been less suitable for the N.S. group.

TA12

(Gap-filling: Identifying Omissions in Text and Completing the Gaps)

Staff were less divided in this task and either thought it a good test or had no opinion either way. Most thought more time was needed. The majority of students thought this a good test though there were some problems with time in the N.N.S. group. The N.S. group appear to have had very little problems with the time allowed in any of the tasks in this Session. The question was raised about the face validity of this task by one tutor:

"I am not too sure of the objectives of Part 1, Task 2. It appears to be a discourse analysis and seems to veer away from the academic study type sections found in the other parts of the test, i.e. the other sections seem more relevant to the needs of students entering colleges and universities in England."

Student comments ranged from lack of familiarity with this type of format to complaints about its length and difficulty. Several asked for the test to be cut down by indicating where the words were missing.

TA13 (Short Answer Reading Comprehension Questions)

Most staff thought this very good and the rest thought it good. Nearly all the staff thought more time was needed for this sub-test in particular. Most students thought it a good test, but of the three reading sub-tests this created the greatest difficulty as regards time. One invigilator commented:

"Even greater emphasis on need to read questions first may have obviated serious timing problems."

TA21 (Short Answer Listening Comprehension Questions)

Most staff thought this was a good E.A.P. test and the time sufficient. The majority of students liked this listening test though about half of the N.N.S. group had problems with time. The general feeling among staff and students was that TA21 was easier than TB21.

One teacher thought:

"Questions were detrimental to listening for gist and 'keeping the thread', but it is realistic of a discussion, which is good."

One tutor thought the time given for reading the outline was too generous. Of the four listening tests, TA21 provoked the least adverse comments. A selection of the limited number of comments made by the N.N.S. students on this task are included below.

"There should be pauses during the listening comprehension (during the time the tape is playing)."

"The interview was good and the questions clear, but it would have been better if we were able to make our notes as if we were receiving lectures in class, then answer the questions later. Having to search for the lines on which to make the appropriate notes was difficult and details are left out easily."

"I would have found it easier to make my own notes. Therefore a blank sheet would have been more useful for my own notes. Listening comprehension much clearer than the one given last week (TB21) and this seems to make the test easier."

"On the listening test the tape was blotted out momentarily by people turning the page. This ended in me missing an answer."

"The tape recording was unclear and as I was sitting near the back of the hall the recording caused sound reverberations."

"Woman's voice unclear."

"The outline in 'listening', help to follow the speakers, but in a normal lecture we would not have this, so to listen and understand is more difficult."

TA31 (Summary Writing)

Staff reactions to this test were mixed varying from very good to bad, but the time was considered sufficient. The majority of students thought it a good test, though about a quarter had problems with the time. One tutor felt there was not sufficient information to be extracted from the third reading passage. Another thought more space was needed for the answer.

Some tutors felt that it would be difficult for the N.S. group to give their own opinions on the problems of overseas students in this task.

TA32 (Editing Task)

Similar reactions as in TA31 from staff, estimates varying from very good to bad. Time was considered sufficient. Students in general thought this a good test and there is little difference in their reaction to this and TA31. They seem to have had less problems with this in terms of time.

One student made the valid criticism:

"It is not very clear how far just only correction or rewriting is supposed to be done."

5.1.6.2.3 TB

TB11 (Multiple Choice Reading Comprehension Questions)

Staff thought this task good to very good and the time allowed sufficient. The majority of students agreed, but about one fifth of the N.N.S. group had trouble with the time allowed. Native speakers had very few problems with the time limits in any of the tasks in this Session (v. Table 5EEE, p.1069).

TB12

(Gap-filling: Identifying Omissions in Text and Completing the Gaps)

Staff either thought this task was very good or good and that sufficient time had been allowed. N.S. students on the whole thought it was a good test though slightly fewer favoured it as compared to TB11. About one third of the N.N.S. group had problems with the time on this. One teacher commented that it was, in his opinion, a test of writing ability rather than reading because it required a "high explicit ability in grammar/cohesion". A few students commented that they would prefer this task if the places where words had been deleted were indicated.

TB13 (Short Answer Reading Comprehension Questions)

Staff either considered the task good or very good, but felt the time was not sufficient. The N.N.S. liked this least of the reading tasks in this Session and also thought they had done worse on it.

About two fifths of the N.N.S. group appear to have had insufficient time to complete the task.

TB21 (Short Answer Listening Comprehension Questions)

Staff opinions were mixed on the listening component ranging from very good to bad. A number of staff did not think there was sufficient time allowed. This was the task the N.N.S. group liked least in this Session with more critical opinions of the test than complimentary. Staff made several comments on this test. These are listed below:

"Need instruction to pause tape before discussion starts (after 3 minutes pause for reading) so student can turn back to the beginning and gear up for listening etc."

"Time pressure distorted the authenticity of the task. In general Part 2 was seen as extremely difficult."

"The arrival of Concorde overhead forced me to stop the tape in mid-interview."

"In an examination hall such as this where the acoustics are bad, certain candidates might be thrown by this listening task."

"Although the interview had been prepared by the third reading passage, I thought there was a very heavy scientific bias in it which would make it very difficult for non-science students. On the other hand I feel the questions were not as 'disturbing' as those in the equivalent part of Session I."

On the whole, staff thought the tests in TB were more difficult for the N.N.S. group than those in TA. This again raises the serious problem when constructing parallel forms of gauging the complexity and functional and referential range of texts satisfactorily.

Quite often extraneous noise interfered with the listening task and further complicated the issue. In my own experience building work on the floor above and excessive traffic noise from the main road outside did little to aid the concentration and comprehension of the examinees. Also on two occasions, candidates who were having problems drew attention to this aloud, thereby affecting other candidates. All these performance constraints become serious difficulties when the tape is only played once.

As regards the TB test format, nearly all the candidates' comments offered in the follow-up questionnaire related to TB21, the listening component. Quite a number of students commented on the lack of time allowed and several on the quality of the recording. We include below a representative selection from the comments made.

"Tape should be run at a slower speed then stopped for taking notes."

"In relation to Question I, part 2, my only comment is that there are too many factors, personal as well as environmental that may cause disturbance in concentration and hearing of the tape, thus affecting to a large degree the results which need not then be a fair reflection of the candidates' ability to understand and grasp the dialogue."

"Sometimes we have to contend with noise from outside the classroom, which makes listening to the taped interview difficult. I find it hard to concentrate on the speakers in the interview and what they are actually hitting at or trying to drive home. All in all, it is a good test of our capabilities in the language, as listening and comprehending what is said is the main feature of an academic course."

"Too many questions were set for the test within the time available. One reading of the recorded passage which was very fast was not enough for one to grasp the main stream of the argument."

"Should be played twice not once. If this is done, the fifteen minutes to answer the questions might be enough."

"To understand the spoken part of the test there should be less tasks to do. You can't read the question, take notes and listen at the same time over a long period. Shorter lectures would be more adequate. The number of lectures could increase therefore."

"When the candidate tries to concentrate both on the tape and the short notes in the book it is difficult to follow the tape in part two."

"Live voice rather than tape recording would be preferred."

"The tape was too fast to comprehend. In any case it was on the wall. Lecturers often speak whilst in front. If the loudspeaker was in front, it would have been better."

"No cassettes please. I prefer to hear and look at the same time when someone speaks."

"The poor tape recording made people's accents sound worse than they really were. I'm sure if those speakers had been here in person, I would have no problem grasping the subject and details. With a tape, a person's attention can also tend to wander."

TB31 (Summary Writing)

Most staff thought this to be a very good task or a good one. Time allowed generally thought to be sufficient. This was the task the N.S. group liked the most and the N.N.S. group on balance thought it good rather than bad. About half of the N.N.S. group had problems with the time allowed here.

TB32 (Editing Task)

Staff were less enthusiastic about this task, opinions ranging from good to bad. About half the staff replies indicated more time was needed on this. The N.N.S. group preferred this to TB31. Again, about half the N.N.S. group had problems with the time allowed.

5.1.6.3 Conclusions

This concludes the description of our attempts to get feedback on students' responses to the tests they had sat. These data provided us with a wider perspective for interpreting the more objective analysis described in the following section. They serve as a reminder that the performance of individual candidates in tests will be influenced both by environmental factors independent of the test material itself and the affective responses of candidates to the test materials. These factors are likely to have introduced a certain degree of error into the measurements under consideration.

Whereas the majority of students taking the test completed a questionnaire, we managed to get feedback only from a very limited number of staff as most of the invigilations were conducted by the writer. Staff who filled in a follow-up questionnaire were tutors on pre-sessional courses at the Universities of Exeter, Lancaster, Reading and Southampton. The comments these tutors made on the difficulty their students had in coping with various tasks must be understood in the context of their teaching situation. The few strong criticisms of the tests, e.g. on complexity and time allowed, all came from teachers in charge of classes with extremely limited proficiency. Their students were in the bottom classes of pre-sessional courses where the students were already pre-selected on

the basis of their need for remedial English language tuition.

T.E.A.P. is a proficiency test and not an achievement test. Its purpose is to establish whether students possess the baseline competence to cope with the language activities involved in an academic course of study. It is not the function of the proficiency tester to aim the level of the test at the standard of the weaker students who might be taking it. The proficiency test should attempt to mirror an external standard which has to be reached irrespective of the differing abilities of the students who might be sitting it.

This is not to say we were unreceptive to the comments of these teachers and where corroborative evidence from the wider based student sample and the data analysis (v. Section 5.2 below) supported their criticisms, the necessary corrective action was taken in preparing the final version of the test.

Timing is obviously a factor that had to be monitored carefully. The pre-test N.S. group, as with the N.S. students we had trialled the tests on in the summer of 1982, encountered hardly any problems with the time allowed, in fact most N.S. students had more than enough time. In general, for a majority of the N.N.S. group (v. Appendix 5.6.2, p.1069), the time allotted was sufficient for most of the tasks.

Where difficulty arose for the N.S. group at the item level (v. Section 5.2, p.446 et seq.) the necessary steps were taken to remedy the deficiencies. Where difficulty occurred for large numbers of the N.N.S. group in completing tasks in the time allowed, steps have been taken to resolve the problem. These amendments are referred to in our discussion of the test analysis of the data relating to specific tasks in Section 5.2 below.

The comments that were made by both staff and students were greatly appreciated by the writer and were all borne in mind when T.E.A.P. was revised and shortened for the final piloting.

5.1.7 Handling the Data

The T.E.A.P. data were keyed between November 1982 and January 1983 via an interactive terminal linked to the Honeywell, Level 64 mainframe. The only exception to this was the O.T. item response data which, since they had been coded on standard A.E.B. O.T. sheets, were entered on to the computer via a document reader. Wherever it was possible the data were validated subsequent to being entered into the computer.

The computer analysis was performed entirely on the mainframe using a variety of programs which fall into three groups.

- (i) The first of these was the Honeywell written statistical package STATPAC. The hierarchical cluster analyses and the principal component analyses were performed using STATPAC.
- (ii) A small number of standard departmental programs were written in FORTRAN. In particular, these were used to produce the correlation matrices and the factor rotations on the principal components produced by STATPAC.
- (iii) A large number of 'ad hoc' programs written in FORTRAN by Mr. K. Trinder and Mr. J. Wilmut, members of staff of the Board's Research and Statistics Division, were used to manipulate and combine files and to produce frequency tables, simple statistics (such as means and standard deviations), cross-tabulation frequencies, item and task analyses and many other specific analyses.

The types of analyses which were performed were dependent on the structure and size of the particular data set under examination and although other sophisticated statistical analyses could have been employed (such as multiple regression and oblique factor rotations), it was felt that these were not likely to yield any further information of significant value.

5.2 DATA ANALYSIS AT THE ITEM LEVEL

5.2.1 Introduction

Before examining the pre-test scores of candidates at the task and task composite level, we first carried out an individual analysis of every item in the pre-test. The tables relating to the analysis we carried out at the item level are included as Appendix 5.8, pages 1076-1095.

We were concerned with how individual items, in each of the macro-skill areas, performed in each of the Sessions. The analysis was dealt with in the following order:

1. Reading

- 1.1 T012 Short answer questions on reading comprehension
- 1.2 TA11 Multiple choice questions on reading comprehension
- 1.3 TA12A Gap-filling: identifying omissions in text
- 1.4 TA12B Gap-filling: completing the gaps
- 1.5 TA13 Short answer questions on reading comprehension
- 1.6 TB11 Multiple choice questions on reading comprehension
- 1.7 TB12A Gap-filling: identifying omissions in text
- 1.8 TB12B Gap-filling: completing the gaps
- 1.9 TB13 Short answer questions on reading comprehension

2. Listening

- 2.1 T021 Dictation
- 2.2 T031 Short answer questions on listening comprehension
- 2.3 TA21 Short answer questions on listening comprehension
- 2.4 TB21 Short answer questions on listening comprehension

3. Linguistic Competence

- 3.1 T041 Multiple choice questions on knowledge of grammar

4. Writing tasks

- 4.1 T011, T032, TA31, TB31 Summary based on information extracted from written and/or spoken sources
- 4.2 TA32, TB32 Editing tasks

For all items in each of the tasks, where it was appropriate to do so, we computed measures of item difficulty and of item discrimination in the manner described below. These statistics were calculated separately for the N.S. (native speaker) and N.N.S. (non-native speaker) sub-groups.

As a measure of item difficulty we calculated the proportion of marks available for each item actually gained by the candidates. This measure is called the facility index. Thus, a facility index of 0.92 indicates that the group of candidates concerned were successful in gaining 92% of the marks. In the case of an item scored dichotomously one-nought, the mean mark for the item was 0.92.

The measure of item discrimination used was the point bi-serial coefficient of correlation (r_{pbi}) described by Guilford (1965, pp. 322-325). It is the appropriate correlation coefficient where one of the variables is dichotomous. Used as an index of discrimination r_{pbi} indicates for the item the correlation between success in the item and total score in the task of which the item is a part. These data are shown in Appendix 5.8, Tables 5HHH-5XXX (pp.1076-1095). In each table, columns 1 and 2 show the facility indices for the N.S. and N.N.S. groups respectively. Column 3 contains the discrimination index (r_{pbi}) for the N.S. group and column 4, that for the N.N.S. group. Columns 5 and 6 show the omits for the two groups.

At the foot of each table the final row of figures shows:

- (1) the number of N.S. students taking the task,
- (2) the number of N.N.S. students taking the task,
- (3) the mean mark gained in the task by the N.S. group,
- (4) the mean mark gained in the task by the N.N.S. group,
- (5) the standard deviation of marks gained in the task by the N.S. group and
- (6) the standard deviation of marks gained in the task by the N.N.S. group.

Emphasis was placed, in the analysis, on the facility index rather than on the discrimination index for each item. We first checked the facility index of each item to determine whether there were any

cases where N.N.S. did better than N.S., where they performed equally or where there was a facility value lower than 0.8 for the N.S. sub-group. In these cases the item is reviewed below. We then checked the N.N.S. r_{pbi} figures as a confirmatory procedure to see how much individual items were contributing to the power of the task as a whole. Though we include the r_{pbi} figures for the N.S. group we feel they should be interpreted with a great deal of caution as the high N.S. facility indices inevitably distort the discrimination index for the N.S. group.

For the multiple choice items in tasks T041, TA11 and TB11, we also looked at the response patterns for the N.N.S. sub-group only, to see how well the distractors were performing. We noted those cases where distractors were strong and where there was a response rate of less than 3% for an option. Where the facility value for an item was above 0.8 we were not unduly worried about those effects.

The omit rates for each item were also considered to determine how that item had behaved and also to help us in decisions about whether the tasks should be reduced in size, or provision made for extra time.

We now look at the results of the item analyses for each of the tasks grouped under the broad macro-skill headings of reading, listening, writing and grammar.

5.2.2 Reading

T012 Short Answer Questions on Reading Comprehension (v. Table 5HHH, p.1077)

The facility and discrimination statistics seem acceptable for all items. There are high omit rates on items 6 and 7 and more critically on items 12, 14 and 15. Items 6 and 7 are suspect in that we are not, with the advantage of experience and hindsight, convinced whether the form they are couched in actually measures the ability to work out the meaning of words in context. In the opinion of the Project Working Party, this skill was adequately covered in

Tasks TA12/TB12; the detecting of omissions and gap-filling was seen as a more appropriate format for testing lexical competence (v. Hawkey 1982). We feel the high omit rate on items 12, 14 and 15 was due to candidates running out of time. These later items are testing 'higher order' enabling skills and this, in our view, justifies their retention. We advise, therefore, that items 6 and 7 should go, especially as informal observation during the pre-testing showed some students spending anything up to five minutes trying to locate the answer for each.

Item 5 also proved difficult for both groups of students. Since we do not think the framing of the question was responsible for this, but rather the difficulty of the higher level skill involved, this item is to remain.

TA11 Multiple Choice Questions on Reading Comprehension
(v. Tables 5III and 5JJJ, pp.1078-1079)

Items 2, 11 and 13 presented a lot of difficulty to the N.S. group, possibly because of inadequacies in the wording and they should be reviewed. The r_{pbi} for these same items accord with those of other items in the task for the N.N.S. group. The omit rate rises on items 11, 12 and 13 for the N.N.S. group, but not to an unacceptable level. The options seem to be fulfilling the conditions we have laid down. Only in the case of item 7, which had a very high facility value for the N.N.S. group, did distractors fail to attract less than 3% of candidates.

TA12A Gap-filling: Identifying Omissions in Text
(v. Table 5KKK, p.1080)

Item 8 proved the most difficult for the N.N.S. group and it was by far the most difficult item for the N.S. group. Why this should be so is not immediately apparent. The items seemed to discriminate well between the N.S. and N.N.S. groups and the majority of r_{pbi} 's for the N.N.S. group were above 0.5. The facility values for the N.S. group, with the exception of item 8, were satisfactory. Omit rates are not indicated for this task as it would have dramatically increased the time needed to mark this exercise if the marking template had to be removed from the script on every item.

TA12B Gap-filling: Completing the Gaps

(v. Table 5LLL, p.1081)

Items 5, 8 and 21 proved very difficult even for the N.S. group. The successful completion of item 5 required reading beyond the immediate context and, given the sentence-based nature of the rest of the gaps, this item is perhaps inappropriate in this context. The number of acceptable responses to item 8 were limited and this may account for a lot of the difficulty in coping with item 8. Similarly item 21 seems just to have been a difficult item. It is suggested that they be omitted from the final version, especially in view of the high N.N.S. omit rate on the last three items of the battery.

There was a wide range in the facility indices of both groups. As a test it spreads out both N.S. and N.N.S. The majority of the r_{pbi} 's are above 0.5 for the N.N.S. group.

There was a feeling among students and teachers that this sub-test was both too difficult and too long and a decision was taken by the Project Working Party to reduce the length of the text and excise a number of the more difficult items.

TA13 Short Answer Questions on Reading Comprehension

(v. Table 5MMM, p.1082)

The gap in difficulty levels between the N.S. and N.N.S. groups was narrower on this than on any other task, with a lot of overlap between the distributions.

The omit rate was very high on items 11 and 12, particularly for the N.N.S. group, indicative of students running out of time. Given that we are reluctant to increase the time allotted for this task, we would suggest taking out items 2 and 11. We suggest item 11 be deleted rather than 12 as we are not happy with the very low N.S. facility index on this. Item 2 should be deleted for the same reasons as those advanced for deleting items 6 and 7 in T012 (v. pp.448-449).

TB11 Multiple Choice Questions on Reading Comprehension

(v. Tables 5NNN and 5000, pp.1083-1084)

Items 8 and 13 presented the N.S. group with undue difficulty and need reformulating. As regards response frequencies, distractors 4B and 13C need to be reviewed.

The facility indices for the N.S. and the N.N.S. groups are closer on this task than on many of the others. The gap between the N.S. and N.N.S. groups is much closer here than it was in TA11. There is not the increase in omits towards the end of the task for the N.N.S. group that was evident in TA11.

TB12A Gap-filling: Identifying Omissions in Text

(v. Table 5PPP, p.1085)

All items seem to be working well here. The facility index for the N.S. group does not fall below 0.73 and that of the N.N.S. group ranges from 0.34 to 0.86. The majority of the r_{pbi} s for the N.N.S. group are above 0.5. As for TA12A we have no record of omits in this task because of the difficulties this would have presented to the markers who were using a template for correcting the items.

TB12B Gap-filling: Completing the Gaps

(v. Table 5QQQ, p.1086)

There seem to be no obvious problems here. The overall mean for the N.N.S. group is higher in TB12 than in TA12. The omit rate was far lower for TB12B as against TA12B, with only four item omit rates of over fifty candidates as against sixteen. If we take out the items causing the most problems in TA12 and reduce the length of that passage, as members of the Project Working Party suggested, then we will have to take out an equivalent number of items here to keep the numbers equal and the balance of skills tested roughly the same. To this effect we suggest omitting item 10 which proved the most difficult for the N.S. group and the requisite number of additional items with the proviso that this should not disturb unduly the balance of the enabling skills being tested.

TB13 Short Answer Questions on Reading Comprehension

(v. Table 5RRR, p.1087)

Item 1 proved a very difficult item for the N.N.S. group and because of its important initial position in the task, we feel it should be taken out or moved to the end of this sub-test. Items 9, 10 and 11 proved difficult for both N.S. and N.N.S. groups and though the wording of 11 was potentially ambiguous, we can find no similar explanation for 9 and 10. There was a very high omit rate on items 10, 11 and 12 and in the follow-up questionnaires the highest complaints about lack of time in this Session were recorded for this task.

We would recommend omitting item 11, as well as item 1, from the final version.

The gap between the N.S. and N.N.S. groups is far greater on this task than is the case in TA13. The N.N.S. group do far worse on this task than on TA13, the N.S. group perform about the same.

5.2.2.1 Conclusions

Where it is necessary to remove items to reduce the length of a sub-test we would recommend that items which seem to focus on lower rather than higher order skills should be removed. If the items focusing on the higher order skills can be reformulated this would be preferable to removal.

At the item level it would be difficult to state that any of the reading sub-tests are working badly. Items with r_{pbi} lower than 0.3 are extremely rare for the N.N.S. group. This is indicative that the items are discriminating well amongst the N.N.S. group.

It does appear though that the items in both multiple choice formats have, on the whole, lower r_{pbi} s than occur in the other formats.

The data analysis revealed a steep rise in the omit rates for both N.S. and N.N.S. groups at the ends of tasks TA13 and TB13. This ties in with the students' criticisms of shortage of time for these tasks discussed in Section 5.1.6.2 above. The reduction in the

size of these tasks that we have recommended, should help here. It may well be that the occurrence of high omit rates in the N.N.S. group reflects the relatively large number of post-graduate students on pre-sessional courses that our sample contains. The N.N.S. population as a whole might not have found so much difficulty in completing the tasks in the time allotted.

In general, the facility indices for the N.S. group hardly ever falls below 0.7 on any item and for the N.N.S. the indices seldom dropped below 0.2. The majority of the facility indices for the N.N.S. group were between 0.3 and 0.8. The items would, on the whole, seem to have been pitched at a suitable level of difficulty.

5.2.3 Listening

T021 Dictation (v. Table 5SSS, p.1088)

Items 12 and 14 proved to be unduly difficult for the N.S. group and need modifying. There was a large gap between the N.S. and N.N.S. groups in performance on this task.

T031 Short Answer Questions on Listening Comprehension (v. Table 5TTT, p.1089)

Item 4 had the highest drop out rate amongst the N.N.S. group and caused the greatest problems for the N.S. group. Our earlier suspicions that the meaning of the idiom involved could not actually be worked out from the context were confirmed and as it is unlikely that the item is testing listening comprehension, we suggest it be removed from the final version. Items 2, 3, 4, 6, 9, 12 and 13 also proved difficult for some of the N.S. group, but the feeling amongst the Project Working Party was that they should be retained as they were not unduly difficult for this group. The facility indices for the N.N.S. group were very low on most of these items. The gap between the N.S./N.N.S. groups was even wider on this task than in T021.

TA21 Short Answer Questions on Listening Comprehension
(v. Table 5UUU, p.1090)

Item 9 proved to be unacceptably difficult for the N.S. group. Re-phrasing of the question is necessary. There are increased omits on the last two items, but this may reflect that these items are testing more difficult skills rather than insufficient time. Of the four listening tasks, this task proved to be the easiest for the N.N.S. group.

TB21 Short Answer Questions on Listening Comprehension
(v. Table 5VVV, p.1091)

By contrast with TA21 this proved to be a very difficult task for the N.N.S. group and fewer questions had been set than in T021, T031 and TA21. It contained a greater range of accents which, though intelligible, might have added to candidates' difficulties. It was also considered to be more dense than TA21 by a number of invigilators and candidates. The task is based on an interview recorded from the radio and the quality, though deemed satisfactory, was perhaps not as good as that of TA21 recorded live in a studio. It illustrates again the problems of getting parallel texts of equivalent difficulty for different Sessions. We noted in Chapter 4 the theoretical and practical deficiencies which marked our attempts to construct an a priori specification of text in terms of complexity and functional and referential range.

There is a big disparity in the performance levels reached by the students in TA21 and TB21. In the latter, the students have been asked to do a lot less and yet have done far worse. Given the greater difficulty of TB21 for both N.S. and N.N.S. groups and the high omit rate for the N.N.S. group, it would probably be advisable to take out item 1 and possibly item 5 and make the number of questions the same as in TA21 by adding 3-5 easier questions.

The follow-up questionnaire indicated that many overseas students felt there was inadequate time for them to cope with this task.

5.2.4 Linguistic Competence

TO41 Multiple Choice Questions on Knowledge of Grammar

(v. Tables 5WWW and 5XXX, pp.1092-1095)

Items 21, 45 and 57 were much more difficult for the N.S. group than any of the other items in this sub-test. All three items were testing the same area of number agreement and it may be that this presents great difficulty to N.S. as well as N.N.S. students. It is debatable whether they should be removed from this part of the test. The r_{pbi} for each of these three items is not noticeably different from that of other items in the task.

As regards response frequencies (v. Table 5XXX, p.1094) the following distractors need to be reviewed since they attracted very few candidates in items where a substantial proportion got the item wrong.

<u>Item No.</u>	<u>Options</u>
1	B, C
9	A, B
11	A
15	A
22	A
25	B, C
41	D
43	D
50	B

5.2.5 Writing

5.2.5.1 Introduction

For the reasons described in Chapter 4, it was decided that all the writing tasks would be marked using all or part of a set of seven analytic criteria. Details of these were given in Table 4J (p.392) and are repeated in Table 5C below. This table shows that each criterion takes the form of a four point attribute writing scale, 0-3. The points on these scales are designated by a behavioural description of the appropriate level of performance in the attribute.

The pre-test battery contained six writing tasks (T011, T032, TA31, TB31, TA32 and TB32). The first four of these are exercises in summary; the last two are editing tasks.

Table 5D below shows the group means for each of the criteria applied to the six writing tasks. Only the accuracy criteria 5-7 were applicable to the editing tasks.

The writing tasks were marked by three markers chosen on the basis described in Chapter 4 (p.399). Each marked two of the writing tasks as indicated in Table 5D below.

TABLE 5C
T.E.A.P. ATTRIBUTE WRITING SCALES

1. Relevance and Adequacy of Content
 0. The answer bears almost no relation to the task set. Totally inadequate answer.
 1. Answer of limited relevance to the task set. Possibly major gaps in treatment of topic and/or pointless repetition.
 2. For the most part answers the task set, though there may be some gaps or redundant information.
 3. Relevant and adequate answer to the task set.
2. Compositional Organisation
 0. No apparent organisation of content.
 1. Very little organisation of content. Underlying structure not sufficiently apparent.
 2. Some organisational skills in evidence, but not adequately controlled.
 3. Overall shape and internal pattern clear. Organisational skills adequately controlled.
3. Cohesion
 0. Cohesion almost totally absent. Writing so fragmentary that comprehension of the intended communication is virtually impossible.
 1. Unsatisfactory cohesion may cause difficulty in comprehension of most of the intended communication.
 2. For the most part satisfactory cohesion though occasional deficiencies may mean that certain parts of the communication are not always effective.
 3. Satisfactory use of cohesion resulting in effective communication.
4. Adequacy of Vocabulary for Purpose
 0. Vocabulary inadequate even for the most basic parts of the intended communication.
 1. Frequent inadequacies in vocabulary for the task. Perhaps frequent lexical inappropriacies and/or repetition.
 2. Some inadequacies in vocabulary for the task. Perhaps some lexical inappropriacies and/or circumlocution.
 3. Almost no inadequacies in vocabulary for the task. Only rare inappropriacies and/or circumlocution.
5. Grammar
 0. Almost all grammatical patterns inaccurate.
 1. Frequent grammatical inaccuracies.
 2. Some grammatical inaccuracies.
 3. Almost no grammatical inaccuracies.
6. Mechanical Accuracy I (Punctuation)
 0. Ignorance of conventions of punctuation.
 1. Low standard of accuracy in punctuation.
 2. Some inaccuracies in punctuation.
 3. Almost no inaccuracies in punctuation.
7. Mechanical Accuracy II (Spelling)
 0. Almost all spelling inaccurate.
 1. Low standard of accuracy in spelling.
 2. Some inaccuracies in spelling.
 3. Almost no inaccuracies in spelling.

TABLE 5D CONTINGENT MEANS FOR EACH OF THE CRITERIA APPLIED IN THE ASSESSMENT OF WRITING

TASK	T011		T032		TA31		TA32		TB31		TB32	
	<u>NNS</u>	<u>NS</u>	<u>NNS</u>	<u>NS</u>	<u>NNS</u>	<u>NS</u>	<u>NNS</u>	<u>NS</u>	<u>NNS</u>	<u>NS</u>	<u>NNS</u>	<u>NS</u>
1	2.11	2.63	2.00	2.67	1.60	2.00	1.83	2.51	1.83	2.51		
2	1.80	2.42	1.73	2.35	2.03	2.63	2.14	2.59	2.14	2.59	Not applicable	
3	1.85	2.53	1.68	2.41	2.15	2.73	2.13	2.72	2.13	2.72	Not applicable	
4	2.20	2.86	1.93	2.74	2.29	2.89	2.58	2.83	2.58	2.83		
5	1.98	2.79	1.82	2.81	1.55	2.68	1.59	2.16	1.59	2.16	0.73	1.73
6	2.27	2.57	2.06	2.45	2.30	2.49	1.98	1.77	1.98	1.77	2.14	2.20
7	2.11	2.42	1.91	2.45	2.37	2.46	2.34	2.09	2.34	2.09	2.06	2.20
	n=321	n=125	n=308	n=120	n=400	n=76	n=415	n=75	n=284	n=123	n=293	n=119
	Marker One			Marker Two			Marker Three					

Overall, the means of the N.S. and the N.N.S. groups on each criterion are noticeably different. The gap between the N.S. and the N.N.S. group appears to be consistently the widest on criterion 5, grammatical accuracy.

A surprisingly low mark for the N.S. group is recorded on criterion 1 in TA31 and merits further examination of both the task and the examiner's interpretation of the relevance and adequacy of the answer to the question set. It may well be that as this question relates partly to the problems overseas students see themselves as having had, it was unsuitable for a native speaker population.

5.2.5.2 Summary Tasks: T011, T032, TA31 and TB31

It was regarded as important to investigate the relationship between the criteria to see whether it would be profitable to merge any of them, for this would simplify the marking of the writing tasks. One way of doing this was to study the correlation matrix for the four writing tasks which were assessed using all seven criteria. These matrices are shown in Table 5E below. Interpretation of these data is made difficult by the fact that the same criteria correlate differentially across markers as well as across tasks. The most striking difference is the greater homogeneity of the correlations in the two writing tasks assessed by marker one. The lowest correlation is 0.54 between criteria 1 and 7, whereas these criteria only correlate at 0.18 and 0.00 for markers two and three respectively. Most of the correlations between the criteria scores awarded by markers two and three are below the 0.5 level.

It is possible that marker one is applying the assessment criteria in a different fashion to markers two and three, though given that they were marking different tasks written by different candidates, we cannot be certain of this.

TABLE 5E
 PEARSON PRODUCT MOMENT COEFFICIENTS FOR THE SEVEN
 ASSESSMENT CRITERIA APPLIED TO THE WRITING TASKS

<u>MARKER ONE</u>								
T011								
Criterion	1	2	3	4	5	6	7	
1	-	.74	.63	.64	.57	.57	.54	
2		-	.73	.71	.64	.63	.62	
3			-	.77	.71	.67	.67	
4				-	.75	.64	.71	
5					-	.64	.66	
6						-	.71	
7							-	
n = 321								
T032								
Criterion	1	2	3	4	5	6	7	
1	-	.79	.68	.73	.66	.59	.59	
2		-	.83	.80	.79	.68	.68	
3			-	.83	.82	.68	.69	
4				-	.86	.76	.77	
5					-	.72	.74	
6						-	.81	
7							-	
n = 308								
<u>MARKER TWO</u>								
TA31								
Criterion	1	2	3	4	5	6	7	
1	-	.47	.42	.30	.25	.19	.18	
2		-	.71	.44	.52	.46	.39	
3			-	.49	.48	.45	.42	
4				-	.43	.31	.23	
5					-	.33	.35	
6						-	.39	
7							-	
n = 400								
TA32								
						5	6	7
						-	.22	.35
							-	.18
								-
n = 415								

TABLE 5F
CLUSTER ANALYSIS DENDROGRAPHS

T011 Non-native							T032 Non-native										
		1	2	3	4	7	5	6			1	2	3	4	5	6	7
		C	C	C	C	C	C	C			S	S	S	S	S	S	S
		A	A	A	A	A	A	A			1	2	3	4	5	6	7
		S	S	S	S	S	S	S									
LEVEL	SIZE	1	2	3	4	7	5	6	LEVEL	SIZE	1	2	3	4	5	6	7
		I	I	I	I	I	I	I	67.00	2	I	I	I	I	I	I	I
113.0	2	I	+-+						72.00	2	I	+-+					
119.0	2	I	I	+-+							I	I					
		I	I		I	I			96.00	2	I	I					
138.3	3	I	I		+-+				110.7	4	I	+-+					
		I	I			I	I				I						
159.7	4	I	I			+-+					I		I				
		I	I				I				I		I				
		I	I					I			I		I				
183.3	3	+-+							180.9	5	+-+						
		I					I				I						
		I					I				I						
		I					I				I						
		I					I				I						
250.2	7	+-+	-----						241.5	7	+-+	-----					
						I											I

TA31 Non-native							TB31 Non-native										
		1	5	2	3	4	6	7			1	2	3	4	7	6	5
		C	C	C	C	C	C	C			C	C	C	C	C	C	C
		A	A	A	A	A	A	A			A	A	A	A	A	A	A
		S	S	S	S	S	S	S			S	S	S	S	S	S	S
LEVEL	SIZE	1	5	2	3	4	6	7	LEVEL	SIZE	1	2	3	4	7	6	5
		I	I	I	I	I	I	I	127.0	2	I	I	I	I	I	I	I
142.0	2	I	I	+-+							I	+-+					
		I	I		I	I					I	I					
		I	I		I	I					I	I					
28.0	2	I	I		I	+-+			267.7	3	+-+						
305.5	4	I	I		+-+				290.0	2	I		+-+				
		I	I			I					I		I				
401.5	5	I	I		+-+				392.7	3	I		+-+				
415.0	2	+-+					I				I			I			
		I					I				I			I			
		I					I				I			I			
		I					I				I			I			
		I					I				I			I			
		I					I				I			I			
770.7	7	-	-	-	-	-	-	-	897.7	7	-	-	-	-	-	-	-
						I									I		

Given the fairly featureless correlations for the T0 writing tasks, in particular it is, perhaps, not surprising that there is no clustering for quite some time. This is certainly the case in T011, though they do start clustering slightly earlier in T032. In TA and TB writing tasks the criteria begin to cluster slightly later, but more important the ultimate clustering occurs at a much later stage than T0. This can be seen by looking at the level on the left hand side of the diagrams, which indicate the distance between the two classes that have been clustered. This earlier ultimate clustering in T0 indicates that there is greater similarity in the marks awarded for each criterion by the examiner responsible for assessing this Session and conversely the later ultimate clustering in TA and TB indicates greater dissimilarity between the marks awarded for each criterion in these two Sessions.

In all, this method proved to be very uneconomical in terms of computer time and each of the dendographs took over three hours of real time to run. The results are not very revealing. We might tentatively suggest that they show that criteria 2 and 3 always go together and 6 and 7 normally end up with each other.

The distinction we thought might emerge between the communicative effectiveness criteria, and those more concerned with accuracy, only materialised to a certain extent in tasks T011 and TB31. This was probably due to marker/candidate/item interaction.

A necessary condition for merging the criteria is that most of the candidates achieve the same mark on each of the merged criteria and an alternative way of approaching the data is to analyse the percentage of candidates getting the same mark on different criteria. Table 5G below shows the results of doing this for the four summary writing tasks. This table may be interpreted by reference to an example. The top left hand entry indicates that, for example, 63% of candidates got the same mark in criterion 1 as in criterion 2. A comparison of the totals for each entry shows that candidates were most likely across all four tasks, to get the same mark in criteria 2 and 3 and least likely to do so in criteria 1 and 7.

TABLE 5G

THE PERCENTAGE OF CANDIDATES GETTING THE SAME MARKS ON THE
DIFFERENT CRITERION DESCRIPTORS IN EACH OF THE WRITING TASKS

Criterion	1	2	3	4	5	6	7
1	-	63	55	62	56	58	54
		68 [233]	59 [205]	66 [207]	63 [195]	61 [192]	60 [183]
		50	42	37	40	36	32
2		52	49	42	36	37	37
			70	58	60	49	58
	-		79 [288]	74 [234]	73 [218]	57 [202]	63 [211]
3			69	56	51	52	48
			70	46	34	44	42
				65	66	54	62
4				73 [241]	76 [221]	58 [209]	64 [223]
				58	47	53	50
				45	32	44	47
5					70	64	67
					82 [217]	68 [231]	71 [240]
					40	56	51
6					25	43	52
						61	63
						68 [201]	72 [202]
7						37	39
						35	28
							68
7							74 [245]
							54
							49

KEY T011
T032 [totals]
TA31
TB3†

It is interesting to note that there are far higher occurrences of candidates getting the same mark for different criteria in T0.

Of the criteria in which candidates were most likely to score the same marks, criteria 2 and 3 and criteria 6 and 7 had been highlighted in the previous analysis. This analysis shows, however, that in the majority of cases only slightly over half the candidates got the same mark when two criteria were compared.

As a result of these analyses we decided not to merge any of the writing criteria.

5.2.5.3 Editing Tasks: TA32 and TB32

Originally the intention had been that these tasks would be objectively scored by summing the number of errors left uncorrected or corrected wrongly together with any new errors of commission by the candidate. This figure was to be deducted from a fixed total to give a final mark for the task. Unfortunately, due to lack of clarity and precision in the rubric (v. Appendices 5.4.2 and 5.4.3, p.1020 and p.1036) candidates sometimes changed the actual wording used and re-wrote parts of the passage, rather than simply editing it. This meant that the error count system of marking originally intended was no longer possible.

We had to compromise our marking procedures and use instead criteria 5-7 of the assessment criteria employed for marking the writing tasks T011, T032, TA31 and TB31 (v. Table 5C, p.457 above). For each of these criteria, two errors or less were taken to equal a Level 3 in the criterion, three to four errors a Level 2, five to eight errors a Level 1 and nine or more errors a Level 0.

For the future versions of T.E.A.P. the rubric will be changed so that candidates realise they must not alter the actual wording used.

As the task is to be changed in future versions of the test, less attention was given to the analysis of data relating to these than to other writing tests.

We did, however, correlate the levels gained on TA32 and TB32 and those achieved for the same criteria 5, 6 and 7 on each of the other four writing tests and these are included in Table 5H below.

TABLE 5H
CORRELATION BETWEEN ACCURACY COMPOSITES
(ASSESSMENT CRITERIA 5 + 6 + 7) IN THE VARIOUS TASKS

	<u>T011</u>	<u>T032</u>	<u>TA31</u>	<u>TA32</u>	<u>TB31</u>	<u>TB32</u>
T011	-	.57	.41	.38	.23	.31
T032		-	.41	.45	.44	.39
TA31			-	.53	.41	.32
TA32				-	.28	.20
TB31					-	.31
TB32						-

CONTINGENT NUMBERS

-	330	136	139	109	109
	-	136	139	109	109
		-	443	101	99
			-	101	99
				-	321
					-

The low correlations between the accuracy criteria (5-7) in the marking of the six writing tasks may be due to marking reliability problems, some of which are discussed later. It is noteworthy that the editing tasks correlate no better with each other than they do with the other writing tasks. Indeed, the lowest correlation is that between TA32 and TB32; this is particularly worrying as these were intended as comparable tasks. The most likely explanation is that the marking scheme was ill-fitted for its purpose and difficult to apply consistently; otherwise the possibility must exist that the two editing tasks are not doing the same job.

No further separate analyses were carried out on the marking of the editing tasks.

5.2.5.4 Inter- and Intra-Marker Reliability

Earlier (p.378 et seq.) we discussed the recognised problems of assessing written production in a reliable manner. Because of these problems we decided to conduct, as part of the development work in the project, a small scale mark/re-mark reliability study for three of the extended writing tasks (T011, TA31 and TB31).

In this study, before scripts were sent out for the first marking, we randomly selected thirty scripts from each Session from two representative Centres, one 'A' level and one post-graduate, and photocopied them. Having marked a whole set of scripts from one Session, all three examiners were subsequently asked to second mark each of the three sets of thirty scripts, thus the original marker marked these selected scripts twice.

In Table 5I we list the coefficients of correlation between the total task scores on the three second markings and single original marking. This gives us a picture of the inter- and intra-marker reliability of the examiners for each Session. The statistics are based on an n of approximately thirty for each writing task. The markers are designated M1, M2 and M3.

We breakdown the mark/re-mark data more specifically in Table 5J where we detail, for each of the seven criteria, the coefficients of correlation between each of the second markings and the original rating for that particular criterion.

Caution must be exercised in interpreting the coefficients of correlation in Table 5J due to the low n's involved and the restricted mark range of the 0-3 rating scale employed.

TABLE 5I
 COEFFICIENTS OF CORRELATION BETWEEN THE TOTAL TASK SCORES
 AWARDED BY THE THREE SECOND MARKERS AND THE ORIGINAL MARKERS

<u>TO11</u>	M2	M3	M1*	M1†	
M2	-	.72	.77	.61	
M3		-	.80	.73	Original Marker M1
*M1			-	.83	
†M1				-	
<u>TA31</u>	M2*	M3	M1	M2†	
*M2	-	.63	.85	.77	
M3		-	.71	.73	Original Marker M2
M1			-	.78	
†M2				-	
<u>TB31</u>	M2	M3*	M1	M3†	
M2	-	.68	.68	.55	
*M3		-	.72	.86	Original Marker M3
M1			-	.56	
†M3				-	

† First marking

* Second marking

TABLE 5J COEFFICIENTS OF CORRELATION BETWEEN THE GRADES AWARDED ON EACH CRITERION BY THE THREE SECOND MARKERS AND THE FIRST MARKER

	CRITERION 1			CRITERION 2			CRITERION 3			CRITERION 4			CRITERION 5			CRITERION 6			CRITERION 7									
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4				
<u>TO11</u>	-	.38	.28	.21	-	.49	.37	.46	-	.41	.41	.66	-	.62	.47	.55	-	.74	.77	.58	-	.51	.23	.41	-	.56	.53	.29
M2 1	-	.64	.69		-	.57	.58		-	.24	.32		-	.37	.37		-	.53	.66		-	.52	.46		-	.53	.13	
M3 2	-			.67	-			.58	-			.40	-			.71	-			.56	-			.37	-			.61
*M1 3	-				-				-				-				-				-				-			
†M1 4	-				-				-				-				-				-				-			
<u>TA31</u>	-	.49	.52	.34	-	.30	.60	.67	-	.32	.46	.38	-	.27	.59	.59	-	.45	.51	.67	-	.14	.51	.44	-	.55	.74	.76
*M2 1	-	.36	.64		-	.49	.40		-	.30	.58		-	.47	.62		-	.45	.67		-	.45	.67		-	.54	.40	
M3 2	-			.34	-			.53	-			.65	-			.55	-			.62	-			.56	-			.62
M1 3	-				-				-				-				-				-				-			
†M2 4	-				-				-				-				-				-				-			
<u>TB31</u>	-	.64	.76	.69	-	.53	.60	.47	-	.57	.41	.41	-	.37	.13	.43	-	.51	.52	.26	-	.20	.27	.10	-	.46	.51	.39
M2 1	-	.70	.94		-	.51	.67		-	.51	.53		-	.25	.41		-	.59	.68		-	.41	.81		-	.62	.77	
*M3 2	-			.74	-			.37	-			.31	-			.22	-			.26	-			.36	-			.66
M1 3	-				-				-				-				-				-				-			
†M3 4	-				-				-				-				-				-				-			

† First marking

* Second marking

A better indication of the differences amongst the three sets of remarks and the original marks can be gained from the printout of marks awarded, included in Appendix 5.9, pp.1096-1101.

We have attempted to summarise these data in Table 5K below for each of the writing tasks: T011, TA31 and TB31. We have calculated the aggregate differences for all candidates in the marking awarded for each criterion between the three second markers and the examiner who first marked the scripts. The fourth mark is the original assessment. Marker two was originally responsible for TA31, Marker three for TB31 and Marker one for T011. Criteria 3, 5 and 6 on T0, criteria 1 and 5 on TA and criterion 5 on TB occasion the most disagreement between the three markers. This is even clearer in Table 5L below where we tabulate the total of mark differences on each criterion in each task.

In Table 5M we summarise the differences between the marks awarded by each of the three markers for each criterion on the first and second marking of the scripts from their own particular Session. All three markers were less consistent on criterion 3 (cohesion) and criterion 5 (grammar).

TABLE 5K
TOTAL DIFFERENCE IN MARKS FOR ALL CANDIDATES AWARDED ON EACH CRITERION BY THE FOUR MARKERS

CRITERION 1	TO11				T	TA31				T	TB31				T		
	M2	M3	*M1	†M1		*M2	M3	M1	†M2		M2	*M3	M1	†M3			
M2	0	10	13	20	43	*M2	0	14	28	15	57	M2	0	16	10	14	40
M3	10	0	13	14	37	M3	14	0	30	11	55	*M3	16	0	16	4	36
*M1	13	13	0	13	39	M1	28	30	0	29	87	M1	10	16	0	14	40
†M1	20	14	13	0	47	†M2	15	11	29	0	55	†M3	14	4	14	0	32
M2	0	15	17	18	50	*M2	0	18	13	10	41	M2	0	14	11	15	40
M3	15	0	10	17	42	M3	18	0	11	14	43	*M3	14	0	13	11	38
*M1	17	10	0	15	42	M1	13	11	0	15	39	M1	11	13	0	14	38
†M1	18	17	15	0	50	†M2	10	14	15	0	39	†M3	15	11	14	0	40
M2	0	24	15	9	48	*M2	0	13	13	15	41	M2	0	14	15	17	46
M3	24	0	23	23	70	M3	13	0	14	14	41	*M3	14	0	11	13	38
*M1	15	23	0	14	52	M1	13	14	0	8	35	M1	15	11	0	12	38
†M1	9	23	14	0	46	†M2	15	14	8	0	37	†M3	17	13	12	0	42
M2	0	10	11	10	31	*M2	0	15	8	9	32	M2	0	5	10	5	20
M3	10	0	11	12	33	M3	15	0	13	10	38	*M3	5	0	9	4	18
*M1	11	11	0	7	29	M1	8	13	0	9	30	M1	10	9	0	9	28
†M1	10	12	7	0	29	†M2	9	10	9	0	28	†M3	5	4	9	0	18
M2	0	17	17	13	47	*M2	0	30	16	16	62	M2	0	20	16	26	62
M3	17	0	32	22	71	M3	30	0	40	22	92	*M3	20	0	28	14	62
*M1	17	32	0	14	63	M1	16	40	0	20	76	M1	16	28	0	34	78
†M1	13	22	14	0	49	†M2	16	22	20	0	58	†M3	26	14	34	0	74
M2	0	24	15	11	50	*M2	0	18	9	11	38	M2	0	19	12	24	55
M3	24	0	25	25	74	M3	18	0	23	17	58	*M3	19	0	17	7	43
*M1	15	25	0	14	54	M1	9	23	0	12	44	M1	12	17	0	20	49
†M1	11	25	14	0	50	†M2	11	17	12	0	40	†M3	24	7	20	0	51
M2	0	11	14	18	43	*M2	0	7	4	5	16	M2	0	7	16	9	32
M3	11	0	13	19	43	M3	7	0	7	10	24	*M3	7	0	11	4	22
*M1	14	13	0	10	37	M1	4	7	0	7	18	M1	16	11	0	9	36
†M1	18	19	10	0	47	†M2	5	10	7	0	22	†M3	9	4	9	0	22

n = 26

n = 28

n = 28

† First marking * Second marking

TABLE 5L
TOTAL MARK DIFFERENCES BETWEEN MARKERS
ON EACH CRITERION IN EACH OF THE THREE TASKS

<u>Criterion</u>	<u>T011</u>	<u>TA31</u>	<u>TB31</u>	<u>Overall</u>
1	166	254	148	568
2	184	162	156	502
3	216	154	164	534
4	122	128	84	334
5	230	288	276	794
6	228	180	198	606
7	170	80	112	362

n = 28 n = 28 n = 26

TABLE 5M
DIFFERENCES IN CRITERION GRADES AWARDED WHEN
EXAMINERS MARKED SCRIPTS A SECOND TIME

<u>Criterion</u>	<u>T011</u>	<u>TA31</u>	<u>TB31</u>	<u>Total</u>
1	15	13	4	32
2	10	15	11	36
3	15	14	13	42
4	9	7	4	20
5	16	14	14	44
6	11	14	7	32
7	5	10	4	19

M1 M2 M3

5.2.5.5 Conclusions

The intra-marker data showed a high correlation, but inter-marker reliability was lower. A possible reason for this was the variation of academic and teaching backgrounds of the markers concerned, a choice that had been deliberately made to test the transmissibility of the marking criteria. All three markers had been subjected to the Board's normal standardisation procedures and particular attention had been paid to the use of the analytic criteria for assessing written production. It may be that if a more homogeneous body of examiners were appointed the inter-marker figures would improve. This will be a concern of future research in connection with T.E.A.P.

If the anticipated improvement does not materialise, we feel that serious consideration should be given to the double marking of essays in T.E.A.P., although it is recognised that to do so has administrative and financial implications for the examining body.

5.3 DATA ANALYSIS AT THE TASK LEVEL (INTERNAL)

5.3.1 Introduction

Our first task was to establish that native speakers could cope with the tasks that were set in each of the three Sessions. We examine in Table 5N below the differences between the mean scores and the standard deviations of the N.S. and N.N.S. groups for each task in each of the three Sessions.

It is evident, even without tests of significance, that the N.S. group, whilst not all scoring 100%, coped very well with all tasks and their overall performance as a group was noticeably different from that of the N.N.S. group. With the exception of TB32 the N.S. group mean was over 75% for each task. The N.S. means were higher for each task and the standard deviations lower than those of the N.N.S. group.

The information is further broken down in Appendix 5.10, pages 1102-1114 by level: Advanced (A), Undergraduate (U) and Post-graduate (P) and by subject area: Science, Engineering and A.S.A.B.S., for those students on whom we had complete background information. A similar pattern of difference between native and non-native speaker performance is evident there also.

It is to be expected that the range of scores of the N.S. group in a test designed for second language users would be narrow and that most of the items would prove fairly easy for this group. The earlier trialling had seen the elimination of most of the items that had proven difficult for the N.S. group as a whole, since such items may be suspect. No assumptions were made of perfect or near perfect performance by the N.S. group, but it was expected that there would be a substantial superiority as compared to the N.N.S. group in their control of the skills being tested (v. Davies 1965).

TABLE 5N
COMPARISON OF N.S. AND N.N.S. PERFORMANCE ON EACH OF THE
TASKS IN THE PRE-TEST VERSIONS OF T.E.A.P.

Task Code	Max. Poss. Mark	NUMBER OF CANDIDATES		OVERALL MEAN				OVERALL S.D.		CRONBACH α
		N.S.	N.N.S.	N.S.	%	N.N.S.	%	N.S.	N.N.S.	N.N.S.
READING										
TO12	23	125	330	18.55	81	12.59	55	3.82	5.88	0.77
TA11	17	71	438	13.11	77	9.04	53	2.89	3.37	0.55
TA12A	21	72	444	18.46	88	10.72	51	2.30	5.37	0.88
TA12B	21	72	439	16.33	78	8.42	40	3.42	5.04	0.88
TA13	18	76	435	13.65	76	10.27	57	2.42	3.54	0.69
TB11	17	124	321	13.85	81	11.80	69	2.08	3.28	0.61
TB12A	21	125	325	17.83	85	12.39	59	4.36	4.97	0.86
TB12B	21	125	326	17.45	83	10.16	48	2.87	5.01	0.86
TB13	19	125	325	14.43	76	8.56	45	3.15	4.08	0.68
WRITING										
TO11	21	125	330	18.22	87	13.97	67	2.07	5.24	0.94
TO32	21	125	330	17.74	84	12.29	59	2.96	6.00	0.97
TA31	21	76	443	17.89	85	12.89	61	2.37	5.54	0.92
TA32	9	75	446	7.90	88	5.83	65	1.02	2.18	0.73
TB31	21	124	324	16.53	79	12.65	60	2.97	5.25	0.87
TB32	9	124	322	5.89	65	4.48	50	1.94	2.08	0.66
LISTENING										
TO21	15	125	328	13.31	89	7.02	47	1.26	4.18	0.87
TO31	18	125	328	14.42	80	6.31	35	2.67	4.09	0.78
TA21	20	76	446	16.95	85	11.53	58	2.84	4.70	0.74
TB21	13	125	325	10.28	79	4.33	33	2.32	3.27	0.71
GRAMMATICAL KNOWLEDGE										
TO41	60	133	333	57.11	95	41.79	70	2.96	10.63	0.92

The listening tests in general evidence the biggest gaps between N.S. and N.N.S. groups of any of the study modes.

It is clear that the N.S. performance in all the tests was superior to that of the N.N.S. group and that there was an obvious difference in proficiency between the two groups within the limits of the samples tested. Because of the magnitude of the differences in means we did not feel it necessary to carry out any tests of significance.

Reliability coefficients, based on the N.N.S. raw scores, were also calculated for each of the test tasks using Cronbach's Alpha (v. Anastasi 1982, p.117). The coefficients for the N.N.S. group are to be found in the final column of Table 5N above. The coefficients for the N.S. group have been omitted because the high facility indices on all tasks for this group mean there is little variance in performance.

It should be pointed out that the Cronbach Alpha's for the writing tasks must be interpreted in a different way from the others. They are not measures of internal consistency between individual items in a task, but measures of the internal consistency between different parts of the marking scheme used to assess performance on a single item. It is noteworthy that for the four summary writing tasks the α 's are very high indicating that candidates were scoring fairly consistently across the seven criteria, although not consistently enough to enable us to merge any of these (v. p.465 above). The α 's for the two editing tasks were amongst the lowest. This may be in line with expectations when one bears in mind the fact that only three criteria were involved and, furthermore, as was suggested above (p.465) these criteria may have been unsuitable for marking these tasks.

Cronbach Alpha is a measure of internal consistency and we would expect a lower value for tests comprising items which individually tested different abilities or ability composites, one item from another, than for tests where all items were testing the same abilities or ability composites. This view is supported by the

evidence in Table 5N. The highest α 's are associated with tasks of a comparatively homogeneous nature, such as T021 (dictation) or TA12A or TA12B (gap-filling) and the lowest α 's with tasks where each item was devised to test one of a range of abilities such as TA11 (multiple choice questions on reading comprehension) and TA13 (short answer questions on reading comprehension). Thus Cronbach α is not a measure of the suitability of a task, but rather says something about its homogeneity.

5.3.2 Approaching the Test Data

The differences between the mean scores of the N.S. and N.N.S. groups detailed in Table 5N, page 475 are sufficient to warrant us considering them as separate groups for the purpose of analysis of test scores. Given the practical limitations that would be placed on the analysis, we wished to examine whether we should keep the further sub-divisions of the N.S. and N.N.S. groups into 'A' level, undergraduate and post-graduate; into the disciplines of science, engineering and arts, social, administrative and business studies. An analysis of the performance of eighteen sub-groups would be fairly demanding in terms of resources. Particularly since it was not envisaged that separate versions of the test would be devised for separate educational levels.

An examination of the mean scores of the post-graduate groups showed them, in most cases, to be lower than those of undergraduate groups in the same broad discipline areas. This reflects the fact that many of the graduate students in our sample were on pre-sessional courses because of problems with their English language proficiency. The differences between the other two levels were not so obvious, so we carried out a series of t-tests detailed in Table 50 below, to establish whether we needed to treat them as separate groups. Details of the means and standard deviations of these separate groups are given in Appendix 5.10, pages 1102-1114.

We are aware of the problems of using t-tests on these samples, as we are probably violating assumptions of both normality and equal variance. Given the limited sizes of the groups, departure from these assumptions was inevitable. To go some way to meeting this criterion, a significance level of 1% was selected as a conservative but sensible estimate.

We carried out t-tests on both N.N.S. and N.S. group mean scores in each of the three discipline areas, where there were sufficient numbers to warrant this.

It was important to establish whether the test was likely to be equally suitable for students at the three educational levels. The data in Table 50 broadly suggests that it might be. Most of the statistically significant differences between the means occur where the two groups are small. The exceptions are amongst the A.S.A.B.S. N.N.S. group in Sessions T0 and TA. It was noticeable that there were significant differences at 1% between the 'A' level and undergraduate students in the N.S. group. These occurred in Tasks T032, T041, TB11 and TB13. Indeed, in all cases, the means for the undergraduate group were higher than those for the 'A' level group: a result that might be expected in view of the greater maturity and presumably higher general level of education of the former group. It is difficult to be sure why these differences were greater for the four tasks cited above; the textual material could be partly responsible in one or two instances.

In the construction of Session TB, texts were selected from the disciplines of science and engineering to produce a version of the test which would have greater face validity for students in these areas. The research, reported earlier in Chapter 3, had indicated that the activities and performance constraints in the various study modes were reasonably similar for students in these disciplines, at the different levels. We carried out a series of t-tests on the mean scores of the various N.S. and N.N.S. Science and Engineering groups in each of the tasks in Sessions T0, TA and TB where there were sufficient numbers to warrant this. These are shown in Table 5P below.

TABLE 50

THE RESULTS OF t-TESTS ON 'A' LEVEL AND UNDERGRADUATE STUDENTS' MEAN SCORES BY TASK AND SESSION

	<u>SESSION TO</u>				<u>SESSION TA</u>				<u>SESSION TB</u>			
	<u>Non-native Speakers</u>		<u>Native Speakers</u>		<u>Non-native Speakers</u>		<u>Native Speakers</u>		<u>Non-native Speakers</u>		<u>Native Speakers</u>	
	A/U	A/U	A/U	A/U	A/U	A/U	A/U	A/U	A/U	A/U	A/U	A/U
	Sci	Eng	ASABS	Sci	Eng	ASABS	Sci	Eng	ASABS	Sci	Eng	ASABS
	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%
T011	NS NS	NS ✓	✓ ✓	NS ✓	NS NS	✓ ✓	NS NS	NS NS	NS NS	NS NS	NS NS	NS NS
T012	NS NS	✓ ✓	✓ ✓	NS ✓	NS NS	✓ ✓	NS NS	NS NS	NS NS	NS NS	NS NS	NS NS
T021	NS NS	✓ ✓	NS NS	NS NS	NS NS	✓ ✓	NS NS	NS NS	NS NS	NS NS	NS NS	NS NS
T031	NS NS	✓ ✓	✓ ✓	NS NS	NS NS	✓ ✓	NS NS	NS NS	NS NS	NS NS	NS NS	NS NS
T032	NS NS	NS NS	✓ ✓	NS ✓	NS NS	✓ ✓	NS NS	NS NS	NS NS	NS NS	NS NS	NS NS
T041	NS NS	✓ ✓	✓ ✓	NS ✓	NS NS	NS ✓	NS NS	NS NS	NS NS	NS NS	NS NS	NS NS
	n =	n =	n =	n =	n =	n =	n =	n =	n =	n =	n =	n =
	50/8	10/6	21/29	35/35	46/6	14/16	73/31	58/9	4/4	52/6	52/51	52/51

KEY: NS: Non-significant
 ✓: Significant
 1%: Significant at the 1% level
 5%: Significant at the 5% level
 A: GCE Advanced level
 U: Undergraduate level
 n: number of students

TABLE 5P

THE RESULTS OF t-TESTS ON SCIENCE AND ENGINEERING STUDENTS' MEAN SCORES BY TASK AND SESSION

	<u>SESSION TO</u>				<u>SESSION TA</u>				<u>SESSION TB</u>				
	<u>Non-native Speakers</u>		<u>Native Speakers</u>		<u>Non-native Speakers</u>		<u>Non-native Speakers</u>		<u>Non-native Speakers</u>		<u>Non-native Speakers</u>		
	S/E	S/E	S/E	S/E	S/E	S/E	S/E	S/E	S/E	S/E	S/E	S/E	
	A	U	P	U	A	U	P	A	U	P	A	U	P
	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%	1% 5%
T011	✓	NS NS	NS NS	NS NS	TA11	NS NS	NS NS	NS NS	TB11	NS NS	NS ✓	NS ✓	NS ✓
T012	✓	NS NS	NS NS	NS NS	TA12A	✓	NS NS	NS NS	TB12A	NS NS	NS NS	NS NS	NS NS
T021	✓	NS NS	NS NS	NS NS	TA12B	✓	NS NS	NS NS	TB12B	NS NS	NS NS	NS NS	NS NS
T031	✓	NS NS	NS NS	NS NS	TA13	NS NS	NS NS	NS NS	TB13	NS NS	NS NS	NS NS	NS NS
T032	NS ✓	NS NS	NS NS	NS NS	TA21	NS NS	NS NS	NS NS	TB21	NS NS	NS NS	NS NS	NS NS
T041	✓	NS NS	NS NS	NS NS	TA31	✓	NS NS	NS NS	TB31	NS ✓	NS NS	NS NS	NS NS
	n =	n =	n =	n =	TA32	NS NS	NS ✓	NS NS	TB32	NS NS	NS NS	NS NS	NS NS
	50/10*	8/6	90/13	35/29		n =	n =	n =		n =	n =	n =	n =
						46/14*	6/16	66/23		58/4	9/4	98/25	

* extremely poor group from one Centre only

KEY: S: Science students
 E: Engineering students
 n: number of students
 A: GCE Advanced level
 U: Undergraduate level
 P: Post-graduate level

NS: Non-significant
 ✓: Significant
 1%: Significant at 1% level
 5%: Significant at 5% level

With the exception of differences occurring due to an extremely poor 'A' level engineering group, the overall impression is that there are few significant differences in the performance of the two groups. Thus, for the purpose of this analysis, we took a decision to collapse the two into a composite Science and Engineering group. We were further encouraged in this direction by the small size of the individual groups in these discipline areas. Thus, in the analysis which follows, they are treated as one group, though we have preserved the separate identities of the 'A' level, undergraduate and post-graduate groups within the science and engineering composite. The means and standard deviations by task and Session for this combined Science and Engineering group are located in Appendix 5.10, pages 1109-1114.

5.3.3 Relationship Between Tasks

In Section 5.2 above we analysed the various Sessions of the test at the item level. In this section we wish to examine whether there is any redundancy amongst the tasks which comprise the test battery. One way of investigating this matter was to study the inter-task correlations.

5.3.3.1 Correlational Data

In Table 5Q below we list again the tasks to be found in the three Sessions of the test: T0, TA and TB, grouped according to the construct the task focuses on. In all, there were twenty tasks. Tasks 1-9 constitute the reading tasks, 10-15 the writing tasks, 16-19 the listening tasks and 20 the test of grammatical competence.

TABLE 5Q

INTER TASK CORRELATIONS

Task Code	20 tasks	4 amalgamations	8 amalgamations	10 amalgamations
TO12	1	1	1	1
TA11	2	2	2	
TA12A	3	3	3	2
TA12B	4			
TA13	5	4		
TB11	6	5	4	3
TB12A	7	6	5	
TB12B	8			
TB13	9	7		
TO11	10	8	6	4
TO32	11			
TA31	12	9	7	5
TA32	13			
TB31	14	11	8	6
TB32	15	12		
TO21	16	13	9	7
TO31	17			
TA21	18	14	10	8
TB21	19	15	11	9
TO41	20	16	12	10

KEY: R: Reading
W: Writing
L: Listening
G: Grammar

All test task correlations are to be found in Appendix 5.11, pages 1115-1135. In Appendix 5.11.1, pages 1118-1126, we list the correlations between tasks for the series of 4, 8 and 10 amalgamations described in Table 5Q above. In these amalgamations of tasks we gradually increased the size of the skill composites and reduced the number of separate task components within individual Sessions until, with 10 amalgamations, there was only a reading, listening and writing composite for each Session plus the separate grammar task. This allowed us to see the possible effects of excising certain tasks from the battery.

In Appendix 5.11.2 we focused on the various groups of students who took the combinations of Sessions: TO + TA, TO + TB and TA + TB. For the purposes of analysis we treated separately the reading, listening and writing tasks in each of the various combinations.

On pages 1127-1129, we have summarised all the available data on the relationship between the reading tasks for students taking the different combinations of Sessions TO + TA, TO + TB and TA + TB.

In Parts A and B of these tables we detail the correlations for the 'A' level and post-graduate groups in the Sci./Eng. and A.S.A.B.S. groups. No correlations are included for undergraduates as the numbers were too small.

In Part C we present the correlations for all the students in each of the two subject area groupings, Sci./Eng. and A.S.A.B.S.

In Part D, we list the correlations between tasks, for all the non-native speakers regardless of discipline, taking all the components in a particular combination of Sessions, including N.N.S. students not separately identified by level or discipline. Inspection of this part of the table should indicate whether any tasks within constructs are redundant.

In Part E we have amalgamated all the reading scores for each candidate within a Session and calculated an overall correlation between Sessions for that particular construct.

A similar analysis was carried out on the listening and writing tasks for the various combinations of Sessions. The correlations for the listening and writing tasks are laid out in the same manner as those for the reading tasks and are to be found on pages 1130-1135.

5.3.3.1.1 Reading

In the TO/TA combination TA11, the multiple choice format for testing reading comprehension, appears to correlate the least with the other tasks for all the sub-groups of candidates. The highest correlations are between the two sets of marks for the two parts; TA12A/TA12B of the gap-filling format. It may be worth considering whether we could reduce the task by showing where the omissions occur and having students supply only the words that are omitted. In the TO/TB combination we find a similar picture with TB11 correlating the lowest with other components in all the sub-groups of candidates. The two parts of the gap-filling component again correlate highly.

In the TA/TB correlation the multiple choice formats TA11 and TB11, in general, have low correlations with other components. The low correlations of these multiple choice tasks may be due to format effect, though the effect does not appear to be mirrored in the case of T041 (V. Table 5ZZZ, p.1118). Neither do the two multiple choice reading comprehension tasks correlate with each other. The earlier item analysis revealed certain inadequacies in these tasks but not in our opinion sufficient to make them that different from the short answer questions which similarly contain a restricted number of items.

With regard to the inter-session correlations in Part E one might expect the correlation between the reading tasks in TO and TA to be somewhat similar to that between those in TO and TB, since both TA and TB are intended as parallel sessions. For the same reason one might expect the TA/TB correlation to be high. These expectations

must be tempered by the number and length of tasks in each of the Sessions and by the marking reliability of the tasks. It is not therefore surprising that in reading the TA/TB correlation is higher than the TO/TA or TO/TB ones; it contains more reading comprehension tasks.

5.3.3.1.2 Writing

The inter-task correlations for this construct are all fairly low and the highest correlation is to be found between the two writing tasks T011 and T032 in the same Session where the examiner is common to both tasks. As well as the possibility of differential student performance on different tasks, the generally recognised difficulty of marking reliably productive writing referred to above (v. Section 4.5.3) leads to greater measurement error which will serve to depress the correlation coefficients between the writing tasks.

The editing tasks TA32 and TB32 in general correlated poorly with other writing tasks; perhaps limited evidence that they are measuring a different element of what we are labelling writing ability and should therefore remain in the battery.

All three inter-session correlations for this construct are lower than the corresponding ones for reading comprehension probably due, at least in part, to the greater degree of measurement error involved in assessing writing. The poorer correlations between TB and the other Sessions in writing may be due to the difficulties noted earlier in respect of TB13 and TB21. As these are integrated with the writing task TB31 then these failings might have had some effect on performance in this latter task also. The very difficult nature of TB32 as against TA32 may also be contributing to the lower correlations achieved by the TB writing construct. This is further evidence of the difficulty there is in constructing parallel tests especially as regards task dimensions.

5.3.3.1.3 Listening

The listening construct shows inter-task correlations somewhat between those for reading comprehension and for writing. The highest correlations were achieved between the two listening tasks T012, the

dictation, and T031, the listening to a lecture and note-taking exercise, presumably because both are ability-related and capable of being marked reliably, and the possibility of differential performance due to a practice effect from Session to Session is absent. The values of the inter-session correlations for listening are also between those for reading comprehension and writing with that of TA/TB lowest of the three. It should be borne in mind, however, that this latter finding may be due to the fact that only two listening tasks occurred in these Sessions; the TO/TA and TO/TB correlations involved three tasks.

5.3.3.1.4 Linguistic competence

The correlations of the grammar sub-test T041 with other parts of the test can be found in Appendix 5.11.1, pages 1118-1125. They show that T041 achieves correlations of between 0.3 and 0.7 with all other tasks, perhaps indicative that linguistic competence is a part of what these other tasks are measuring. This can be seen most clearly in Table 5CCCC, page 1124. When we amalgamated the reading, listening and writing test scores into three composite skill scores for each of the Sessions TO, TA and TB and correlated these with T041, slightly higher correlations resulted.

5.3.4 Factor Analysis Data

It was felt that factor analysis techniques might provide a further perspective on the relationships between test tasks emerging from the correlational data.

Using the Honeywell package STATPAC (Honeywell 1981) principal components were calculated from the various correlation matrices (v. pp.1136 - 1172) for pairs of Session and individual Sessions. Characteristically the principal component analysis (P.C.A.) revealed the existence of one 'general ability' factor, though in no case did this component account for more than two thirds of the variance. In order to clarify the data it was decided to employ the Varimax rotation technique. This technique reorganises the data from the principal component analysis giving a new set of factors which may

serve to divide the test tasks into more easily interpreted groupings.

Three sets of data were involved in this analysis and we list these below. Set 2 contains all the candidates in Set 1 and Set 3 contains all those in Set 2.

Set 1. Sci./Eng. and A.S.A.B.S. Candidates Taking Various Combinations of Session (v. Appendix 5.12.1, pp.1137-1155)

TO + TA	Sci./Eng. candidates	n = 80
TO + TA	A.S.A.B.S. candidates	n = 40
TO + TB	Sci./Eng. candidates	n = 76
TO + TB	A.S.A.B.S. candidates	n = 21
TA + TB	Sci./Eng. candidates	n = 43
TA + TB	A.S.A.B.S. candidates	n = 48

Set 2. All N.N.S. Taking Various Combinations of Session (v. Appendix 5.12.2, pp.1156-1165)

TO + TA		n = 132
TO + TB	All N.N.S.	n = 103
TA + TB		n = 95

Set 3. All N.N.S. Taking Single Sessions (v. Appendix 5.12.3, pp.1166-1172)

TO		n = 328
TA	All N.N.S.	n = 435
TB		n = 321

The general procedure followed was to select those components which had eigenvalues greater than one after the P.C.A. and then rotate these components through Varimax procedure (v. Alderson 1978a). In the TO+TA, TA+TB combinations of Session for both Sci./Eng. and A.S.A.B.S. candidates in Set 1, this meant that three components were rotated. In the case of TO+TB for both Sci./Eng. and A.S.A.B.S. groups this criterion would have resulted in fewer components being rotated. For reasons of consistency and to allow comparison between combinations, it was decided to rotate three factors on this combination of Session also. The printouts of these principal

component analyses and accompanying rotations are included as Appendix 5.12.1, pages 1137-1155. A similar pattern was adopted in Set 2 for the analyses carried out on the correlation matrices obtained for all N.N.S. candidates, irrespective of discipline, on these three combinations of Session (v. Appendix 5.12.2, pp.1156-1165).

In the combinations of Session taken by various groupings of candidates the samples were limited whereas in individual Sessions the number of candidates was far greater. We therefore decided to perform these analyses on the T0, TA and TB single Sessions and the printouts for these can be found in Appendix 5.12.3, pages 1166-1172. In all cases in the single Sessions in Set 3 there was only one component with an eigenvalue greater than one after the principal component analysis. This single component never accounted for more than two thirds of the variance. In Session T0 only two components had eigenvalues greater than 0.5 after the P.C.A. and therefore rotation of more than two components was not considered worthwhile. For the purposes of consistency and comparison we decided therefore to rotate only two components in each of the single Sessions.

To make the data contained in Appendix 5.12 more accessible we have abstracted those cases where the factor loadings after the Varimax rotation are greater than 0.6 and detail them in chart form below for the three sets of data. We feel some of the results may well have been affected by the low n's and the natures of the samples involved.

Set 1. Sci./Eng. and A.S.A.B.S. Candidates Taking Various
Combinations of Session
T0 + TA (v. Appendix 5.12.1, pp.1138-1143)

In Table 5R below the results of the Varimax procedure are summarised for this combination of Sessions. For both groups of students the first factor would appear to be a comprehension factor embracing reading and listening tests and the second a writing factor. The loading of TAll on a third factor in both groups is difficult to explain with no other task loading very heavily on this. The same does not appear to be true for TB11 in the other combinations which

TABLE 5R

SET 1 FACTOR LOADINGS GREATER THAN 0.6 AFTER VARIMAX ROTATION.
A.S.A.B.S. AND SCI./ENG. STUDENTS ON COMBINATIONS OF SESSION.

	TO/TA			TO/TB			TA/TB					
	A.S.A.B.S.			A.S.A.B.S.			A.S.A.B.S.					
	Sci/Eng	1	2	3	Sci/Eng	1	2	3	Sci/Eng	1	2	3
TO12 1	*			TO12 1	*				TA11 1			
TA11 2	*			TB11 2	*				TA12A 2	*		
R TA12A 3	*	*		R TB12A 3	*	*			TA12B 3	*	*	
TA12B 4	*	*		TB12B 4	*	*			TA13 4	*	*	
TA13 5				TB13 5	*			R	TB11 5	*	*	
TO11 6	*			TO11 6	*				TB12A 6	*	*	
TO32 7	*	*		TO32 7	*	*			TB12B 7	*	*	
W TA31 8	*	*		W TB31 8	*	*			TB13 8	*	*	
TA32 9	*			TB32 9	*	*			TA31 9	*		
TO21 10	*			TO21 10	*				TA32 10	*	*	
L TO31 11	*	*		L TO31 11	*	*		W	TB31 11	*	*	*
TA21 12	*	*		TB21 12	*	*			TB32 12	*	*	*
G TO41 13	*			G TO41 13	*			L	TA21 13	*	*	
									TB21 14	*	*	

* = factor loading > 0.6

v. Appendix 5.12.1, pp.1138-1155

in general loads on the same factor as other reading comprehension tasks.

That T011 and T012 load heavily on the same factor is due to them both having in common the same reading passage as stimulus material. This effect is repeated in the T0/TB combination for the A.S.A.B.S. group. The loading of TA12B in the Sci./Eng. data on what we have tentatively labelled a writing factor may be due to the productive nature of the second part of this integrative gap-filling measure.

TO + TB (v. Appendix 5.12.1, pp.1144-1149)

For the Sci./Eng. group the first factor appears to be a comprehension factor. The integrated nature of tasks TB13, TB21 and TB31 appears to be contributing to the similarity of the loadings. The writing task TB31 is based on the stimulus material presented in TB13 and TB21. The second factor shows the dictation task T021, the listening task T032 and the test of grammar T041 both loading heavily on the same factor which might be a further comprehension factor. The third factor appears to be a writing factor.

For the A.S.A.B.S. group, the first factor also appears to be a comprehension factor. That T011 loads on this factor is probably due to the fact that its successful completion is also dependent on comprehension of the reading passage that task T012 is based on. As we noted above, this could also explain the similar occurrence in the T0/TA combination of Sessions. As for the Sci./Eng. group, T021 and T041 load together as a second factor though for this group TB32 does also. This may well mean that it is another comprehension factor, given that the nature of TB32 is editing rather than productive writing. The third factor again appears to be a writing factor.

The differences in the factor loadings in T0 + TA as compared with T0 + TB might be taken as evidence that these combinations are not identical in what they are testing.

TA + TB (v. Appendix 5.12.1, pp.1150-1155)

In TA/TB a similar picture appears after the Varimax rotation (v. Table 5R above). The first factor for both groups appears to be a

comprehension factor. The heavy loadings in tasks 2, 3, 8 and 13 are common to both groups. The loading of TA32 on this factor in both groups is perhaps due to the receptive editing as against productive writing focus of the task. The third factor for the Sci./Eng. group also appears to be a comprehension factor. The second factor for both groups would seem to be a writing factor as does factor three for the A.S.A.B.S. group.

Set 2. All N.N.S. Taking Various Combinations of Session (v. Appendix 5.12.2, pp. 1156-1165)

Similar factor analyses were carried out on the combinations of Session using the data from all the N.N.S. students considered as a single group. The results are presented in summary form in Table 5S below where we have tabulated those cases where the factor loadings are greater than 0.6. Again the different factor loadings in TO + TA as compared with TO + TB might suggest these combinations are not parallel tests.

In TO/TA the first factor would appear to be a comprehension factor with strong loadings on both reading and listening tasks. The second appears to be a writing factor. The link between T011 and T012 arising out of common stimulus material is again in evidence. The third factor which TAll loads heavily on is not easily explained.

In the TA/TB combination the first factor would again seem to be a comprehension factor with most of the reading and listening tests loading heavily on this. The second factor might tentatively be described as a writing factor. TB11 loads on the same factor as TB31 which again is not easily susceptible to explanation. In the TO/TB combination the pattern is not so clear. Factor one might be seen as a comprehension factor and factor three as a writing factor. Factor two is difficult to interpret with writing, reading and listening tasks loading heavily on it.

TABLE 5S
 SET 2 FACTOR LOADINGS GREATER THAN 0.6 AFTER VARIMAX ROTATION
 ALL N.N.S. CANDIDATES ON COMBINATIONS OF SESSION

	TO/TA			TO/TB			TA/TB		
	1	2	3	1	2	3	1	2	3
TO12 1	*			TO12 1	*		TA11 1		
TA11 2		*		TB11 2	*		TA12A 2	*	
R TA12A 3	*			R TB12A 3			TA12B 3	*	
TA12B 4	*			TB12B 4			R TA13 4		
TA13 5				TB13 5	*		TB11 5		*
TO11 6	*			TO11 6		*	TB12A 6	*	
TO32 7	*			TO32 7		*	TB12B 7	*	
W TA31 8	*			W TB31 8		*	TB13 8	*	
TA32 9	*			TB32 9		*	TA31 9		
TO21 10	*			TO21 10		*	TA32 10		
L TO31 11	*			L TO31 11			W TB31 11		*
TA21 12	*			TB21 12	*		TB32 12		*
G TO41 13	*			G TO41 13		*	TA21 13	*	
							L TB21 14	*	

* = factor loading > 0.6

v. Appendix 5.12.2, pp.1156-1165

Set 3. All N.N.S. Taking Single Sessions (v. Appendix 5.12.3, pp.1166-1172)

Given that the size of the samples was greater in the separate Sessions T0, TA and TB, we decided to carry out additional principal component analyses on the tasks within each of these Sessions to be followed by Varimax rotation. The full printouts of the analyses are to be found in Appendix 5.12.3, pages 1166-1172. We detail the results of the Varimax rotations below for those cases where the factor loadings are greater than 0.6. In this instance we have only taken the first two components as for the remaining cases, where eigenvalues fell below 0.5, it was not thought worthwhile to include such components in the Varimax rotation. The results illustrate clearly the presence of two main factors which we might tentatively label comprehension and written production.

TABLE 5T
FACTOR LOADINGS GREATER THAN 0.6 AFTER VARIMAX
ROTATION, ALL N.N.S. ON SINGLE SESSIONS

	T0		TA		TB	
	<u>1</u>	<u>2</u>	<u>1</u>	<u>2</u>	<u>1</u>	<u>2</u>
R T012	*		R TA11		R TB11	*
W T011	*		R TA12A	*	R TB12A	*
W T032	*		R TA12B	*	R TB12B	*
L T031	*		R TA13	*	R TB13	*
L T031	*		W TA31	*	W TB13	*
G T041	*		W TA32	*	W TB32	*
			L TA21	*	L TB21	*

In general the reading and listening tasks seem to load on the same factor across Sessions and the writing tasks on a different factor. TA11 is still behaving quite oddly though. There is also quite a heavy loading in T012 on what might be described as a writing factor (factor two) and, as we noted above, this is probably due to T011 and T012 having the same reading text as stimulus. There is no comparable evidence from TA or TB that the reading and/or listening tests were

heavily influencing the writing tasks TA31 or TB31. The pattern of loadings in TA and TB are similar in these data, which was expected given our attempts to set up parallel tests in the two Sessions.

5.3.4.1 Conclusions

In general we might conclude that our tests do not seem to respond to these factor analytical techniques. This could arise from the integrated nature of certain of the tasks or the error variance associated with imperfect task reliabilities. The interpretation of the factor structures is further complicated by the sizes and the natures of the samples involved.

We might tentatively suggest that two factors seem to emerge from the data, namely comprehension and written production.

5.4 DATA ANALYSIS AT THE TASK LEVEL (EXTERNAL)

5.4.1 Introduction

On page 418 we described how we proposed to validate T.E.A.P. scores against a variety of external criteria. Amongst these were tutors' estimates of the proficiency in English of students taking the pre-test and self-assessments by the students of their own ability in English. The findings of this part of the validation exercise are shown below.

5.4.2 Tutors' Estimates (Global and Communicative) of Students' Language Proficiency

5.4.2.1 Background

Traditionally (cf. Davies 1965 and Moller 1982), testing specialists have only asked teachers to make 'global' assessments of students' ability either in aggregate or separately for listening, reading, writing and speaking. However, given that we were experimenting in our pre-tests with certain tasks which might claim to have some 'communicative relevance' to the types of activities students encounter in English medium study, we felt it necessary to ask teachers to assess their students and students to assess themselves more specifically in terms of proficiency in these activities; two tutor rating forms were administered. The first, included as Appendix 5.5.2, pages 1050-1051 was derived from the categories established during our earlier investigation of students' language activities and associated difficulties described in Chapter 3 above. The three sections on reading, listening and writing were broken down into specific parts of the overall construct for rating purposes. Three composite scores of the estimates of communicative proficiency were then calculated, one for each skill, and used for correlational purposes in the analysis below.

The second tutor rating form used for recording the tutors' global estimates derives largely from the work of Hawkey (1982) and Moller (1982); in it teachers were asked to produce a global estimate of a student's ability in reading, listening and writing on a four point scale (v. Appendix 5.5.3, p.1053).

A category for 'don't knows' was included in both versions and if this occurred the teacher's rating for that particular macro-skill was not included in the analysis.

In Tables 5U and 5V below, we list the Pearson product moment correlation coefficients between task scores in the various Sessions and the following tutors' estimates:

1. TCR: Teachers' communicative assessments of students' reading ability.
2. TCL: Teachers' communicative assessments of students' listening ability.
3. TCW: Teachers' communicative assessments of students' writing ability.
4. TGR: Teachers' global assessments of students' reading ability.
5. TGL: Teachers' global assessments of students' listening ability.
6. TGW: Teachers' global assessments of students' writing ability.

These are shown for the whole N.N.S. group for whom we have data as well as for five Centres for which estimates for more than twenty students were available. These Centre data are indicated C3, C11, C14, C18 and C20 respectively.

TABLE 5U CORRELATIONS BETWEEN TASK SCORES AND TEACHERS' COMMUNICATIVE ASSESSMENTS OF STUDENTS' READING, LISTENING AND WRITING ABILITIES

TASK CODES	<u>TCR</u>					<u>TCL</u>					<u>TCW</u>							
	ALL NNS	C3	C11	C14	C18	C20	ALL NNS	C3	C11	C14	C18	C20	ALL NNS	C3	C11	C14	C18	C20
<u>Reading</u>																		
TO12	.58	.67			.46													
TA11	.30	.25	.43	.39	.08	.42												
TA12A	.57	.69	.59	.54	.49	.61												
TA12B	.60	.74	.70	.62	.48	.56												
TA13	.52	.59	.60	.44	.48	.53												
TB11	.36	.41		.01	.39	.56												
TB12A	.49	.67		.27	.43	.55												
TB12B	.49	.45		.30	.47	.61												
TB13	.38	.28		.34	.33	.30												
<u>Listening</u>																		
TO21							.53	.57			.45							
TO31							.52	.59			.41							
TA21							.54	.46	.69	.68	.44							
TB21							.46	.24		.54	.19							
<u>Writing</u>																		
TO11													.39	.32			.41	
TO32													.39	.56			.33	
TA31													.41	.20	.63		.35	.27
TA32													.44	.44	.58		.38	.29
TB31													.38	.38			.40	.33
TB32													.21	.26			.27	.06

For Contingent Numbers see Appendix 5.13.1, page 1176

TABLE 5V CORRELATIONS BETWEEN TASK SCORES AND TEACHERS' GLOBAL ESTIMATES
OF STUDENTS' READING, LISTENING AND WRITING ABILITIES

TASK CODES	<u>TGR</u>					<u>TGL</u>					<u>TGW</u>				
	ALL NNS	C3	C11	C14	C18 C20	ALL NNS	C3	C11	C14	C18 C20	ALL NNS	C3	C11	C14	C18 C20
<u>Reading</u>															
TO12	.60	.67													
TA11	.34	.12	.44	.41	.36										
TA12A	.51	.57	.76	.47	.39										
TA12B	.50	.53	.80	.53	.34										
TA13	.43	.49	.69	.53	.23										
TB11	.30	.52		.18	.34										
TB12A	.45	.42		.42	.38										
TB12B	.44	.31		.43	.43										
TB13	.33	.14		.33	.28										
<u>Listening</u>															
TO21						.56	.54								
TO31						.51	.49								
TA21						.46	.32	.76	.66	.31					
TB21						.41	.15		.52	.25					
<u>Writing</u>															
TO11											.25	.22			
TO32											.59	.61			
TA31											.31	.19	.53	.43	.16
TA32											.31	.32	.69	.31	.22
TB31											.24	.38		.21	.01
TB32											.36	.07		.31	.52

For Contingent Numbers see Appendix 5.13.1, page 1177

5.4.2.2 Reading

Both sets of data on teachers' estimates of students' proficiency, communicative (TCR) and global (TGR), show similar characteristics; TA11, TB11 and TB13 being the tasks which correlate poorly with the teachers' estimates overall. These lower correlations in TA11 and TB11 confirmed the earlier finding in the inter-task correlations that these tasks are behaving differently from the other reading measures.

It is noteworthy that, in general, the correlations are lower for the TB tasks than for the TA ones. This may be explained, in part, by the fact that the A.S.A.B.S. students taking a module designed for the Science/Engineering group may have found themselves in unfamiliar territory resulting in erratic performances.

Neither set of data shows strong agreement between estimates and task scores, although the majority of tasks exhibit a level of validity of over +.45, regarded as satisfactory by Davies (1965).

5.4.2.3 Listening

Tutors' estimates for listening bore the same relation to test scores for communicative proficiency (TCL) as for global proficiency (TGL). The correlations were found to be, overall, slightly higher than for reading. Again, tutors were less successful in predicting students' listening scores in Session TB than in Session TA.

5.4.2.4 Writing

When the whole N.N.S. group is considered, the correlations between writing task scores and tutors' estimates of students' writing ability are slightly lower than those for reading and markedly lower than those for listening. As was the case for the other skills there is little to choose between the communicative and global estimates (TCW and TGW). The lower correlations between the Session TB tasks, previously noted, is not so apparent here. The global estimate for T032 was amongst the highest across the three skills.

The above observations have been made on the data relating to all the N.N.S. group. A study of the range of values for the individual Centres shows a wide divergence in some cases, indicating that tutors may have been variously competent in making estimates of their students' language skills, there may have been wide fluctuations in the ranges of ability in the various groups of students estimates were made for, or the tasks set in T.E.A.P. may not have matched well some tutors' perceptions of what the skill involves.

Davies (1965, pp.150-151) pointed out:

"... even after validation studies have been made, the inevitable lack of uniformity between one situation and another ... is likely to cause wide fluctuations among validity coefficients ... Validity, it must be expected, will vary: the fact that it does vary may well, for a proficiency test validated in widely diverse situations, itself be evidence of the test's validity."

5.4.3 Students' Self-Assessment of Language Proficiency

The students' self-assessment rating form is shown as Appendix 5.5.1., pages 1047-1048. It was developed on the same basis as the one used for the tutors' communicative assessment (p.418).

In Table 5W below, we list the Pearson product moment correlation coefficients between task scores in the various Sessions and the following student assessments:

1. SR: Students' self-assessments of reading ability.
2. SL: Students' self-assessments of listening ability.
3. SW: Students' self-assessments of writing ability.

As was the case for the tutors' estimates, these are shown for the N.N.S. group as a whole and for the five Centres for which more than twenty self-assessments were available.

TABLE 5W CORRELATIONS BETWEEN TASK SCORES AND STUDENTS' SELF ASSESSMENTS OF READING, LISTENING AND WRITING ABILITIES

TASK CODES	<u>SR</u>					<u>SL</u>					<u>SW</u>				
	ALL NNS	C3	C11	C14	C18 C20	ALL NNS	C3	C11	C14	C18 C20	ALL NNS	C3	C11	C14	C18 C20
<u>Reading</u>															
T012	.21	.29			.15										
TA11	.18	.08	.42	.18	(-.01)	.11									
TA12A	.25	.41	.57	.22	.15	.05									
TA12B	.28	.44	.65	.31	.17	.10									
TA13	.22	.20	.49	.07	.04	.20									
TB11	.22	.29		.19	.09	.34									
TB12A	.31	.02		.09	.25	.20									
TB12B	.29	.05		.08	.16	.21									
TB13	.25	.29		.33	.23	.05									
<u>Listening</u>															
T021							.68			.20					
T031							.56			.20					
TA21							.37	.50		.28					.45
TB21							(-.29)			.34					.27
<u>Writing</u>															
T011												.28			.22
T032												.26			.23
TA31											.52	(-.13)		.40	.20 (-.03)
TA32											.49	.20		.13	.07 .03
TB31											(-.03)			.25	.17 .31
TB32											.34			.11	.12 (-.02)

For Contingent Numbers see Appendix 5.13.1, page 1175

On the whole, the student self-assessments were poorly correlated with their task scores in T.E.A.P.; much less well for all three skills than were the tutors' estimates. This is to be expected in the case of overseas students who may have difficulty in assessing their own proficiency in a foreign language (cf. Chaplen 1970; Sen 1970; Jordan 1977a and Walker 1978) and some may be tempted to make optimistic assessments because they think it expedient to do so. As with the tutors' estimates, there were considerable variations across individual Centres. It is interesting to note that the C11 students' self-assessments were much higher than average in all three skills for the tasks on which data were provided. This may be explained, in part, by the wide range of ability of this small group of candidates from a language school in Hove. Having already been divided into classes for examination purposes, students were selected from three classes to take our experimental test. Students were perhaps more aware of their ability by virtue of having been previously placed in these classes.

5.4.4 Students' and Tutors' Estimates of Language Proficiency Compared with Aggregated Task Scores

Table 5X shows the correlation between students' pre-test scores for reading, listening and writing, aggregated across tasks within each Session and the various linguistic proficiency measures elicited from students and tutors.

These data are of interest mainly insofar as they show that tutors' estimates accord better with students' scores in Session T0 than in Sessions TA and TB. The data display the same characteristics noted previously in that tutor estimates correlate more highly than do student self-assessments and there is little difference in the strength of the correlations between tutors' communicative and global estimates of the three skills, although TCR and TCL (the tutors' communicative estimates of reading and listening) do correlate better with test scores than do the more impressionistic global estimates.

TABLE 5X CORRELATIONS BETWEEN READING, LISTENING AND WRITING TASK SCORES,
AMALGAMATED BY SESSION, AND STUDENTS' AND TEACHERS' ASSESSMENTS

Session Task	SR	SL	SW	TCR	TCL	TCW	TGR	TGL	TGW
<u>Reading</u>									
TO TO12	.21			.58			.60		
TA TA11) TA12A & B) TA13)	.28			.63			.54		
TB TB11) TB12A & B) TB13)	.38			.53			.45		
<u>Listening</u>									
TO TO21) TO31)		.40			.60			.60	
TA TA21		.22			.54			.46	
TB TB21		.15			.46			.41	
<u>Writing</u>									
TO TO11) TO32)			.25			.45			.53
TA TA31) TA32)			.28			.47			.35
TB TB31) TB32)			.31			.44			.40

For Contingent Numbers see Appendix 5.13.1, page 1178

Interpretation of correlational data which involves comparison should be done with caution when the score ranges being correlated vary across the correlation matrix as they do here. (The various tasks have different maximum possible scores.)

5.4.5 Relationships Within and Between Tutors' and Students' Assessments of Language Proficiency

Table 5Y below shows the Pearson product moment correlations within and between the students' self-assessments and the two sets of tutors' assessments. These data are of interest insofar as they illustrate the relative homogeneity of the students' self-assessments of their proficiency in various language skills. The same pattern is repeated in the tutors' estimates particularly in the communicative assessments (TCR, TCL, TCW).

These data might be taken as evidence of a fairly strong general ability factor at work, resulting in similarity between the estimates in the three skill areas. Alternatively they might be regarded as evidence that the teachers in particular were unable to distinguish any great differences in the various skill areas as exhibited in the proficiency of their students. This may be due to an unawareness of a student's capabilities in different skills if not directly responsible for a particular area of tuition. The higher correlations between reading and listening and also between reading and writing might also suggest, however, that there is a perceived connection between these pairs of skills in the proficiency demonstrated by the students for whom the teachers are responsible. The lower correlations obtaining between listening and writing in the tutors' assessments would tend to support this hypothesis.

In general, the students' estimates of their own proficiency do not correlate very highly with those of their teachers, perhaps further evidence of the difficulty overseas students may have in assessing their own proficiency in a foreign language.

TABLE 5Y COEFFICIENTS OF CORRELATION BETWEEN STUDENTS' SELF ASSESSMENTS AND TEACHERS' GLOBAL AND COMMUNICATIVE ASSESSMENTS OF PROFICIENCY IN READING, LISTENING AND WRITING

	SR	SL	SW	TCR	TCL	TCW	TGR	TGL	TGW
SR		.68	.73	.34			.31		
SL			.68		.39			.38	
SW						.25			.27
TCR					.79	.78	.68		
TCL						.70		.74	
TCW									.67
TGR								.75	.81
TGL									.67
TGW									

For Contingent Numbers see Appendix 5.13.1, page 1179

5.5 COMPILING THE FINAL FORM OF T.E.A.P.

5.5.1 Introduction

Now that the data analysis had been completed at both item and task levels (pp.446-505), we were able to take decisions about which tasks to include in the final form of the test and to decide how, if at all, these tasks should be amended.

It had been recognised prior to the administration of the pre-tests that the Sessions, as they stood, were perhaps too lengthy but this was considered unavoidable given their experimental nature. The main criticisms voiced by the N.N.S. group sitting the test concerned the length of the test, its involved nature and the amount of examining time involved (v. p.420 et seq.). Whilst accepting that the more samples of each skill we could take the greater the reliability of the test was likely to be, practical considerations such as the effect on candidates and cost had to be borne in mind. Given the integrated nature of certain of the activities, however, and given the need to protect against format effect, we were unwilling to reduce the length of the test too much without justifiable grounds for doing so. The tasks comprising each of the three constructs reading, listening and writing were considered in turn.

5.5.2 Decisions on Which Tasks to Include

5.5.2.1 Reading

As shown in Table 5N, page 475, of the reading tasks the multiple choice questions on reading comprehension, TA11 and TB11 were the least reliable in terms of Cronbach α and, perhaps more importantly, they had the lowest correlations with other reading measures (v. pp.1127-1129) and with tutors' estimates of reading proficiency (Tables 5U and 5V, pp.497-498). In the factor analysis (pp.489-493) Varimax rotations showed these tasks to be sometimes loading on a factor different from those of the other reading tasks. Furthermore,

multiple choice items based on a reading comprehension passage are notoriously difficult to construct. For all these reasons it was decided, after discussion with the Project Working Party, that tasks TA11 and TB11 should be removed from the battery.

The data on the gap-filling tasks TA12A/B and TB12A/B were scrutinised to see whether it would be possible to delete either the task where candidates had to identify the place where words were omitted or the one in which they had to supply the missing word. In the end it was agreed that these tasks should remain, partly on account of their high reliability (internal consistency) measures (p.475) and their strong correlation with tutors' estimates (pp.497-498). It was decided, however, that we would shorten these tasks by reducing the number of items slightly.

5.5.2.2 Listening

Since all four listening tasks correlate fairly well with one another (v. Appendix 5.11.1, p.1118) and have reasonably high values of Cronbach α and appropriate difficulty levels, with the exception of task TB21 which, for the reasons stated earlier (p.454), has been revised to make it easier, it was decided to retain all the listening tasks.

5.5.2.3 Writing

In page 465 we reported a flaw in the rubric of the editing tasks TA32 and TB32. This has been rewritten to make it clear that the task is one of error recognition only. In view of the disparity in difficulty levels between TB32 compared with TA32 steps have been taken to make TB32 easier.

As observed earlier (p.485) the inter-task correlations were low for this construct for the reasons explained there. These correlation data make it clear either that the various tasks make different kinds of demands on the candidates or that performance is highly task-dependent. Marker variability may also be exerting a strong influence. In any case it is right that all the writing tasks should be retained.

5.5.3 Data Analysis at the Construct Level

As it was decided to provide results separately for each construct and likely that candidates will be offered two versions of T.E.A.P., Sessions TO + TA or TO + TB, it was necessary to analyse the pre-test data at the construct level, using only data on those candidates who had taken one of these combinations of Sessions.

In Table 5Z below we describe the means and standard deviations of the scores of all N.S. and N.N.S. candidates in this category. It shows the statistics of their total scores in reading, listening, writing and grammar across Sessions, weighted in accordance with the proposed weighting in the final form of the test. Details of these weightings are given below.

In deciding upon these weightings it was necessary to make sure that they resulted in two versions of the test TO + TA and TO + TB in which the skills and sub-skills were, as far as possible, equally weighted. Because of the lower maximum number of marks available in TB21 as compared to TA21 we multiplied the TB21 raw marks by 20/13 to bring them into line. The maximum possible raw scores for TA13 and TB13 differed by one and the TA13 raw scores were accordingly multiplied by 19/18 to make them equivalent. The raw scores of the reading tasks TA13, TB13 and TO12 were then doubled in order to balance them with the scores of the gap-filling tasks, TA12 and TB12, which were already out of 42. This double weighting was thought necessary given that it was mainly lower order skills that were being tested in TA12 and TB12.

We see from Table 5Z that there is a noticeable difference in the performance of the two groups on all four constructs. Whereas the N.S. group mean is eighty per cent or more, that of the N.N.S. group is substantially lower. This discrepancy must be considered satisfactory. If the N.S. group had not scored highly, the difficulty level of T.E.A.P. would have been in doubt. Another satisfactory feature of these data is the close similarity in construct means between the N.N.S. groups taking TO + TA and TO + TB respectively, indicating that both versions were of comparable difficulty, given that the two groups of N.N.S. students were of comparable ability.

TABLE 5Z

COMPARISON OF NS AND NNS PERFORMANCE ON SKILL COMPOSITES IN VARIOUS
SESSION COMBINATIONS

	Sessions TO + TA		Sessions TO + TB		Sessions TO + TA and TO + TB combined	
	NS n = 41	NNS n = 132	NS n = 46	NNS n = 103	NS n = 87	NNS n = 235
<u>Reading*</u>						
Mean	103.34 (82%)	57.44 (46%)	96.54 (77%)	61.40 (49%)	99.75 (79%)	59.17 (47%)
s.d.	13.97	21.63	15.36	23.68	15.11	22.64
<u>Listening*</u>						
Mean	45.66 (86%)	23.05 (43%)	41.57 (78%)	21.41 (40%)	43.49 (82%)	22.33 (42%)
s.d.	4.23	10.25	5.55	10.68	5.38	10.47
<u>Writing</u>						
Mean	62.73 (87%)	40.93 (57%)	56.46 (78%)	42.90 (60%)	59.41 (82%)	41.80 (58%)
s.d.	6.83	13.58	6.01	13.30	7.13	13.49
<u>Grammar</u>						
Mean	57.66 (96%)	40.02 (67%)	55.43 (92%)	40.69 (68%)	56.48 (94%)	40.31 (67%)
s.d.	2.13	9.53	3.87	10.26	3.36	9.86

* The raw marks in TA and TB were scaled up to be equivalent and all tasks were weighted in accordance with intended weightings in the final version of TEAP.

There is evidence that both N.N.S. and both N.S. groups were of equal ability. All took the same grammar test and the means for that construct were similar, for the two N.N.S. and for the two N.S. groups. Thus it is reasonable to conclude that the two versions of the test are fairly well matched in terms of difficulty.

When a comparison is made between performance in the various constructs, the following points emerge. Whilst the N.S. students scored equally well in the reading, listening and writing constructs, the N.S. group taking TO + TA did rather better than did that taking TO + TB. This pattern was not reflected, however, in the performance of the N.N.S. groups. Both did noticeably least well in listening and best in writing. Since the N.S. students found all three parts of the test of equivalent difficulty, there must be a suggestion that the disparity in N.N.S. performance is due to different levels of attainment in these constructs.

The decision to report the results of T.E.A.P. separately for each construct is supported by the inter-construct correlations shown in Table 5AA below.

This table shows that the four constructs correlate fairly well with one another, but not at a level where reporting them separately would be redundant. The correlation matrices for the two versions of the test are not identical as one might have hoped; the writing score correlates better with the other scores in TO + TB than in TO + TA. No obvious reason for this is discernible.

TABLE 5AA
INTER-CORRELATION BETWEEN CONSTRUCT SCORES

Sessions TO + TA

	R	W	L	G
R	-	.76	.72	.67
W		-	.56	.58
L			-	.69
G				-

Sessions TO + TB

	R	W	L	G
R	-	.78	.81	.67
W		-	.73	.75
L			-	.70
G				-

Sessions TO + TA and TO + TB combined

	R	W	L	G
R	-	.77	.75	.67
W		-	.62	.66
L			-	.69
G				-

The factor analysis suggested that reading might be expected to correlate fairly well with both listening and, to a lesser extent, with writing. This analysis also indicated that the listening tasks were likely to load only with other comprehension tasks. Thus, it may be in line with expectation that Table 5AA above shows the correlations between listening and writing to be rather lower than the others.

5.5.4 Reliability of the Constructs

We were also interested in the reliability of these skill composites in the combinations of Sessions. Using Willmott et al.'s (1975) formula for pooling coefficient alpha on the N.N.S. group's alpha coefficients, the coefficients listed in Table 5BB below were obtained for the reliability (internal consistency) of the skill composites in the two-session combinations. As a result of the decisions taken on the reduction of tasks described in Section 5.5.2 above, TA11 and TB11 were omitted from the calculations for reading.

TABLE 5BB
POOLED CRONBACH α 's FOR SKILL COMPOSITES IN THE
VARIOUS COMBINATIONS OF SESSION

TO + TA	Reading	0.93
TO + TB	Reading	0.93
TO + TA	Listening	0.91
TO + TB	Listening	0.91
TO + TA	Writing	0.96
TO + TB	Writing	0.94

The Cronbach α for the single grammar component T041 was 0.92

The reliabilities of the total scores for reading, listening and writing are computed on the basis of the internal consistency reliabilities of the tasks in each of the Sessions. A problem occurs in connection with the writing tasks because, in the measurement of this construct, there is evidence of error variance due to inter-marker differences which is not accounted for in the coefficient α 's.

The inter-marker reliabilities (v. Table 5I, p.468 above) for each of the three writing tasks in the mark/re-mark experiment described above were calculated at:

T011	0.73
TA31	0.74
TB31	0.65

Using the method proposed by Anastasi (1982) we added together the error variance due to markers and the error variance measured by coefficient α for each of the three writing tests, to give us the new correlation coefficients listed below:

T011	0.67
TA31	0.66
TB31	0.52

Using Willmott et al.'s (1975) formula for pooling coefficient α we then pooled these to get an overall reliability figure for the two writing tasks in each of the combined Sessions. These were:

T011 + TA31	0.77
T011 + TB31	0.70

If we had been able to take T032 and the editing task in TA and TB into account also, we might have obtained improved reliability coefficients for writing in each of the combined Sessions.

One might also expect the reliability of the combined Session marks to improve further with double marking.

As regards the reading and listening tasks, no special mark/re-mark experiment took place, therefore it was not possible to estimate error variance due to marking, but this was not so important in view of the highly structured marking schemes generally employed in these parts of the test. We did, however, re-mark one hundred scripts randomly selected from each Session; fifty from two 'A' level Centres, fifty from two post-graduate Centres. All tasks except the writing and the multiple choice tasks, which had been machine scored, were re-marked. This exercise was conducted to ensure that the first markers had been operating at a consistent standard and following the instructions laid down in the marking schemes. It was not intended as a true mark/re-mark exercise however and whereas in

the writing re-mark exercise described earlier (p.467) the scripts had been clean for both markers, in this later exercise on the non-writing tasks the second marker was obviously aware of the first set of marks. The correlation coefficients between the first and second markings for the tasks listed in Table 5CC below must therefore be treated with due caution.

The conditions under which the marking was carried out were not parallel and the coefficients may well have been different if they had been. Given these provisos, the correlation coefficients would seem to lend some limited support to the view that the reliability of the listening and reading tasks as measured by coefficient α would not be reduced too greatly if one was to add in the error variance arising out of marker unreliability.

TABLE 5CC
PRODUCT MOMENT CORRELATIONS OBTAINED BETWEEN FIRST AND
SECOND MARKINGS OF VARIOUS TEST TASKS

	r	n
T012	0.99	100
T021	0.99	100
T031	0.96	100
TA12A	0.96	98
TA12B	1.00	100
TA13	0.97	99
TA21	0.98	99
TB12A	0.99	98
TB12B	0.97	99
TB13	0.95	98
TB21	0.98	91

5.5.5 The Concurrent and Predictive Validity of the Constructs

Individual tasks may or may not alone show satisfactory coefficients of correlation either with internal or external criteria. What matters to both test constructor and user is the effect of the composite scores in reading, listening or writing. In Tables 5DD and 5EE below the composite scores for reading, listening, writing and grammar are examined in relation to the concurrent and predictive validity criteria we were able to collect (v. Section 5.1.5 above).

Before discussing the general implications arising out of these data, the greater incidence of low correlations between TO + TB and the concurrent and predictive validity measures warrants some explanation. It may be that the teachers making these particular estimates were less reliable than those doing so in respect of the TO + TA combination. It is also likely that the problems of the items in TB13, TB21 and the effect this had on TB31, may have contributed to a greater degree of error variance in this Session. About one third of the students taking the TO+TB combination were A.S.A.B.S. students and it is possible that their performance might have been affected by having to take the Science/Engineering oriented module (TB). This again might have contributed to the amount of error variance in the TB Session.

It is noticeable in the correlation coefficients for TO + TB quoted in Table 5FF, that whereas the coefficients between T.E.A.P. and teachers' concurrent assessments are reasonable, the coefficient of both with subject tutors' language assessments are low. Leaving aside the small n's involved, it does seem likely that English language teachers co-operating in the concurrent validity study were, for these candidates, better able to assess a student's language proficiency than subject tutors, perhaps through far greater exposure to students' language capabilities. The time span between the predictive and concurrent estimates would, of course, also affect the correlational pattern.

TABLE 5DD

CORRELATIONS BETWEEN SKILL COMPOSITE SCORES AND LANGUAGE
TEACHERS' ASSESSMENTS OF STUDENTS' READING, LISTENING
AND WRITING ABILITIES AT THE TIME OF TAKING THE TEST

<u>Session</u>										
<u>TO + TA</u>										
	1	2	3	4	5	6	7	8	9	10
1	-	.76	.72	.67	.67	.50	.66	.64	.59	.62
2		-	.56	.58	.57	.46	.54	.58	.51	.52
3			-	.69	.53	.23*	.63	.45	.44	.52
4				-	.44	.39	.50	.41	.34*	.49
5					-	.79	.88	.68	.64	.63
6						-	.79	.71	.72	.63
7							-	.67	.62	.71
8								-	.90	.83
9									-	.78
10										-
<u>TO + TB</u>										
	1	2	3	4	5	6	7	8	9	10
1	-	.78	.81	.67	.63	.53	.51	.47	.28*	.49
2		-	.73	.75	.65	.61	.67	.68	.34*	.39*
3			-	.70	.60	.47	.61	.24*	.36*	.13*
4				-	.63	.58	.60	.00*	.04*	.30*
5					-	.79	.48	.37*	.48	.49
6						-	.38	.63	.54	.51
7							-	.69	.58*	.69
8								-	.41*	.61
9									-	.25*
10										-
<u>TO + TA/TB</u>										
	1	2	3	4	5	6	7	8	9	10
1	-	.77	.75	.67	.66	.49	.61	.66	.59	.64
2		-	.62	.66	.60	.50	.59	.65	.53	.57
3			-	.69	.54	.36	.63	.48	.47	.53
4				-	.51	.46	.54	.47	.39	.56
5					-	.78	.80	.69	.65	.67
6						-	.67	.71	.71	.63
7							-	.68	.63	.72
8								-	.86	.83
9									-	.75
10										-

* not significantly different from zero at the 5% level

For contingent numbers see Appendix 5.13.2 pages 1181-1182

KEY

- | | |
|--|--|
| 1. Reading composite score | 7. Teachers' assessments listening (communicative) |
| 2. Writing composite score | 8. Teachers' assessments reading (global) |
| 3. Listening composite score | 9. Teachers' assessments writing (global) |
| 4. Grammar score | 10. Teachers' assessments listening (global) |
| 5. Teachers' assessments reading (communicative) | |
| 6. Teachers' assessments writing (communicative) | |

TABLE 5EE
 CORRELATIONS BETWEEN SKILL COMPOSITE SCORES AND SUBJECT TUTORS'
 ASSESSMENTS OF STUDENTS' LANGUAGE ABILITIES AT THE
 END OF THE SECOND TERM

<u>Session</u>							
<u>TO + TA</u>							
	1	2	3	4	5	6	7
1	-	.77	.81	.70	.62	.55	.70
2		-	.66	.64	.63	.61	.70
3			-	.80	.55	.49	.61
4				-	.45	.50	.46
5					-	.85	.89
6						-	.79
7							-

n = 57

<u>TO + TB</u>							
	1	2	3	4	5	6	7
1	-	.78	.80	.60	.15*	.29*	.19*
2		-	.76	.68	.11*	.27*	.29*
3			-	.64	.17*	.27*	.29*
4				-	.23*	.24*	.42
5					-	.78	.86
6						-	.74
7							-

n = 35

<u>TO + TA/TB</u>							
	1	2	3	4	5	6	7
1	-	.79	.81	.71	.48	.47	.55
2		-	.71	.68	.49	.51	.60
3			-	.76	.45	.43	.53
4				-	.43	.43	.48
5					-	.83	.88
6						-	.78
7							-

n = 92

*not significantly different from zero at the 5% level

KEY

- | | |
|------------------------------|---|
| 1. Reading composite score | 5. Subject Tutors' assessments: reading |
| 2. Writing composite score | 6. Subject Tutors' assessments: writing |
| 3. Listening composite score | 7. Subject Tutors' assessments: listening |
| 4. Grammar score | |

In Table 5FF below we have correlated candidates' total test scores across all parts of the test against aggregates of both types of language teachers' concurrent assessments of language proficiency and against aggregates of subject tutors' assessments of language proficiency and academic progress made at the end of Term 2.

It would seem to be the case from the data presented in Table 5FF below and Tables 5DD and 5EE above, that the concurrent validity of T.E.A.P. is on the whole superior to its predictive validity. The correlations of T.E.A.P. with both the subject tutors' language assessments and estimates of academic progress at the end of the second term are lower than those with teachers' estimates of language ability taken at the same time as T.E.A.P. was administered. This is to be expected since T.E.A.P., as a measure of language proficiency, should be nearer to the teachers' estimates of English than to those of academic success with all the other abilities the latter involves (v. Davies 1965).

There is slight evidence from those correlations which are significant that T.E.A.P. might nevertheless be considered a better predictor of academic success than were the teachers' language ratings. These teachers' concurrent assessments of language ability are marginally better than T.E.A.P. in the correlations with subject tutors' assessment of language proficiency at the end of Term 2.

TABLE 5FF
CORRELATIONS BETWEEN TOTAL TEST SCORES, TOTAL TEACHERS'
ASSESSMENTS AND SUBJECT TUTORS' ASSESSMENTS
(LANGUAGE AND ACADEMIC PROGRESS)

Session

TO + TA

	1	2	3	4	5	6
1	-	.73	.73	.64	.70	.66
2		-	.52	.43	.50	.32
3			-	.83	.73	.46
4				-	.65	.38
5					-	.63
6						-

TO + TB

	1	2	3	4	5	6
1	-	.70	.59	.57	.25*	.25*
2		-	.51	.14*	.29*	.09*
3			-	.70*	.29*	.82*
4				-	.19*	.42*
5					-	.13*
6						-

TO + TA/TB

	1	2	3	4	5	6
1	-	.73	.69	.68	.57	.52
2		-	.52	.51	.47	.27
3			-	.82	.70	.39*
4				-	.65	.34
5					-	.48
6						-

* not significantly different from zero at the 5% level

For contingent numbers see Appendix 5.13.3 page 1184

KEY

1. Total test scores (reading + listening + writing)
2. Test scores grammar
3. Total teachers' concurrent assessments, R + L + W (communicative)
4. Total teachers' concurrent assessments, R + L + W (global)
5. Total subject tutors' language assessments, R + L + W, at end of Term 2
6. Subject tutors' assessment of academic progress at end of Term 2

5.5.6 Decision on Which Constructs to Include

After analysis of the experimental pre-test data a decision had also to be taken about the place of T041, the multiple choice test of grammar, in future versions of the battery.

In terms of face validity (v. Section 5.1.6.2.1 and Table 5CCC, p.1067) it would seem that the N.N.S. students thought this a good, if not a better test of E.A.P. proficiency than those tasks we had considered to be more content valid; this is perhaps to be understood in terms of candidates' previous testing experience (v. Porter 1983) and the fact they considered they had performed best on this task (v. Table 5DDD, p.1068). The reactions from the teachers had been less positive. If one's job is to make good deficiencies in students' E.A.P. proficiency, then a more content valid test in terms of the broad macro-activities involved, as against a single test of knowledge of linguistic structures, might appear to be more useful.

In terms of reliability (v. Table 5N , p.474above) T041 was good, if not better than all tests in the battery, with an internal consistency coefficient α of 0.92. Given that it was machine marked, this reliability figure is not subject to the additional error variance noted for the other tests arising out of marker unreliability.

In terms of the purposes for which our test had been constructed and the a priori content validation described earlier, the inclusion of a 'discrete point' multiple choice test of grammatical knowledge was not felt to be a high priority for T.E.A.P. It was included in the experimental pre-test to examine how well it worked as compared to other tasks which were deemed more valid in terms of our specification.

After close attention had been paid to the internal and external analysis of test results it was thought by the Project Working Party that the multiple choice test of grammatical knowledge was not contributing any useful additional information on the constructs being measured and that since grammar was tested elsewhere (for example, in the writing and editing tasks) and there was a pressing

need to reduce the size of the test, this task should be deleted. It was felt by the Working Party that it was not part of our content specification for the test and would not fit easily into the way in which the T.E.A.P. results would be profiled. In this profile each student would receive a separate score for reading, listening and writing; translated into the form of a behavioural grade. Therefore, despite respectable reliability figures and criterion-related validity only slightly poorer than the rest of the T.E.A.P. components combined, the grammar task was to be deleted from the battery.

If the purpose of our proficiency test was solely prognostic, e.g. in terms of academic success and we were not concerned with providing a profile of students' ability in the different macro-skills required in an E.A.P. context, then serious consideration would have to be given to the status of T041. It could be argued that the lower predictive validity of T041 (v. Table 5FF above) might be acceptable given the amount of time, resources and cost, the construction of the alternative, more face valid battery would involve. The greater efficiency and reliability of T041 might well have encouraged us to conclude that, validity considerations notwithstanding, the test of grammar might be a sufficient indicator on its own of a student's ability to cope with the language demands made on students by English medium study.

5.5.7 Conclusions on the Validity and Reliability of the Constructs

Within the limitations of this study and though the evidence is slight, both concurrent and predictive validity may be claimed for T.E.A.P. Significant concurrent validity coefficients over +0.6 were established for the listening and reading composites in both the T0 + TA and T0 + TB combinations in relation to teachers' communicative assessments of students' language proficiency. No significant correlation fell below 0.46 for any of the skill composites when correlated with the teachers' global or communicative ratings for that macro-skill. Lower correlations were also reported but none were significant at the 5% level. Slightly higher concurrent than predictive validity was indicated. The reading tests

appeared to show the highest validity figures followed by the listening with writing tests performing far worse. The poor showing of the latter is due, in part at least, to the lower reliability of the assessment of this construct (v. below). The greater number of discrete items in the reading tasks collectively perhaps explain the greater reliability and the slightly better validity coefficients than those obtained for the listening tasks.

Estimates of reliability, incorporating both internal consistency and mark/re-mark reliability were made for the writing construct. In the case of the other two constructs, only internal consistency estimates were made. The internal consistency measures (Cronbach α) were high and similar for all three constructs. When mark/re-mark reliability is considered, it is likely that, had it been possible to make estimates for this effect for reading and listening, these would have been higher than the ones shown for writing due to the greater objectivity of marking. Thus, if pooled coefficients had been available, these might have been higher than the values 0.77 and 0.70 shown for writing.

Given our reservations about the TB component and concern about its parallelism with TA we would nevertheless consider that the concurrent and predictive validity of T.E.A.P. (TO + TA, TO + TB) are reasonable. We have demonstrated that there was a positive relation between the English proficiency as evidenced by our tests and the teachers' and subject tutors' estimates of language proficiency and the latters' predictions of academic success. We had also demonstrated that all three constructs are acceptably reliable. These construct reliabilities must not be compared with reported reliabilities for tests in which all constructs are combined for the purposes of final reporting.

5.5.8 Analysis of Performance in the Two Versions of the Test

So far the analysis has been concerned with the pre-test version of T.E.A.P. at the item, task and skill composite levels. Earlier

(pp.327-330) we discussed the reasons that led us to devise two versions of the test, one comprising T0 + TA for A.S.A.B.S. students and one comprising T0 + TB for Science and Engineering students. In this section we examine the data to see whether it indicates that one version would be sufficient.

As explained earlier (Section 5.1.2.3, p.410), in pre-testing the two versions we took steps to ensure that Session TA designed for A.S.A.B.S. students was also taken by Science and Engineering students for whom Session TB was intended and vice versa. The group means for Science/Engineering and A.S.A.B.S. are shown in Tables 5GG and 5HH below, which also show that a t-test of significance between the group means indicates no significant difference at the 1% level for any of the tasks. At the 5% level only on the test of grammatical knowledge T041 is there a significant difference between the two groups. The single difference is difficult to explain, but is reflected also in the performance of the N.S. groups on this task, particularly as regards 'A' level students (v. Appendix 5.10, p. 1102-1115). One might speculate that this could arise from the A.S.A.B.S. students having to be more aware than the Science/Engineering students of how language functions in their particular courses of study. Since the T0 tasks were intended to be unbiased towards either group of students and there was no significant difference in performance levels, it may be concluded that the language proficiency of the two groups is similar. It follows that the non-significant results for the TA and TB tasks also indicate no bias.

TABLE 5GG

DIFFERENCES BETWEEN THE PERFORMANCES
OF SCI./ENG. AND A.S.A.B.S. STUDENTS
ON TO/TA COMBINATION

Maximum Possible Score	Task Code	Sci./Eng.		A.S.A.B.S.		t																			
		mean	s.d.	mean	s.d.	1%	5%																		
23	TO12	10.9	5.3	9.9	4.9	NS	NS																		
17	TA11							8.6	3.3	8.0	2.9	NS	NS												
21	TA12A													8.2	4.8	9.0	4.7	NS	NS						
21	TA12B																			6.0	4.0	7.2	4.1	NS	NS
18	TA13																								
21	TO11	12.8	4.8	13.1	4.9	NS	NS																		
21	TO32							10.4	5.6	10.7	5.5	NS	NS												
21	TA31													11.5	5.0	11.8	5.5	NS	NS						
9	TA32																			5.5	1.7	5.6	1.9	NS	NS
15	TO21	5.7	3.9	6.3	3.8	NS	NS																		
18	TO31							4.8	3.5	5.1	3.0	NS	NS												
20	TA21													11.5	4.4	11.6	4.7	NS	NS						
60	TO41	G	37.9	8.6	41.6	10.1	NS	✓																	
			n = 80		n = 40																				

NS non-significant

✓ significant

1% significant at 1% level

5% significant at 5% level

TABLE 5HH
DIFFERENCES BETWEEN THE PERFORMANCES
OF SCI./ENG. AND A.S.A.B.S. STUDENTS
ON TO/TE COMBINATION

Maximum Possible Score	Task Code	Sci./Eng.		A.S.A.B.S.		t		
		mean	s.d.	mean	s.d.	1%	5%	
23	T012	} R	12.9	5.5	12.7	6.6	NS	NS
17	TB11		10.3	2.8	10.4	3.1	NS	NS
21	TB12A		11.5	4.3	11.4	3.5	NS	NS
21	TB12B		9.2	4.2	8.4	3.5	NS	NS
19	TB13		8.3	3.9	8.0	3.8	NS	NS
21	T011	} W	13.8	4.9	13.1	6.5	NS	NS
21	T032		12.5	5.0	13.1	5.3	NS	NS
21	TB31		12.8	4.6	14.5	2.5	NS	NS
9	TB32		4.3	2.2	4.4	1.4	NS	NS
15	T021	} L	7.5	3.8	7.5	4.0	NS	NS
18	T031		7.1	4.0	6.1	3.4	NS	NS
13	TB21		5.0	3.2	4.2	3.5	NS	NS
60	TB41	G	41.0	10.6	41.7	8.5	NS	NS
			n = 76		n = 21			

We also arranged for two groups of students, one Science/Engineering and the other A.S.A.B.S., to take both Sessions TA and TB. Table 5II below shows how the students performed in the two versions. It is noteworthy that some of these findings are slightly at odds with those reported above. In the reading tasks the Science/Engineering group performed significantly better, at the 5% level, on TB11 than on TA11 and on TB12B than on TA12B, but significantly poorer on TB13 than on TA13; this last result is contrary to expectation since TB13 was designed to be appropriate to the Science/Engineering group whereas TA13 was not. There were no significant differences for any of the reading tasks for the A.S.A.B.S. group.

Both groups of students performed significantly better, at the 1% level, on TA32 as compared with TB32 and on TA21 as compared with TB21. These results, however, are not incompatible with the results in Tables 5GG and 5HH above. They merely demonstrate that for both groups of students, the intended equivalence in level of difficulty between the two versions had not been achieved in all tasks; it does not indicate bias for or against one or other group of students.

There is some slight, but contradictory, evidence that the Science/Engineering group might be at a disadvantage as compared with the A.S.A.B.S. group if required to take Session TA rather than TB. It may well be that the background knowledge of the Science/Engineering group played a part in improving some of their test scores. Perhaps it is a question of motivation induced by the greater face validity of Session TB.

TABLE 5II DIFFERENCES IN THE PERFORMANCES ON TA AND TB OF
SCI./ENG. AND A.S.A.B.S. STUDENTS WHO TOOK BOTH

Maximum Possible Score	Task Code	Sci./Eng.			A.S.A.B.S			A.S.A.B.S			
		mean	s.d.	t	1%	5%	mean	s.d.	t	1%	5%
17	TA11)	8.8	3.2		NS	✓	10.4	3.3		NS	NS
17	TB11)	10.4	2.7				10.4	2.4			
21	TA12A)	10.3	4.2		NS	NS	12.0	5.4		NS	NS
21	TB12A)	12.3	5.2				12.4	5.0			
21	TA12B)	7.0	4.3		✓	✓	9.6	5.0		NS	NS
21	TB12B)	10.0	5.4				9.4	5.0			
*19	TA13)	9.9	3.2		✓	✓	11.0	4.3		NS	NS
19	TB13)	7.8	3.8				9.6	4.4			
21	TA31)	14.0	3.5		NS	NS	14.3	5.4		NS	NS
21	TB31)	12.5	5.2				12.0	6.6			
9	TA32)	5.8	2.2		✓	✓	6.4	1.6		✓	✓
9	TB32)	4.2	1.5				4.5	2.2			
20	TA21)	10.02	4.0		✓	✓	11.3	4.4		✓	✓
*20	TB21)	6.4	5.2				6.7	4.7			

n = 43

n = 48

* Scaled to same maximum score for both TA and TB

In the end the decision whether to have separate Sessions for Science/Engineering and the A.S.A.B.S. students was based on considerations of marketing and cost rather than on the research evidence. Given that the J.M.B. offers a test biased towards the Technical/Science student rather than the A.S.A.B.S. one, if we were to offer a more general second paper for all, the Science/Engineering group might prefer to sit the J.M.B. test with its ostensibly greater face validity for them. If we were to offer the Science/Engineering (TB) version only, although we might not, in fact, disadvantage the A.S.A.B.S. students, given the comments made in responding to the follow up questionnaire (v. pp.423-427) by this group of students, there is evidence to suggest they might feel themselves to be disadvantaged. Therefore, a decision was taken that T.E.A.P. will maintain two versions of Session II, one designed principally for the A.S.A.B.S. group and the other for the Science/Engineering group.

5.5.9 External Validation of Sessions and Combinations of Sessions

In addition to teachers' concurrent estimates of students' language proficiency, we had also asked co-operating institutions to let us have details of any other concurrent validity data that were available (v. Section 5.1.5 above).

Given the lengths of the Sessions of the test we wished to trial, it was not possible, except in a limited number of cases, to administer concurrently other well established proficiency tests such as E.L.B.A. or the Davies (E.P.T.B.) Test to those students taking various combinations of the pre-test version of T.E.A.P. As we stressed in Chapter 2, this is an essential step, once the final version of the test is ready. It is hoped that, having secured the necessary authorisation, this might take place in the future. The opportunity to compare T.E.A.P. with the new E.L.T.S. battery did not materialise, but again we hope that it will be possible to arrange this in the long term.

In the event, we were able to collect and compare the E.L.T.S. scores of only a small number of candidates who sat the pre-test version of T.E.A.P. and these are reported in Table 5JJ below, together with a smaller number of those with E.L.B.A. scores. In addition, we collected information from those institutions which administered subject and language tests to students on entry (again low n's) and the comparison of the scores achieved on these with those on T.E.A.P. are also reported in Table 5JJ below. For the purposes of concurrent validation, therefore, we had to rely mainly on comparison of T.E.A.P. scores with teachers' and students' assessments of language proficiency.

Caution must be exercised in interpreting the statistics in Table 5JJ below as some of the measures were not taken at the time T.E.A.P. was administered. The E.L.T.S. scores were often from tests taken six months or more before the administration of T.E.A.P. Most of the external criteria, being in the form of single final grades, were also on a very narrow scale of 0-5 and given that the total scores for T.E.A.P. Sessions were out of quite large numbers this may have served to depress the correlations reported.

Given our reservations about TB due to the unsatisfactory performance of certain of the tasks, in particular TB21, and given the error variance that might also have resulted from A.S.A.B.S. students taking TB, designed for the Science/Engineering group (v. Section 5.1.6 above) and vice versa in the case of TA, the correlations are quite encouraging. One might tentatively suggest that they provide some evidence for the validity of T.E.A.P. on this particular sample.

The amount and quality of external evidence presented here is obviously very restricted. We would advise due caution, therefore, in making any claims for the validity of T.E.A.P. until the revised, final version has been more extensively validated against external criteria. What is required is a validation such as is taking place for the E.L.T.S. battery under the auspices of the Institute for Applied Language Studies, at the University of Edinburgh.

TABLE 5JJ

CONCURRENT VALIDITY

WHOLE SESSION(S) VERSUS OTHER TEST RESULTS

	S E S S I O N S				
	TO	TA	TB	TO+TA	TO+TB
<u>Established Tests</u>					
1) ELTS (pre.Sep)	.48 (39)	.57 (38)	.14*(36)	.84 (17)	.23*(21)
2) CPE (Dec)	.41 (24)	.75 (9)	.33*(23)		.44 (21)
3) FCE (Dec)	-	.65 (18)	-	-	-
4) ELBA (Sep.Oct)	-	.87 (13)	-	-	-
5) JMB (Mar)	.88 (14)	.85 (17)	.70 (20)	-	.73 (11)
<u>Institutional</u>					
<u>Language</u>					
<u>Entry Tests</u>					
CONCORDE COLLEGE					
Grammar (Sep)	.81 (26)	.33*(13)	.38*(14)	.60 (12)	.50*(13)
JMB type (Nov)	.79 (29)	.69 (30)	-	.79 (29)	-
PADWORTH COLLEGE					
Grammar (Sep)	.95 (9)	.59* (9)	-	.84 (9)	-
LOUGHBOROUGH TECH.					
Dictation (Sep)	-	.73 (55)	.68 (50)	-	-
JMB type (Sep)	-	.83 (43)	.65 (39)	-	-
SOUTHAMPTON UNIVERSITY					
Entry Test					
Grammar Writing (Sep)	-	.80 (27)	.37*(23)	-	-

* not significant at the 5% level

The number of students is indicated in brackets.

5.6 INFLUENCE OF BACKGROUND VARIABLES ON ENGLISH PROFICIENCY

In Section 5.1.2.3 above, we described the background details of the N.N.S. group taking the various Sessions and combinations of Sessions in the pre-test. Though it was thought unlikely that the A.E.B. would alter the format of T.E.A.P. on the basis of any findings about the effect of these background variables on proficiency we were nevertheless interested in establishing whether those variables had had a noticeable effect on student performance. This might enable us to make useful suggestions to receiving institutions.

As described above, with the exception of T041, all the tests listed below in Table 5KK will be part of the final version of T.E.A.P. For reference purposes we have detailed the maximum possible marks available on each of the tasks after re-weighting of the pre-test raw scores (v. p. 475).

TABLE 5KK						
MAXIMUM POSSIBLE SCORES ON TEST TASKS AFTER APPLICATION OF INTENDED FUTURE WEIGHTINGS						
	<u>Task</u>	<u>Maximum Possible Mark</u>	<u>Task</u>	<u>Maximum Possible Mark</u>	<u>Task</u>	<u>Maximum Possible Mark</u>
Reading	T012	46	TA12	42	TB12	42
			TA13	38	TB13	38
Writing	T011	21	TA31	21	TB31	21
	T032	21	TA32	9	TB32	9
Listening	T021	15	TA21	20	TB21	20
	T031	18				
Grammar*	T041	60				

* Not to be included in final version of T.E.A.P.

The N.S. and N.N.S. means and standard deviations with regard to each of the skill composites in the combinations of Sessions, were tabulated in Table 5Z, page 509 above.

In Appendix 5.14 we have recorded the means and standard deviations of various sub-groups within the population according to background variables we were able to collect information on through the Background Details Questionnaire. Although some of the comparison groups are small and some of the findings inconclusive, these data do shed some interesting light on the performance of certain sub-groups in the test tasks. The data are considered briefly below.

Age (v. Table 5LLLL, p. 1186)

It is interesting to note that it is the younger age groups that often achieve the highest scores in the various skill composites in the single Sessions and combinations of Sessions. This is perhaps due to the fact that these students, especially undergraduates, are often likely to have obtained qualifications in English language prior to embarking on their courses of study (cf. Chaplen 1970 and Morrison 1974).

There does not seem to be an increasing command of English commensurate with an increase in age for our test population.

Sex (v. Table 5MMMM, p.1187)

Without exception the females performed better as a group than the males in every skill and in every Session or combination of Sessions. As Davies (1965, p.205) pointed out, this is perhaps to be expected:

"... because education for women is at a premium in many countries which send large numbers of students to Britain and the women who achieve that education are accordingly highly selected."

The size of the standard deviations indicate that a number of the women could have serious problems, though, in general, these standard deviations are lower than those for the male groups.

Language Grouping (v. Table 5NNNN, p.1188)

On the whole, the Germanic group came out best on this evidence with the Semitic group doing particularly badly as compared with any other group, in all skills (v. Sen 1970). This finding indicates that institutions should be aware of the serious language problems that applicants from the Middle East Semitic group may have.

The wide ability range in lots of these other groupings indicates that some of the students even in these groups may need as much attention as those in the Semitic group appear to.

Length of Time Spent in Britain (v. Table 5PPPP, p.1190)

The group who had been in Britain less than two months performed second only to the group who had been here for over three years. It is possible the former group, who had only just arrived, had been exempted from pre-sessional language courses or introductory years due to their already proven language ability. Those who have been here for three years are better than the other three groups, particularly as regards listening comprehension. It would have been surprising if this had not been so. This finding might be interpreted therefore as indirect evidence of T.E.A.P.'s validity. The differences between the two middle groups are not as clear cut.

Amount of Time Spent Outside of Class with People who Speak
English (v. Table 5QQQQ, p.1191)

Again the expected pattern emerges. In general the greater amount of time spent outside of class with people who speak English, the higher the mean score obtained. This is particularly true of listening where one would expect increased interaction with native speakers to have an effect on listening and only to a lesser extent on reading. Again the findings might be taken as evidence of the validity of the test.

Number of Years Spent Learning English in Language Classes in
Their Own Countries (v. Table 5RRRR, p.1192)

No totally consistent patterning is in evidence here. One might tentatively suggest that the group which has spent 12+ years are, on

balance, likely to do better than those who have spent up to two years and in most cases better than those who have spent three to five years or six to eight years. The difference between the nine to eleven and twelve plus groups is not so regular. However, the large standard deviations in the twelve plus years group would advise caution as there is evidence that some candidates, even with a substantial number of years of being taught English in language classes abroad fall well below the N.N.S. mean for the whole population. One, of course, has no idea of the standards of language tuition in many of the countries our population originates from and this would clearly affect the relevance of this variable.

Number of Years Spent Learning English in Language Classes
in Britain (v. Table 5SSSS, p.1193)

The picture is more consistent here as one would have expected. Recognised language schools in Britain on the whole might be expected to provide more effective tuition than some schools abroad, if only in their use of trained N.S. as against N.N.S. speakers as teachers and the very presence of the candidates in an English speaking environment would obviously help in the process. At the risk of generalisation it seems from these data that the more language tuition the better the scores attained. Again this might be taken as evidence of the validity of our test.

Previous Experience of Learning a Subject Through
English (v. Table 5TTTT, p.1194)

On the whole those who had previously been taught a subject in English did better than those who had not. The exceptions to this all occur in the reading scores. It is perfectly conceivable that students may have been taught a subject in their native language and yet had to read texts in English. Thus, though their reading skills in English may have been developed, there may have been no need to write or listen in English.

Amount of Previous Education Conducted Through
English Medium (v. Table 5UUUU, p.1195)

The particularly low n's in many of the sub-groups make comparisons difficult here.

Previous Need to Read Books in Their Subject Areas
(v. Table 5VVVV, p.1196)

The n's in the 'never' column are too small to be worth considering. The pattern clearly emerges from the rest of these data that previous reading in English in the subject area has a positive effect on the scores achieved in reading, listening and writing, but that the effect is most noticeable on the reading sub-tests and may be adduced as evidence of the validity of the T.E.A.P. reading constructs.

To examine further what effect these variables were having on various skill composites, we calculated Pearson product moment correlation coefficients between background variables and test scores. The results of this exercise are tabulated in Table 5LL below.

The negative correlations between variable 5 (age and test scores) is to be explained by the fact that proficiency for our sample does not correlate at all well with age, although the statistically significant negative correlation for the TO + TB group is surprising and warrants further investigation. Indeed, the explanation for this may be found in the generally lower language ability of many of the post-graduates in our sample, many of whom were attending pre-session courses.

TABLE 5LL
 BACKGROUND VARIABLE CORRELATIONS WITH SKILL COMPONENT SCORES FOR READING, WRITING, LISTENING AND GRAMMAR

	1	2	3	4	5	6	7	8	9	10	11	12
1		.76	.72	.67	(-.06)*	.22	.08*	.40	.19	.37	.05*	.14*
2			.56	.60	(-.09)*	.23	.02*	.30	.06*	.24*	.09*	.16*
3				.69	.01*	.25	.19	.49	.31	.39	.29	.21
4					(-.21)	.39	.18	.40	.31	.43	.32	.22
Contingent Numbers	126	126	126	126	126	126	119	125	110	41	125	124
1		.77	.82	.68	(-.43)	.17*	.25	.19*	.01*	.52	.02*	.39
2			.73	.76	(-.46)	.24	.14*	.23	.05*	.31*	.04*	.23
3				.71	(-.47)	.04*	.20	.21	.24*	.59	.19*	.36
4					(-.56)	.14*	.20	.17*	.33	.38*	.16*	.26
Contingent Numbers	98	98	98	98	98	98	98	97	78	18	97	97

KEY: 1: Reading 5: Age 9: English language tuition in home
 2: Writing 6: Sex country
 3: Listening 7: Length of time spent in Britain 10: English language tuition in Britain
 4: Grammar 8: Amount of time spent outside of class with people who speak English 11: Previously studied subject in English
 12: Frequency with which read the subject in English

* Not significantly different from zero at the 5% level

Variables 8 (amount of time spent outside of class with people who speak English) and 10 (amount of language tuition in Britain) correlate slightly higher with task scores than do the other background variables for the TO + TA group, with variable 10 doing so also for the TO + TB group. As is clear from the summary data on means and standard deviations in Appendix 5.14, pages 1186-1196, the generally low correlations are probably accounted for, at least in part, by the imprecise nature of the data collected on these background variables. The comparatively high incidence of non-significant correlation coefficients suggests that this is so. The suggestion that many of these background variables are totally uncorrelated with language proficiency is implausible. Even those correlations which are significant offer only very slight evidence of the validity of T.E.A.P.

5.7 SETTING GRADE BOUNDARIES

5.7.1 Introduction

As mentioned earlier (p.508), once the test becomes operational it is intended to issue candidates' results in the form of a profile on each of the three study modes: reading, writing and listening. In each study mode candidates' performances are to be reported in terms of the five grades listed below.

- IV Proficiency in the study mode approaching that of native speaker tertiary level contemporaries; a limited number of weaknesses may be evident, but not sufficient to hamper academic progress seriously.

- III Moderate proficiency; some weaknesses which could affect performance in the study mode; some remedial language tuition would be helpful.

- II Limited proficiency; considerable weaknesses affecting performance in the study mode; some remedial language tuition is necessary.

- I Elementary language level; a large number of weaknesses are evident in performance in the study mode; these could seriously hamper academic progress; considerable remedial language tuition would probably be needed.

- 0 Beginner language level; almost no proficiency; cannot cope at all in the study mode; needs long-term language tuition before starting an academic course of study.

These grades were arrived at after extensive discussion with the Project Working Party and with groups of A.R.E.L.S. teachers. Methods of determining the grades are to be investigated and the grades themselves may be revised in the light of attempts to apply them operationally in the September 1983 piloting of the test.

The use of five grades implies that four boundaries (IV/III, III/II, II/I and I/0) must be fixed for each study mode. A number of ways of doing this are discussed below.

The investigation of the approaches to the problems of fixing grade boundaries was carried out in co-operation with Mr. M.J. Cresswell of the Board's Research and Statistics Division (v. Cresswell 1983). The work reported in this section owes a great deal to his efforts.

Two approaches that might be adopted for setting grade boundaries when T.E.A.P. becomes operational are described in detail. We could ask specialists in the field to make a posteriori judgements based on the performance of groups of candidates or to make a priori judgements about test content. We consider both of these approaches in turn.

5.7.2 Judgements about Groups of Candidates

This approach involves allocating candidates to the various grades using data other than their test scores and then determining the scores which, when used as grade boundaries, produce the lowest number of errors of classification. To establish the IV/III boundary, for example, we would look for the score which was exceeded by most candidates classified as grade IV according to the external criterion (v. Zieky et al. 1977).

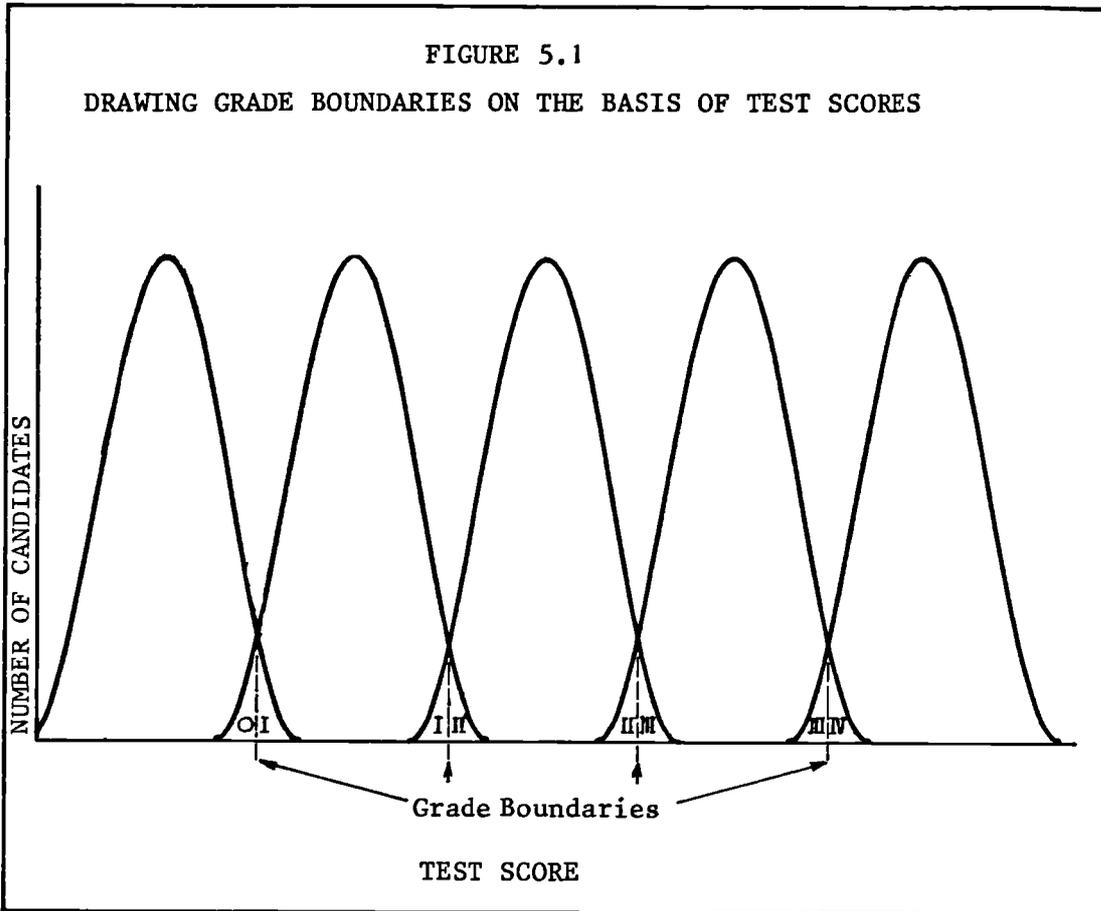
Two procedures could be employed for this purpose. First, a scrutiny of examination scripts could provide information on the basis of which candidates could be classified into grades. This is the usual method employed by examination boards to establish grade boundaries in G.C.E. 'O' and 'A' level examinations. Judges, however, may find it difficult to classify candidates into grades for proficiency in reading and listening by this method, because of the more discrete nature of the items of measurement that collectively are taken to constitute these abilities. Experience in the G.C.E. field suggests that judges find it easier to classify essay papers into grades than they do short answer or multiple choice papers. A second procedure that might be used in making judgements about groups of candidates

would be to employ completely independent data, such as teachers' estimates.

In the first of these methods candidates' scripts are examined by a group of judges who then use these to identify the grade boundaries. For example, in fixing the IV/III boundary the judges would identify two groups of scripts: in the first group minimally qualified candidates with a limited number of weaknesses which are almost, but not quite, "sufficient to hamper academic progress seriously" and in the second group candidates who just fell short of this standard. Cresswell (1983) pointed out that ideally the lowest score in the first group and the highest in the second would be adjacent marks (e.g. 65% and 64% respectively). In practice, though, this is unlikely to occur and the mid-point of a range of uncertainty over which judges might disagree between the two marks would normally be taken as the best estimate of the required grade boundary. For example, if the lowest mark in the grade IV group was 65% and the highest in the grade III group was 60%, then the IV/III boundary would be fixed at 63%/62%.

One final point concerning this method relates to the reference in the description of grade IV to native speakers. In the pre-test native speakers had a mean score of about 80% and a standard deviation of about 10% on each of the reading, writing and listening tests (v. Table 5Z, p.509). In determining the IV/III boundary, judges might therefore reasonably begin by considering scripts with less than about 70% of the available marks in each study mode. (Note that a IV/III boundary above 70% would have to be viewed with suspicion, whatever method gave rise to it, because a sizeable proportion of native speakers would thereby be given grade III.)

In the second method E.F.L. specialists associated with the students taking part in the piloting of the test will be asked to classify their candidates into the five grades according to the descriptions given earlier. The distributions of scores for the five groups of candidates will then be plotted and, ideally, would resemble Figure 5.1 below. Grade boundaries drawn at the points indicated would mean that, at each boundary, a minimum number of candidates may have been misclassified.



It had been hoped that analyses of this type could be successfully employed with the teachers' estimates gathered during the pre-tests, but the results proved to be disappointing for two reasons (v. Cresswell 1983). First, the sample sizes were very small and the teachers indicated that most students were in the upper grades. The result of these two factors was that little information was available about either the O/I or I/II boundary. Secondly, there was often substantial overlap between the distributions for adjacent grade groups and this implies a substantial number of misclassifications wherever the boundaries are positioned. The explanation for the overlaps between the distributions lies with the level of correlation found between the teachers' estimates and the test scores. The extent to which the observed correlations have been reduced by unreliability, either in the teachers' estimates or the test scores, cannot be evaluated. The equivocal results of this exercise are described in detail by Cresswell (1983). They are not reported here

because of their inconclusive nature.

We had also hoped to make use of subject tutors' estimates of the likelihood of students passing or failing their courses, but so few tutors thought their students would fail that we were unable to proceed with this line of enquiry.

5.7.3 Judgements of Test Content

Another possible approach we might adopt involves making direct judgements on test content. Cresswell (1983) describes several methods that are available for converting these judgements into scores on 'discrete' item tests, such as the short answer listening and reading comprehension tasks in T.E.A.P. (v. Angoff 1971). In the writing tasks in T.E.A.P., judges will have to be asked to make their evaluations directly in terms of the criteria for assessment.

Fundamental to this content-based method of setting grade boundaries is the notion of the minimally qualified candidate. A group of examiners will be asked to attempt to form a mental picture of the abilities and achievements of candidates who are just barely above each of the boundaries with which they are concerned. For example, the definition of the top grade (v. p.538 above) contains a reference to "a limited number of weaknesses ... but not sufficient to hamper academic progress seriously", they must decide when the "number of weaknesses" is almost, but not quite large enough to "hamper academic progress seriously". At that point, they are thinking of a minimally qualified candidate in the top grade.

The group of examiners will be asked to apply their conceptions of minimally qualified candidates to each item in the test in turn and to indicate how likely such a candidate is to answer each item correctly. The sum of the probabilities for all the items in a particular skill composite would then form the boundary for the particular grades concerned.

Since the extended writing tasks (TO11, TO32, TA31 and TB31) do not consist of a large number of discrete items, it will be necessary for the judges to work directly with the assessment criteria for this study mode. Each of the criteria is marked on a 0-3 scale (v. p.457 above) and the judges will be asked, for example, to indicate how many criteria would be met above the 2 level by candidates whose weaknesses were just "not sufficient to hamper academic progress". The answers to questions of this sort would define the grade boundaries.

Like the previous one, this procedure is fundamentally subjective and different judges are likely both to have different conceptions of the minimally qualified candidates and to interpret these differently in terms of the test material. In addition, it is likely that no individual judge will be entirely self-consistent in his own conception or interpretation of minimal qualifications. However, the aim is to use a number of judges, initially making a number of decisions independently of each other, so that the effect of these inconsistencies is diminished and an acceptable compromise emerges. A meeting will then be held to reconcile any major differences of opinion which may be present.

The use of norm-referenced methods of determining grade boundaries for T.E.A.P. were rejected at this stage on the grounds, inter alia, that maintaining the same standard would be difficult from test occasion to test occasion, notwithstanding the arbitrary nature of the initial decision about the proportion of candidates in the various grades (v. Houston 1983).

5.7.4 Maintaining Grade Boundaries

Cresswell (1983) has also pointed to the further problem of maintaining grade boundaries at comparable levels in future operational versions of T.E.A.P. So far, we have only addressed ourselves to the problem of establishing grade boundaries on a single occasion. As new versions of T.E.A.P. are to be administered every year, we have to ensure that the grade boundaries in successive years represent the same levels of performance. Statistical methods for equating the

scores on successive forms of the test could be employed, but this would normally involve a random sample of each year's candidates taking two forms of the test. Since each version of the test lasts about five hours, it seems unlikely that candidates and their teachers would be willing to be involved in an additional five hours of testing solely for the Board's equating purposes. However, because the test is to remain closed even after it has been administered, an alternative strategy may be possible.

Statistical equating techniques can be used when two forms of a test have common sections and with a closed test it would be possible to include certain tasks from earlier versions of the test. Statistical equating of the two forms would then be theoretically possible.

A third way of maintaining the equivalence of boundary performance levels across years is to ask judges to scrutinise test content, candidates' responses or both, in order to determine which scores on the new form of the tests are equivalent to the scores at the grade boundaries on the previous form. This is seen as being the principal means of maintaining the grade boundaries at fixed performance levels in the first few years of the test. It would be easier to apply judgements in this way rather than fix the boundaries anew every time by employing one of the methods for this described above, since the necessarily complex and highly subjective judgements involved may be easier to make on a comparative basis.

5.7.5 Future Grade Awarding

The final version of T.E.A.P. is to be piloted in late September 1983 which will present an ideal opportunity to establish grade boundaries. The methods discussed above are difficult to choose between on theoretical grounds and we need experience of applying them to T.E.A.P. before an informed choice can be made. It is proposed that we should try out each of the methods in this pilot examination so that the grade boundaries may be fixed as precisely as possible before the first operational administration of the test planned for May 1984. The onus on those determining grade boundaries for subsequent versions of T.E.A.P. would be to maintain the grade boundary performance levels established at that time.

C H A P T E R S I X

CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES

6. CONCLUSIONS AND SUGGESTIONS FOR FURTHER STUDIES

At the outset we identified three broad discipline areas in which the majority of N.N.S. students, coming to Britain to study at institutions of further and higher education, are likely to enrol from year to year. We examined the nature of the problems other writers had suggested these students faced and discussed the place of language proficiency in relation to other variables affecting academic progress. Though language was by no means the sole determinant of academic success, we argued that its importance justified attempts at producing an instrument to reveal any deficiencies the N.N.S. group might have in this respect.

We next examined the concepts underlying language test construction and emphasised our concern with the a priori validation of test tasks whilst recognising the need for external validation and the importance of making our tests as reliable as possible. Our concern was with the content validity of tests as well as with their prognostic value. It was our contention that a posteriori validation was a necessary but not sufficient procedure of test construction. A priori validation of test tasks especially in terms of a concern for content validity was seen as equally essential for testing within a more 'communicative' paradigm, especially as the test results were to be used for informing those responsible for remedial language teaching about the extent of any weaknesses in various macro-skills as well as enabling administrators to make decisions on the acceptability of candidates. We also felt that it was important that our test should have a suitable 'wash back' effect on language teaching, i.e. that any preparation of students for our test should reflect established good pedagogical practice in pre- and in-session E.A.P. courses.

After careful examination of the various approaches to language testing, we decided that a balanced, eclectic approach involving both 'discrete' and integrative focuses suited our purposes best. We took careful note of recent developments in communicative approaches to language testing and this influenced the design of our framework of categories for description of events and activities

relevant to the test population, our subsequent data collection procedures and our desire to ensure that an appropriate degree of context was evidenced in the tasks set in the test battery.

Very little is known about the language needs and difficulties of overseas students across discipline areas and levels. Most published information relates to very strictly delineated groups of students on particular courses. We on the other hand were interested in students in three broad discipline areas studying at a variety of levels. Given the limited resources at the disposal of those responsible for running remedial language classes and the heterogeneity of the students they have to deal with, remedial tutors have by necessity to attempt to cater for what is common in terms of activities and problems encountered among the more heavily populated disciplines and levels. Given that the organisers of these courses were likely to be major users of the results of our tests, we accordingly set up a framework for establishing E.A.P. language needs across levels and disciplines. We constructed data collection instruments for filling out this framework, namely the observation schedule and staff and student questionnaires. These instruments enabled us to develop an empirically based specification of language needs from which we could construct our Test in English for Academic Purposes (T.E.A.P.).

Whilst much of what we established could, perhaps, have been accomplished by sophisticated armchair speculation, we feel we have provided future research with a firmer empirical base on which to build than existed hitherto.

Our data collection, however, revealed areas where our test specification framework was extremely weak. The part of our framework that deals with the test dimensions of complexity and referential and functional range, exhibits the most serious deficiencies which it is hoped future research may improve upon. As a result, the extent to which we were successful in devising two part-versions of the test, TA and TB which were identical in what they sought to test and different only in the subject context in which it was tested, is open to doubt as the data analysis showed.

There is a need for more precise methods for dealing with task dimensions than the pragmatic ones used in this research. We relied heavily on the judgement of teachers and other experts in the field as well as on the results of small trial administrations to guide us on the appropriacy of task dimensions in the various constructs. Unless finer instruments are developed than these rather coarse subjective estimates, it is difficult to see how fully parallel versions of the test can ever be developed.

This may not be a realisable goal however. Even if we were to develop more suitable procedures for establishing these dimensions, the very complexity and sophistication of the measures required might rule out their use in the design of tests which are annual events. No examining board would contemplate lightly the extended research that might well be necessary to ensure that texts and, for that matter, the questions set upon them, are equivalent in terms of complexity and functional and referential range for a certain intended population, every time a paper is set. The preference will probably be, as in existing G.C.E. examinations, to rely on post examination procedures to rectify any imbalance from year to year.

In terms of that part of our framework entitled "dynamic communicative characteristics", we would again admit to deficiencies in realising these in T.E.A.P. because of the practical constraints which affect test construction. The chimera of full, communicative authenticity is just that. Without a grammar of language in use, it will never be possible to describe fully what authentic situations for a given population might be and even if this were ever devised, the artificiality of the test situation would remain.

This is not to say that a limited set of generalisable "dynamic communicative characteristics" of the type we have so far identified cannot be successfully incorporated into a test framework, thereby adding to the validity of the test operations.

It was only in terms of the general descriptive parameters that we came close to realising our communicative framework in T.E.A.P. We managed to incorporate reasonably representative activities, interactions, settings, instrumentality and enabling skills into our

operational test. However, practical constraints, e.g. time, money, the imagination of the test constructor and the need for reliability, will always temper how realistic we can make our test events. We nevertheless constructed a methodological framework and filled it out through empirical investigation in order to provide a bench mark against which T.E.A.P. and tests designed for similar purposes could be judged. We may neither wish to, nor be able to, realise the specification completely in a test but we needed a better picture of reality than was available to assess our own efforts and shortcomings and to guide us in test task construction. In a small way we hope we have made an initial contribution to establishing what it is that we should want to test and in so doing provided data collection instruments which future investigators can improve upon and expand through further empirical work on the coarse specification we have established.

It may be that no value is seen in attempting to paint such a broad canvas again and that any future effort is best put into investigating a small corner, e.g. a particular course at a particular institution. For certain situations this may be an acceptable viewpoint and our data collection instruments and methodological framework should still provide a useful starting point for these more specific investigations.

It does seem, however, that if E.A.P. pre-sessional and in-sessional courses are seen to be serving a useful purpose then attempts to improve the general description of what constitutes E.A.P. proficiency across disciplines and levels and to construct more valid and reliable measures for specifying a student's ability in regard to this proficiency are worthwhile.

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 E.S.P. tests therefore remains unproven. Given the small numbers involved in our investigation we would, however, advise further research through controlled experimentation on the effect on student performance, of employing texts different from the subject areas they are studying in, as stimuli for test tasks.

As regards test procedures we have not produced any novel panaceas. We have shown how various formats may be utilised for focusing on the testing of certain enabling skills. We have also introduced the idea of more communicative, integrated activities into a test where listening and/or reading texts, as well as providing a stimulus for more discrete testing of skills, are also employed as the stimulus material for a writing test. What is clear is that 'communicative tests' will be more difficult to construct than traditional measures. As well as the need for greater explicitness about what it is that one is trying to test, there are serious problems in successfully realising these specifications in test form and in devising suitable assessment procedures. As with all new departures, integrated, communicative tests are at present difficult to construct, complex to take, difficult to mark and difficult to report results on.

It is felt, however, that the methodological approach we have advocated in this work will help to ensure a greater degree of content and face validity for future E.A.P. tests conceived within this paradigm. The a priori validation of test tasks is considered as a first, but nevertheless an essential, step in test task construction.

In retrospect we would argue that the evidence produced in the pre-test data has also established some validity for our tests. The performance of the N.S. group and the difference in the task means between N.S. and N.N.S. groups and the validity of individual items argue for the content validity of T.E.A.P. We have shown that

there is a significant difference between the English proficiency of overseas students in Britain and that of comparable native English students. The factorial groupings also provide slight evidence that our tests are measuring constructs in accordance with the labels we had given them.

The limited amount of concurrent and predictive validity established so far for T.E.A.P. is satisfying, but we realise the need for more extensive, external, validation studies and it is hoped that these will be conducted over the next few years by the examining board working in concert with selected universities. The accumulation of test statistics of all kinds and their storage on a computer data base over the years from the first proper administration of the test, should also aid future research and development of the test.

We have provided some evidence that background variables, especially length of stay in Britain, amount of time spent outside of class with English speakers, amount of language tuition in Britain and language group, can have an influence on English proficiency. We have suggested that indirectly some of these data also lend support for the validity of T.E.A.P.

By employing a balance of 'discrete point', integrative and integrated tasks in our test battery, we feel that both the validity and reliability of our proficiency test was enhanced. The reliability coefficients obtained from the reading and listening composites were quite encouraging, though those for writing are in need of some improvement. A considerable degree of context can be built into a test if due care is taken in observing the a priori validation procedures we have designed and followed through. Whilst 'authenticity' is an unattainable goal in language testing a 'realistic' context is not. We have demonstrated that we no longer have to rely solely on a test of linguistic competence which, though reliable, may be lacking in validity. We believe that future research and development along the lines we have established will provide ways of building even more validity into our test measures without any unacceptable loss of reliability or efficiency.

We have shown that integrative measures such as cloze and dictation, though reliable and indirectly valid measures in their own right, are not on their own sufficient indicators of a student's E.A.P. proficiency. Factor analysis demonstrated that, though there is one main factor underlying performance in our tests, it does leave a lot of the variance unaccounted for. The Varimax rotations provided some slight evidence that tasks of comprehension and production exhibit different factor loadings. The research reported lends some support to the limited divisibility hypothesis of language proficiency.

As regards more specific aspects of test construction, the value of the multiple choice format for testing reading comprehension was brought into question. Poor correlational data from both internal and external analysis led us to drop multiple choice tests of reading comprehension from our final battery. Practical difficulty in constructing items and the lower content validity of the format also encouraged us in this direction. Limitation of time and resources had prevented us from investigating the effect on student performance of using different formats in the various skill areas, but it is to be hoped that future research will examine the issue of format effect more carefully. It is a complex, vital, but unfortunately neglected area of test construction.

The writing components of the battery did not work as well as we had hoped. Great care was taken in developing valid, analytic criteria for the purpose of rating the writing tasks. Though this helped achieve quite reasonable standards of intra-marker consistency, we feel that the inter-marker figures quoted are still capable of improvement. Either an attempt should be made to increase the reliability of marking by ensuring greater homogeneity of background in the markers and putting them through more rigorous standardisation procedures, or serious consideration should be given to the double marking of writing tasks.

There is some evidence in the factor analysis that, owing to the integrated nature of some of the tasks where reading and/or listening feed into writing, performance on the latter is to a

certain extent influenced by proficiency in the other skills. The writing scores might, therefore, be contaminated by previous performance on listening and/or reading tasks. As these integrated measures reflect the situation students are likely to face in the academic context, we felt that this was acceptable. However, we do feel that it might be advisable in the future research to be carried out on the test to compare candidates' performance on these existing, integrated writing tasks with performance on extended writing tasks which are not linked in this way to other activities. These non-integrated tasks would obviously have less face and content validity than the ones presently employed, but the exercise might serve to demonstrate whether there were any candidates who were being disadvantaged by the present scheme.

We were aware that a potential problem in adopting an integrated approach might be in determining where any breakdown in the process had occurred. For this reason we included 'discrete' tests of reading and listening comprehension in the battery. This was to enable us to examine whether a candidate experienced problems in either of these areas and to report on this in the profile. The question remains for future investigation as to whether a candidate who performs badly in these two areas might nevertheless perform satisfactorily on writing tasks if they were of the non-integrated variety.

We are still not completely satisfied with many of the test tasks in T.E.A.P. and think that they can be improved upon. We are perhaps dissatisfied though because we are better able to see their failings in the light of our specification and the experience gained in establishing it. The data analyses of the pre-test results are far from discouraging.

The present version of T.E.A.P. is certainly as valid, if not more so, than all available alternative tests in terms of our content specification and the reliability figures we quote for the composite task scores are encouraging, though it is felt that both could be improved upon. We would claim to have provided a solid base of

empirical data for the construction of future operational versions of T.E.A.P. which would provide, by means of individual profiles, information on students' understanding and use of written and spoken English in academic situations.

B I B L I O G R A P H Y

BIBLIOGRAPHY

- ADAMS, M.L. and FRITH, J.R. (eds.) (1979). Testing Kit:- French and Spanish. Foreign Service Institute, U.S. State Department.
- ADAMS, R. (1981). 'The Reliability of Marking of Five June 1980 Examinations.' Mimeo. Associated Examining Board, Aldershot.
- AIKEN, E.G., THOMAS, G.S. and SHENNUM, W.A. (1975). 'Memory for a Lecture: Effects of Notes, Lecture Rate and Informational Density.' Journal of Educational Psychology 67.
- AITKEN, K. (1979). 'Techniques for Assessing Listening Comprehension in Second Languages.' Audio-Visual Language Journal 17/3.
- ALBERT, T. (1976). 'If You Were Foreign.' T.H.E.S. 221.
- ALDERMAN, D.L. and HOLLAND, P.W. (1981). Item Performance across Native Language Groups on the Test of English as a Foreign Language. T.O.E.F.L. Research Report 9.
- ALDERSON, J.C. (1978a). 'A Study of the Cloze Procedure with Native and Non-Native Speakers of English.' University of Edinburgh Ph.D. Thesis.
- ALDERSON, J.C. (1978b). 'The Effect of Certain Methodological Variables on Cloze Test Performance and its Implication for the Use of the Cloze Procedure in E.F.L. Testing.' Mimeo. 5th International Congress of Applied Linguistics, Montreal.
- ALDERSON, J.C. (1981a). 'Reaction to the Morrow Paper (3).' in ALDERSON and HUGHES (eds.) 1981.
- ALDERSON, J.C. (1981b). 'Report of the Discussion on Communicative Language Testing.' in ALDERSON and HUGHES (eds.) 1981.
- ALDERSON, J.C. (1981c). 'Report of the Discussion on Testing English for Specific Purposes.' in ALDERSON and HUGHES (eds.) 1981.
- ALDERSON, J.C. and HUGHES, A. (eds.) (1981). E.L.T. Documents 111 - Issues in Language Testing. The British Council.
- ALDERSON, J.C. and URQUHART, A.H. (1982). 'The Effect of Student Background Discipline on Comprehension: A Pilot Study.' Mimeo. University of Lancaster.
- ALDERSON, J.C. and WATERS, A. (1983). A Course in Testing and Evaluation for E.S.P. Teachers or 'How Bad Were My Tests?' Mimeo. University of Lancaster.
- ALEXANDER, L.G. (1971). Guided Composition Writing. Longman.

- ALEXANDER, L.G., STANNARD-ALLEN, W., CLOSE, R.A. and O'NEILL, R.J. (1975). English Grammatical Structure. Longman.
- ALLEN, J.P.B. and CORDER, S.P. (eds.) (1974). 'Techniques in Applied Linguistics.' The Edinburgh Course in Applied Linguistics 3.
- ALLEN, J.P.B. and DAVIES, A. (eds.) (1977). 'Testing and Experimental Methods.' The Edinburgh Course in Applied Linguistics 4.
- ALLEN, J.P.B. and WIDDOWSON, H.G. (1974a). 'Teaching the Communicative Use of English.' I.R.A.L. 12/1.
- ALLEN, J.P.B. and WIDDOWSON, H.G. (1974b). English in Physical Science. Oxford University Press.
- ALLWRIGHT, J. and ALLWRIGHT, R. (1977). 'An Approach to the Teaching of Medical English.' in HOLDEN (ed.) 1977.
- ALTMAN, H. (1980). 'Foreign Language Teaching: Focus on the Learner.' in ALTMAN and JAMES (eds.) 1980.
- ALTMAN, H. and JAMES, C. (eds.) (1980). Foreign Language Learning: Meeting Individual Needs. Pergamon.
- ANASTASI, A. (1982). Psychological Testing (5th edition). Macmillan: New York.
- ANDERSON, D.F. (1953). 'Tests of Achievement in the English Language.' English Language Teaching 7/2.
- ANGOFF, W.H. (1971). 'Scales, Norms and Equivalent Scores.' in THORNDIKE (ed.) 1971.
- ARCHER, M. and NOLAN-WOODS, E. (1976). Practice Tests for Proficiency. Nelson.
- ASSESSMENT OF PERFORMANCE UNIT (A.P.U.) (1983). Language Performance in Schools: Secondary Survey Report 2. H.M.S.O.
- AUSTIN-WARD, B.J. (1979). 'An Enquiry into the Attitudes and Abilities of Entrants to Colleges of Further Education and Technical Colleges in Regard to Use and Study of English Language.' University of London Ph.D. Thesis.
- BACHMAN, L.F. and PALMER, A.S. (1981a). 'Basic Concerns in Test Validation.' in READ (ed.) 1981a.
- BACHMAN, L.F. and PALMER, A.S. (1981b). 'The Construct Validation of the F.S.I. Oral Interview.' Language Learning 31/1. University of Michigan.
- BARNES, D. and SEED, J. (1981). Seals of Approval: An Analysis of English Examinations at Sixteen Plus. University of Leeds, School of Education.

- BARRETT, T.C. (1972). 'Taxonomy of the Cognitive and Affective Dimensions of Reading Comprehension.' Table in CLYMER, T. 'What is Reading? - Some Current Concepts.' in MELNIK AND MERRITT (eds.) 1972.
- BARTOLIC, L. (1975). 'Technical English: A Method of Teaching the Cause-Effect Relation as Applied to a Diagram.' E.L.T. Journal 29/2.
- BATES, M. and DUDLEY-EVANS, A. (1974). 'Notes on the Introductory English Courses for Students of Science and Technology at the University of Tabriz, Iran.' E.L.T. Documents 4.
- BATES, M. and DUDLEY-EVANS, A. (1976). General Science. Longman.
- BEARD, R. (1970). Teaching and Learning in Higher Education. Penguin.
- BEARDSMORE, H.B. (1974). 'Testing Oral Fluency.' I.R.A.L. 12/4.
- BEIGHLEY, K.C. (1954). 'An Experimental Study of the Effect of Three Speech Variables on Listener Comprehension.' Speech Monographs 21.
- BENSOUSSAN, M. (1982). 'Testing the Test of Advanced E.F.L. Reading Comprehension: To What Extent Does the Difficulty of a Multiple-Choice Comprehension Test Reflect the Difficulty of the Text?' System 10/3. Pergamon.
- BERETTA, A. (1983). 'A Comparison of Three Tests of Listening in the Context of English for Academic Purposes.' University of Edinburgh M.Sc. Applied Linguistics Assignment.
- BEST, J. and ILYIN, D. (1976). Structure Tests - English Language Advanced Forms 1 and 2. Newbury House.
- BLACK, C. (1971). 'Graded Practice in Advanced Listening Comprehension.' English Language Teaching XXV/2.
- BLIGH, D.A. (1972). What's the Use of Lectures? Penguin.
- BLUE, G. (1979). 'Diagnostic English Language Test.' Mimeo. Southampton University Language Centre.
- BORMUTH, J.R. (1963). 'Cloze as a Measure of Readability.' Yearbook of the International Reading Association 8.
- BORMUTH, J.R. (1970). On the Theory of Achievement Test Items. University of Chicago Press.

- BORMUTH, J.R., MANNING, J., CARR, J. and PEARSON, D. (1970). 'Children's Comprehension of Between and Within Sentence Syntactic Structures.' Journal of Educational Psychology 61/5.
- BREW, A. (1980). 'Responses of Overseas Students to Differing Teaching Styles.' in GREENALL and PRICE (eds.) 1980.
- BRIGHT, J.A. and MCGREGOR, G.P. (1970). Teaching English as a Second Language. Longman.
- BRISTOW, R. (1977). Trends and Developments 1974/78. Coombe Lodge Reports 10/6.
- BRITISH COUNCIL (1975). English for Academic Study with Special Reference to Science and Technology. Problems and Perspectives. E.T.I.C. Occasional Paper. British Council.
- BRITISH COUNCIL (1978a). Statistics of Overseas Students in Britain 1976/1977. British Council.
- BRITISH COUNCIL (1978b). English for Specific Purposes. E.L.T. Documents 101. British Council.
- BRITISH COUNCIL (1978c). Pre-sessional Courses for Overseas Students. E.T.I.C. Occasional Paper. British Council.
- BRITISH COUNCIL (1979). Statistics of Overseas Students in Britain 1977/1978. British Council.
- BRITISH COUNCIL (1980a). Statistics of Overseas Students in Britain 1978/1979. British Council.
- BRITISH COUNCIL (1980b). Projects in Materials Design. British Council.
- BRITISH COUNCIL (1981a). Statistics of Overseas Students in Britain 1979/1980. British Council.
- BRITISH COUNCIL (1981b). Survey of the Factors Affecting the Performance of O.D.A.-Sponsored Study Fellows. English Language Division, British Council.
- BRITISH COUNCIL (1981c). The Teaching of Listening Comprehension. E.L.T. Documents Special. Central Information Service, English Language Division, British Council.
- BRITISH COUNCIL (1982). Statistics of Overseas Students in Britain 1980/1981. British Council.
- BRITTON, J.N. (1963). 'Experimental Marking of English Compositions Written by Fifteen Year Olds.' Educational Review 16/1.

- BRITTON, J.N., MARTIN, N.C. and ROSEN, H. (1966). Multiple Marking of English Compositions: An Account of an Experiment. Schools Council Examinations Bulletin 12. H.M.S.O.
- BROOKS, V. (1980). Improving the Reliability of Essay Marking. A Survey of the Literature with Particular Reference to the English Language Composition. University of Leicester.
- BROWN, G. (1977). Listening to Spoken English. Longman.
- BROWN, G. (1978). Lecturing and Explaining. Methuen.
- BROWN, J.D. (1979). 'A Correlational Study of Four Methods for Scoring Cloze Tests.' Paper presented at the 1977 T.E.S.O.L. Convention, Boston.
- BRUMFIT, C.J. (1977). 'The Teaching of Advanced Reading Skills in Foreign Languages, with Particular Reference to E.F.L.' Language Teaching and Linguistics: Abstracts 10/2.
- BRUMFIT, C.J. (1978). 'The Teaching of Advanced Reading Skills in Foreign Languages, with Particular Reference to English as a Foreign Language.' in KINSELLA (ed.) 1978.
- BRUMFIT, C.J. (1980). 'From Defining to Designing: Communicative Specifications Versus Communicative Methodology in Foreign Language Teaching.' Mimeo. University of London, Institute of Education.
- BRUMFIT, C.J. and JOHNSON, K. (eds.) (1979). The Communicative Approach to Language Teaching. Oxford University Press.
- BRUNER, J.S. (1975). 'Language as an Instrument of Thought.' in DAVIES (ed.) 1975.
- BURNS, D.G. (ed.) (1965). Travelling Scholars: An Enquiry into the Adjustment and Attitudes of Overseas Students holding Commonwealth Bursaries in England and Wales. National Foundation for Educational Research.
- BURSTALL, C. (1975). 'Factors Affecting Foreign Language Learning: A Consideration of some Recent Research Findings.' Language Teaching and Linguistics: Abstracts 8/1.
- BYRNE, D. (1979). Teaching Writing Skills. Longman.
- CAMPBELL, R. and WALES, R. (1970). 'The Study of Language Acquisition.' in LYONS (ed.) 1970.
- CAMPBELL, V. (1974). Pilot Report on the Communication Problem of Overseas Students in British Technical Education. Mimeo. University of London.

- CANALE, M. (1982). 'Evaluating the Coherence of Students Writing in L1 and L2.' Paper presented at the 1982 T.E.S.O.L. Convention, University of Hawaii.
- CANALE, M. and SWAIN, M. (1980). 'Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing.' Journal of Applied Linguistics 1/1.
- CANDLIN, C.N. (1977). Questionnaire on the English Language Problems Faced by Overseas Students Studying in Britain. Mimeo. University of Lancaster, Institute of English Language Education.
- CANDLIN, C.N., KIRKWOOD, J.M. and MOORE, H.M. (1975). 'Developing Study Skills in English.' in BRITISH COUNCIL 1975.
- CANDLIN, C.N., KIRKWOOD, J.M. and MOORE, H.M. (1978). 'Study Skills in English: Theoretical Issues and Practical Problems.' in MACKAY and MOUNTFORD (eds.) 1978.
- CANDLIN, C.N. and MURPHY, D.F. (1976). Engineering Discourse and Listening Comprehension. Mimeo. University of Lancaster.
- CARROLL, B.J. (1978a). 'An English Language Testing Service: Specifications.' British Council.
- CARROLL, B.J. (1978b). Guidelines for the Development of Communicative Tests. Mimeo. The Royal Society of Arts.
- CARROLL, B.J. (1980). Testing Communicative Performance: An Interim Study. Pergamon.
- CARROLL, B.J. (1982). 'Language Testing - Is There Another Way?' in HEATON (ed.) 1982.
- CARROLL, J.B. (1961). 'Fundamental Considerations in Testing for English Language Proficiency of Foreign Students.' Testing the English Proficiency of Foreign Students. Center for Applied Linguistics, Washington D.C.
- CARROLL, J.B. (1968). 'The Psychology of Language Testing.' in DAVIES (ed.) 1968.
- CARROLL, J.B. (1972). 'Defining Language Comprehension: Some Speculations.' in CARROLL and FREEDLE (eds.) 1972.
- CARROLL, J.B. (1973). 'Foreign Language Testing: Will the Persistent Problems Persist?' in CONCANNON-O'BRIEN (ed.) 1973.
- CARROLL, J.B. (1979). 'Psychometric Approaches to the Study of Language Abilities.' in FILLMORE, KEMPLER and WANG (eds.) 1979.
- CARROLL, J.B. and FREEDLE, R.O. (eds.) (1972). Language Comprehension and the Acquisition of Knowledge. N.Y: Halsted.

- CAST, B.M.D. (1939). 'The Efficiency of Different Methods of Marking English Composition.' British Journal of Educational Psychology 9/1.
- CAST, B.M.D. (1940). 'The Efficiency of Different Methods of Marking English Composition.' British Journal of Educational Psychology 10/2.
- CHAPLEN, E.F. (1970). 'The Identification of Non-Native Speakers of English Likely to Under-Achieve in University Courses Through Inadequate Command of the Language.' University of Manchester Ph.D. Thesis.
- CHAPLEN, E.F. (1971). 'The Reliability of the Essay Sub-Test in a University Entrance Test in English for Non-Native Speakers of English.' in PERREN and TRIM (eds.) 1971.
- CHAPLEN, E.F. (1976). 'Measuring Achievement in Adult Language Learning.' Committee for Out-of-School Education and Cultural Development. Strasbourg.
- CHAPLEN, E.F. (1982). 'Measuring Student Achievement: Some Practical Considerations.' in HEATON (ed.) 1982a.
- CHEUNG, V.K.C. (1978). 'Testing the English Language Proficiency of Overseas Post-Graduate Students - A Tentative Model.' University of Exeter M.A. Thesis.
- CLAPHAM, C.M. (1981). 'Reaction to the Carroll Paper (1).' in ALDERSON and HUGHES (eds.) 1981.
- CLARK, J.L.D. (1975). 'Theoretical and Technical Considerations in Oral Proficiency Testing.' in JONES and SPOLSKY (eds.) 1975.
- CLOSE, R.A. (1965). The English We Use for Science. Longman.
- CLYMER, T. (1972). 'What is "Reading"?: Some Current Concepts.' in MELNIK and MERRITT (eds.) 1972.
- COFFMAN, W.E. (1971). 'Essay Examinations.' in THORNDIKE (ed.) 1971.
- COFFMAN, W.E. and KURFMAN, D. (1968). 'A Comparison of Two Methods of Reading Essay Examinations.' American Educational Research Journal 5/1.
- CONCANNON-O'BRIEN, M. (ed.) (1973). Testing in Second Language Teaching: New Dimensions. A.T.E.S.O.L., Ireland.
- COOPER, C.R. (1975). 'Measuring Growth in Writing.' English Journal 64.
- COOPER, C.R. (1977). 'Holistic Evaluation of Writing.' in COOPER and ODELL (eds.) 1977.

- COOPER, C.R. and ODELL, L. (eds.) (1977). Evaluating Writing: Describing, Measuring, Judging. Urbana Illinois: National Council of Teachers of English.
- COOPER, R.L. (1968). 'An Elaborated Language Testing Model.' in UPSHUR and FATA (eds.) 1968.
- CORDER, S.P. (1973). Introducing Applied Linguistics. Penguin.
- COULTHARD, M. (1975). 'Discourse Analysis in English - a Short Review of the Literature.' Language Teaching and Linguistics: Abstracts 8/2.
- COWIE, A.P. and HEATON, J.B. (eds.) (1977a). English for Academic Purposes. University of Reading: B.A.A.L./S.E.L.M.O.U.S. Publication.
- COWIE, A.P. and HEATON, J.B. (1977b). 'Preparing a Writing Programme for Students of Science and Technology.' in COWIE and HEATON (eds.) 1977a.
- CRESSWELL, M.J. (1983). 'T.E.A.P.: Setting the Grade Boundaries.' Mimeo. Associated Examining Board.
- CROFTS, J. (1976). 'Review of JUPP, T.C. et al. (1975).' E.S.P.M.E.N.A. Bulletin 5.
- CRONBACH, L.J. (1970). Essentials of Psychological Testing. N.Y.: Harper and Row.
- CRONBACH, L.J. (1971). 'Test Validation.' in THORNDIKE (ed.) 1971.
- CRYSTAL, D. and DAVY, D. (1975). Advanced Conversational English. Longman.
- CULHANE, T., KLEIN-BRALEY, C. and STEVENSON, D.K. (eds.) (1982). Practice and Problems in Language Testing. University of Essex, Department of Language and Linguistics, Occasional Papers 26.
- CUMMINS, J. (1980). 'The Cross-Lingual Dimensions of Language Proficiency: Implications for Bi-Lingual Education and the Optimal Age Issue.' T.E.S.O.L. Quarterly 14/2.
- CZIKO, G.A. (1981). 'Psychometric and Edumetric Approaches to Language Testing: Implications and Applications.' Applied Linguistics II/1.
- CZIKO, G.A. (1982). 'Improving the Psychometric, Criterion-Referenced and Practical Qualities of Integrative Language Tests.' T.E.S.O.L. Quarterly 15.
- DAKIN, J. (1969). 'The Teaching of Reading.' in FRASER and O'DONNELL (eds.) 1969.

- DANIEL, N. (1975). The Cultural Barrier. Edinburgh University Press.
- DARNELL, D.K. (1968). The Development of an English Language Proficiency Test of Foreign Students Using a Clozentropy Procedure. Boulder Co., Department of Speech and Drama, University of Colorado.
- DARNELL, D.K. (1970). 'Clozentropy: A Procedure for Testing English Language Proficiency of Foreign Students.' Speech Monographs 37/1.
- DAVIES, A. (1965). 'Proficiency in English as a Second Language.' University of Birmingham Ph.D. Thesis.
- DAVIES, A. (1967). 'The English Proficiency of Overseas Students.' British Journal of Educational Psychology 37/2.
- DAVIES, A. (ed.) (1968). Language Testing Symposium. A Psycholinguistic Approach. Oxford University Press.
- DAVIES, A. (1969). 'The Notion of Register.' Educational Review 22.
- DAVIES, A. (1973). 'Language Proficiency Testing and the Syllabus.' in CONCANNON-O'BRIEN (ed.) 1973.
- DAVIES, A. (ed.) (1975). Problems of Language Learning. Heinemann.
- DAVIES, A. (1977a). 'Do Foreign Students Have Problems?' in COWIE and HEATON (eds.) 1977a.
- DAVIES, A. (1977b). 'The Construction of Language Tests.' in ALLEN and DAVIES (eds.) 1977.
- DAVIES, A. (1978). 'Language Testing.' Survey Article Parts I and II. Language Teaching and Linguistics: Abstracts 11 3/4.
- DAVIES, A. (1981a). 'A Review of Communicative Syllabus Design.' T.E.S.O.L. Quarterly 15/3.
- DAVIES, A. (1981b). 'Reaction to the Palmer and Bachman and the Vollmer Papers (2).' in ALDERSON and HUGHES (eds.) 1981.
- DAVIES, A. (1982). 'Criteria for Evaluation of Tests of English as a Foreign Language.' in HEATON (ed.) 1982a.
- DAVIES, A. (forthcoming 1983). 'The Validity of Concurrent Validation.' Paper presented at the B.A.A.L. Seminar, University of Reading, December 1981. To appear in HUGHES and PORTER (eds.) 1983.
- DAVIES, A. and ALDERSON, J.C. (1977). English Proficiency Test Battery. Short Version Form D. British Council.

- DAVIES, A., MOLLER, A.D. and ADAMSON, D.M. (1975). The English Proficiency of Foreign Students in Higher Education (Non-university) in Scotland. A Report to the Scottish Education Department. Department of Linguistics, University of Edinburgh.
- DAVIES, A. and WIDDOWSON, H.G. (1974). 'Reading and Writing.' in ALLEN and CORDER (eds.) 1974.
- DAVIS, F.B. (1944). 'Fundamental Factors of Comprehension in Reading.' Psychometrika 9.
- de GREVE, M. et al. (eds.) (1973). Modern Language Teaching to Adults: Language for Special Purposes. Second A.I.M.A.V. Seminar 1972. Brussels: A.I.M.A.V. Paris: Didier.
- DEPARTMENT OF EDUCATION AND SCIENCE (D.E.S.) (1975). Language for Life. Report of the Committee of Inquiry appointed by the Secretary of State for Education and Science under the Chairmanship of Sir Alan Bullock, F.B.A.
- DIEDERICH, P.B. (1974). Measuring Growth in English. Urbana, Illinois: National Council of Teachers of English.
- DINNING, U. (1977). 'An Experiment on the Memory of Discourse.' University of Lancaster M.A. Thesis.
- DUDLEY-EVANS, A. (1977). 'Planning a Course for Science and Engineering Students.' in HOLDEN (ed.) 1977.
- DUDLEY-EVANS, A. (1978). 'Report Writing.' in BRITISH COUNCIL 1978c.
- DUDLEY-EVANS, A. (1980). 'The Overseas Student and Intercommunication Needs.' in HALL (ed.) 1980.
- DUDLEY-EVANS, A. and SWALES, J. (1980). 'Study Modes and Students from the Middle East.' in GREENALL and PRICE (eds.) 1980.
- DUDLEY-EVANS, A. and JOHNS, T.F. (1981). 'A Team Teaching Approach to Lecture Comprehension for Overseas Students.' in BRITISH COUNCIL 1981c.
- DUNLOP, F. (1966). Europe's Guests, Students and Trainees: A Survey on the Welfare of Foreign Students and Trainees in Europe. Council of Cultural Co-operation of the Council of Europe.
- DUNLOP, I. (1969). 'Tests of Writing Ability in E.F.L.' English Language Teaching 24/1.
- EBEL, R.L. (1972). 'Why is a Longer Test Usually a More Reliable Test?' Educational and Psychological Measurement 32.

- EDWARDS, P.J. (1978). 'The Problems of Communication Facing Overseas Nurses in Training in England and Wales.' University of London Ph.D. Thesis.
- EDGEWORTH, F.Y. (1888). 'The Statistics of Examinations.' Journal of the Royal Statistical Society 11/III.
- EGGLESTONE, J.F., GALTON, M.J. and JONES, M.E. (1975). A Science Teaching Observation Schedule. Schools Council Research Studies. MacMillan Educational.
- ENGELHART, M.B. (1972). Methods of Educational Research. Rand McNally: Chicago.
- ENGINEER, W.D. (1977). 'An Investigation of a Reading Model for English as a Second Language.' University of Edinburgh Ph.D. Thesis.
- ESKEY, D.E. (1973). 'A Model Program for Teaching Reading to Advanced Students of English as a Foreign Language.' Language Learning 23/2.
- EWER, J.R. and LATORRE, G. (1967). 'Preparing an English Course for Students of Science.' English Language Teaching 21/3.
- EWER, J.R. and LATORRE, G. (1969). A Course in Basic Scientific English. Longman.
- FARHADY, H. (1979). 'The Disjunctive Fallacy between Discrete-Point and Integrative Tests.' T.E.S.O.L. Quarterly 13/3.
- FERGUSON, N. and O'REILLY, M. (1977). Listening and Note-Taking. Evans Brothers Ltd.
- FILLMORE, C.J., KEMPLER, D. and WANG, W.S.Y. (eds.) (1979). Individual Differences in Language Ability and Language Behaviour. Academic Press.
- FINLAYSON, D.L. (1951). 'The Reliability of the Marking of Essays.' British Journal of Educational Psychology 21/2.
- FISHER, M.S.S. and HUBBARD, R. (1979). 'The Reliability of R.S.A., F.C.E. and A.R.E.L.S. Oral Examinations.' University of Birmingham B.Phil. Ed. Thesis.
- FISHMAN, J.A. (1969). 'The Sociology of Language.' in GIGLIOLI (ed.) 1976.
- FISHMAN, J.A. (1972). Language in Sociocultural Change. California: Stanford University Press.

- FITCH, R.K. (1980). 'Experiences of a Post-Graduate Physics Course Tutor with Overseas Students.' in GREENALL and PRICE (eds.) 1980.
- FOK, C.Y.Y.A. (1981). 'The Testing of Listening Proficiency at the Tertiary Level.' in READ (ed.) 1981b.
- FOK, A., LORD, R., LOW, G. and T'SOU, B.K. (eds.) (1981). Language Testing. Working Papers in Linguistics and Language Teaching 4. Language Centre of University of Hong Kong.
- FORTUNE, A. (1977). 'Making Courses in Science and Technology More Relevant to the Needs of the Student.' in HOLDEN (ed.) 1977.
- FOWLER, W.S. and COE, N. (1976). Nelson English Language Tests. Book 3 Advanced. Nelson.
- FRANCIS, J.C. (1977). 'Impression and Analytic Marking Methods.' Mimeo. Associated Examining Board.
- FRASER, H. and O'DONNELL, W.R. (eds.) (1969). Applied Linguistics and the Teaching of English. Longman.
- FRY, D.B. (1963). Teaching Faster Reading - A Manual. Cambridge University Press.
- FURTHER EDUCATION STAFF COLLEGE (1971). Study Conference 70/50A. Overseas Students in F.E. Colleges. Coombe Lodge Report 4/4.
- FURNEAUX, C. (1982). 'Comparing Listening Recall and Dictation as Possible Sub-tests for Inclusion in the A.E.B.'s Test in English for Academic Purposes.' University of Reading M.A. Thesis.
- GALTON, M. (1979). 'Systematic Classroom Observation: British Research.' Educational Research 21/2.
- GARDNER, R.C. and LAMBERT, W.E. (1969). 'Motivational Variables in Second Language Acquisition.' Canadian Journal of Psychology 13
- GARDNER, R.C. and LAMBERT, W.E. (1972). Attitudes and Motivation in Second Language Learning. Newbury House.
- GARWOOD, C.H. (1970a). 'The Teaching of English to the Non-English Speaking Technical Student. 1: The Context of Situation.' English Language Teaching 24/2.
- GARWOOD, C.H. (1970b). 'The Teaching of English to the Non-English Speaking Technical Student. 2: Relating the Structures to the Contexts of Situation.' English Language Teaching 24/3.
- GEDDES, M. (1977). 'Reference Skills.' in HOLDEN (ed.) 1977.

- GIGLIOLI, P.P. (ed.) (1976). Language and Social Context. Penguin.
- GIPPS, C. and EWEN, E. (1974). 'Scoring Written Work in English as a Second Language: the Use of the T-Unit.' Educational Research 16/2.
- GODSHALK, F.I., SWINEFORD, F. and COFFMAN, W.E. (1966). The Measurement of Writing Ability. E.T.S. Research Monograph 6. Princeton: College Entrance Examination Board.
- GOODMAN, K.S. (1967). 'Reading: A Psycholinguistic Guessing Game.' Journal of the Reading Specialist 6.
- GORMAN, T.P., WHITE, J., ORCHARD, L. and TATE, A. (eds.) (1981). Language Performance in Schools Primary Survey Report No. 1. D.E.S. H.M.S.O.
- GOROSCH, M. (1973). 'On Defining Linguistic Objectives for the Teaching/Learning of Modern Languages for Adults.' in de GREVE et al. (eds.) 1973.
- GOROSCH, M. (1976). 'Modern Language Teaching to Adults for Professional Use.' A.L.S.E.D.-L.S.P. Newsletter 5.
- GOVE, P.S. (1981). 'Aspects of Communicative Testing with Special Reference to English for Science and Mathematics 'A' Level.' University of Exeter M.A. Thesis.
- GRAY, W.S. (1960). 'The Major Aspects of Reading.' in ROBINSON (ed.) 1960.
- GREENALL, G.M. (1980). 'Some Code Features of Examination Scripts: A Comparison between Native Speaker and Non-Native Speaker Answers.' in GREENALL and PRICE (eds.) 1980.
- GREENALL, G.M. and PRICE, J.E. (eds.) (1980). 'Study Modes and Academic Development of Overseas Students.' E.L.T. Documents 109. British Council.
- GRIMES, J.E. (1975). The Thread of Discourse. Mouton.
- GUILFORD, J.P. (1965). Fundamental Statistics in Psychology and Education. 4th Edition. McGraw Hill.
- GURNEY, R. (1973). Language, Brain and Interactive Processes. Arnold.
- HALE REPORT (THE) (1964). Report of the Committee on University Teaching Methods. University Grants Committee.
- HALL, M. (ed.) (1980). Strangers in Birmingham. A Survey of Overseas Students at Birmingham University and Polytechnic. Birmingham International Council for Overseas Student Affairs.

- HALLIDAY, M.A.K. and HASAN, R. (1976). Cohesion in English. Longman.
- HALLIDAY, M.A.K., McINTOSH, A. and STREVENS, P. (eds.) (1964). The Linguistic Sciences and Language Teaching. Longman.
- HARRIS, D.P. (1969). Testing English as a Second Language. McGraw Hill.
- HARRISON, A. (1982). 'Review of Graded Tests.' Schools Council Examinations Bulletin 41.
- HARTLEY, J. (1976). 'Lecture Handouts and Student Note-Taking.' Programmed Learning and Educational Technology 13/2.
- HARTLEY, J. and DAVIES, I.K. (1978). 'Note-Taking: A Critical Review.' Programmed Learning and Educational Technology 15/3.
- HARTLEY, J. and MARSHALL, S. (1974). 'On Notes and Note-Taking.' Universities Quarterly 28.
- HARTLEY, J. and TRUEMAN, M. (1978). 'Note-Taking in Lectures: A Longitudinal Study.' Bulletin of the British Psychological Society 31.
- HARTOG, P.J., RHODES, E.C. and BURT, C. (1936). The Marks of Examiners. Macmillan.
- HARTOG, P.J. et al. (1941). The Marking of English Essays. Macmillan.
- HAUG, U.A. (1981). 'Student Selection with Particular Reference to Applicants Whose Native Language is not English.' Polytechnic of Central London in Collaboration with London University Ph.D. Thesis.
- HAWKEY, M. (1979). 'E.S.P. and the Case for Tailor-Made Courses: An Investigation and a Possible Framework for Assessing Purpose-Specific Language Teaching Materials.' University of London Institute of Education M.A. Thesis.
- HAWKEY, R. (1982). 'An Investigation of Inter-Relationships between Cognitive/Affective and Social Factors and Language Learning. A Longitudinal Study of 27 Overseas Students Using English in Connection with Their Training in the United Kingdom.' University of London Ph.D. Thesis.
- HEAD, J.J. (1966). 'Multiple Marking of an Essay Item in Experimental 'O' Level Nuffield Biology Examinations.' Educational Review 19/1.
- HEATON, J.B. (1970). 'Tests and Courses in English Language: The Leeds Approach.' U.K.C.O.S.A. News.

- HEATON, J.B. (1975a). Writing English Language Tests. Longman.
- HEATON, J.B. (1975b). Studying in English. Longman.
- HEATON, J.B. (1980). 'English Language Test for Overseas Students.' Mimeo. University of Leeds.
- HEATON, J.B. (ed.) (1982a). Language Testing. Modern English Publications Ltd.
- HEATON, J.B. (1982b). 'Writing in Perspective: Some Comments on the Testing and Marking of Written Communication.' in HEATON (ed.) 1982a.
- HEATON, J.B. and PUGH, A.K. (1974). 'A Study of the Relationship between Scores Obtained by Overseas Students on a Test of English Proficiency and the Examination Results in Their University Courses.' Mimeo. University of Leeds.
- HEATON, J.B. and PUGH, A.K. (1975). 'Real Test of Success is How Well They Read English.' T.H.E.S. 178.
- HEBRON, C.C. (1967). 'The Performance of Overseas Students in Rutherford College.' Mimeo. University of Newcastle-upon-Tyne.
- HENDERSON, W. (1979). 'Teaching Economics through the Medium of a New Second Language.' Economics Autumn 1979.
- HENNING, G. (1982). 'Growth Referenced Evaluation and the Critical Intervention Index.' Paper presented at the 1982 T.E.S.O.L. Convention, Hawaii.
- HEWITT, E.A. (1967). 'The Reliability of G.C.E. 'O' Level Examinations in English Language.' Mimeo. Joint Matriculation Board.
- HOLDEN, S. (ed.) (1977). English for Specific Purposes. Modern English Publications Ltd.
- HOLDS, C.D. (1972). 'An Investigation into Some Aspects of the English Language Problems of Two Groups of Overseas Post-Graduate Students at Birmingham University.' University of Birmingham M.A. Thesis.
- HOLLAND, P.W. and RUBIN, D.B. (1982). Test Equating. Academic Press.
- HOME OFFICE (Cmd. 4298) (1970). Instructions to Immigration Officers. Commonwealth Immigrants Acts 1962 and 1968. H.M.S.O.
- HONEYWELL (1981). 64 DPS STATPAC/64 User Manual.

- HOUGHTON, D. (1980). 'Writing Problems of Iranian Students.' in GREENALL and PRICE (eds.) 1980.
- HOUSTON, J.G. (1974). 'The Effects of Different Verbal Teaching Styles on Attainment in Physics of Educational Objectives in the Cognitive and Affective Domains.' University of Edinburgh Ph.D. Thesis.
- HOUSTON, J.G. (1981). 'National Criteria.' Mimeo. Associated Examining Board.
- HOUSTON, J.G. (1982). 'Sampling.' Mimeo. Associated Examining Board.
- HOUSTON, J.G. (1983). 'Norm and Criterion Referencing of Performance Levels in Tests of Educational Attainment.' Mimeo. Associated Examining Board Research Advisory Committee Paper 241.
- HOWATT, A.P.R. and DAVIES, A. (1979). E.L.B.A. Testing 1973-1977 Interim Report on Analysis of Results. Institute for Applied Language Studies, University of Edinburgh.
- HOWE, M.J.A. (1970). 'Note-Taking Strategy, Review and Long-Term Retention of Verbal Information.' Journal of Educational Research 63/285.
- HOWE, P. (1980). English Language Entrance Examination. Mimeo. University College at Buckingham.
- HUDDLESTON, R.D., HUDSON, R.A., WINTER, E.O. and HENRICI, A. (eds.) (1968). Sentence and Clause in Scientific English. Report of the O.S.T.I. Research Project. Communication Research Centre, Department of General Linguistics, University College London.
- HUGHES, A. (1981a). 'Reaction to the Palmer and Bachman and the Vollmer Papers (1).' in ALDERSON and HUGHES (eds.) 1981.
- HUGHES, A. (1981b). 'Conversational Cloze as a Measure of Oral Ability.' E.L.T. Journal XXXV/2.
- HUGHES, G. and KNIGHT, M. (1977). 'Student Specific English - One-to-One Courses.' in HOLDEN (ed.) 1977.
- HUGHES, A. and PORTER, D. (eds.) (1983). Current Developments in Language Testing. Academic Press.
- HUTCHINSON, T. and WATERS, A. (1979). 'Performance and Competence in English for Specific Purposes. The Communicative Needs of the Technical Student.' Mimeo. University of Lancaster.
- HUTCHINSON, T. and WATERS, A. (1980). 'Communication in the Technical Classroom - "You Just Shove This Little Chappie in Here Like That".' in BRITISH COUNCIL 1980b.

- HUTCHINSON, T. and WATERS, A. (1981). 'Performance and Competence in English for Specific Purposes.' Applied Linguistics II/I.
- HYMES, D.H. (1972). 'On Communicative Competence.' in PRIDE and HOLMES (eds.) 1972.
- IMHOOF, M. and HUDSON, H. (1975). From Paragraph to Essay. Longman.
- INGRAM, D.E. (1981). 'The Australian Second Language Proficiency Ratings.' in READ (ed.) 1981a.
- INGRAM, E. (1964). English Language Battery (E.L.B.A.). University of Edinburgh, Department of Linguistics.
- INGRAM, E. (1968). 'Attainment and Diagnostic Testing.' in DAVIES (ed.) 1968.
- INGRAM, E. (1970). 'Report on the Marking of English Compositions.' Mimeo. University of Edinburgh, Department of Linguistics.
- INGRAM, E. (1973). 'English Standards for Foreign Students.' University of Edinburgh Bulletin 9 9/12.
- INGRAM, E. (1977). 'Basic Concepts in Testing.' in ALLEN and DAVIES (eds.) 1977.
- IRVINE, P., ATAI, P. and OLLER, J.W. (1974). 'Cloze, Dictation and the Test of English as a Foreign Language.' Language Learning 24/2.
- JACOBS, H.L., ZINKGRAF, S.A., WORMUTH, D.R., FAYE HARTFIEL, V. and HUGHEY, J.B. (1981). English Composition Program. Testing E.S.L. Composition: A Practical Approach. Newbury House.
- JAKOBOVITS, L.A. (1970). Foreign Language Learning: A Psycholinguistic Analysis of the Issues. Newbury House.
- JAMES, G.C.R. (forthcoming). 'The Testing of Spoken English.' University of Exeter Ph.D. Thesis (in preparation).
- JAMES, A.J., ROBINSON, J.L. and SCARRETT, H. (1977). 'What Queen's English do Universities Accept?' T.H.E.S. 289.
- JAMES, K. (1972). 'Overseas Students' Listening Problems: An Enquiry.' Mimeo. University of Manchester.
- JAMES, K. (1977). 'Note-Taking in Lectures: Problems and Strategies.' in COWIE and HEATON (eds.) 1977a.
- JAMES, K. (1980a). 'Seminar Overview.' in GREENALL and PRICE (eds.) 1980.

- JAMES, K. (1980b). 'Survey of University of Manchester Overseas Post-Graduate Students: Initial Level of Competence in English and their Subsequent Academic Performance: Calendar Year 1977.' in GREENALL and PRICE (eds.) 1980.
- JAMES, K., JORDAN, R.R. and MATTHEWS, A.J. (1979). Listening Comprehension and Note-Taking Course. Collins.
- JOHNS, A.M. (1981). 'Necessary English: A Faculty Survey.' T.E.S.O.L. Quarterly 15/1.
- JOHNS, T.F. (1978). 'The Text and its Message: An Approach to the Teaching of Reading Strategies for Students of Development Administration.' Mimeo. University of Birmingham.
- JOHNS, T.F. (1979). University of Birmingham Assessment and Diagnostic Test Question Booklet. English for Overseas Students Unit, University of Birmingham.
- JOHNS, T.F. (1980). 'English for Overseas Students: A Revision of the Birmingham Assessment and Diagnostic Test.' University of Birmingham Teaching News 10
- JOHNS, T.F. (1981). 'Testing a Test: A Reply to Allanson, J.' University of Birmingham Teaching News 12.
- JOHNS, C.M. and T.F. (1977a). 'Seminar Discussion Strategies.' in COWIE and HEATON (eds.) 1977a.
- JOHNS, T.F. and C.M. (1977b). 'The Current Programme of Materials Development in English for Academic Purposes at the Universities of Birmingham and Aston.' in COWIE and HEATON (eds.) 1977a.
- JOHNSON, K. (1976). 'Communicative Syllabus Design for Written English.' Paper presented at the B.A.A.L. Annual General Meeting. University of Reading.
- JOHNSON, K. (1977a). 'Why Are Foreign Students Incoherent?' in HOLDEN (ed.) 1977.
- JOHNSON, K. (1977b). 'Teaching Virtuous Writing.' in HOLDEN (ed.) 1977.
- JOHNSON, K. (1979). 'Communicative Writing Practice and Aristotelian Rhetoric.' Paper given at the Conference on the Teaching of Writing, held at University of Carleton, U.S.A.
- JOHNSON, K. (1981). Communicate in Writing. Teachers Book. Longman.
- JOHNSON, K. and MORROW, K.E. (1976). 'Meeting Some Social Language Needs of Overseas Students.' E.T.I.C., E.L.T. Documents 76/2.

- JOHNSON, K. and MORROW, K.E. (eds.) (1978). Functional Materials and the Classroom Teacher. University of Reading, Centre for Applied Language Studies.
- JOHNSON, K. and MORROW, K.E. (1980). Communicative Language Teaching: Issues and Applications. Longman.
- JOHNSON, K. and PORTER, D. (1983). Perspectives in Communicative Language Teaching. Academic Press.
- JONES, K. (1974). 'The Role of Discourse Analysis in Devising Undergraduate Reading Programmes.' Mimeo. E.T.I.C., British Council.
- JONES, K. and ROE, P. (1975). 'Designing English for Science and Technology (E.S.T.) Programmes.' in BRITISH COUNCIL 1975.
- JONES, K. and ROE, P. (1976). 'Problems in Designing Programmes in English for Science and Technology.' in RICHARDS (ed.) 1976.
- JONES, R.L. and SPOLSKY, B. (eds.) (1975). 'Testing Language Proficiency.' Center for Applied Linguistics: Arlington.
- JOOS, M. (1967). The Five Clocks. New York: Harcourt, Brace and World Inc.
- JORDAN, R.R. (1977a). 'Identification of Problems and Needs: A Student Profile.' in COWIE and HEATON (eds.) 1977a.
- JORDAN, R.R. (1977b). 'Study Skills and Pre-sessional Courses.' in HOLDEN (ed.) 1977.
- JORDAN, R.R. (1980). Academic Writing Course. Collins.
- JORDAN, R.R. and MACKAY, R. (1973). 'A Survey of the Spoken English Problems of Overseas Post-Graduates at the Universities of Manchester and Newcastle-upon-Tyne.' Journal (Durham and Newcastle Institutes of Education) 25/125.
- JUPP, T.C. and HODLIN, S. (1975). Industrial English: an Example of Theory and Practice in Functional Language Teaching. Heinemann Educational.
- JUPP, T.C. and MILNE, J. (1972). Guided Paragraph Writing. Heinemann.
- KAHN, R.L. and CANNELL, C.F. (1957). The Dynamics of Interviewing. New York: Wiley.
- KAPLAN, R.B. (1966). 'Cultural Thought Patterns in Intercultural Education.' Language Learning XVI/1 and 2.
- KELLY, R. (1978). 'On the Construct Validation of Comprehension Tests: An Exercise in Applied Linguistics.' University of Queensland Ph.D. Thesis.

- KELLY, R. (1981). 'Aspects of Communicative Performance.' Applied Linguistics 11/2.
- KENNEDY, G.D. (1978). The Testing of Listening Comprehension. S.E.A.M.E.O. Regional Language Centre, Singapore University Press.
- KENNEDY, G.D. (1981). 'Performance Variability and Listening Proficiency.' in READ (ed.) 1981b.
- KING, R. (1980). 'Real Groups and False Aggregates in Educational Research.' Educational Studies 6/3.
- KINSELLA, V. (ed.) (1978). Language Teaching and Linguistics: Surveys. Cambridge University Press.
- KLEIN-BRALEY, C. (1981). 'Empirical Investigations of Cloze Tests.' Universität Duisburg Ph.D. Thesis.
- KLEIN-BRALEY, C. and STEVENSON, D.K. (eds.) (1979). Practice and Problems in Language Testing I. Verlag Peter D. Lang.
- LACKSTROM, J., SELINKER, L. and TRIMBLE, L. (1973). 'Technical Rhetorical Principles and Grammatical Choice.' T.E.S.O.L. Quarterly 7/2.
- LADO, R. (1961). Language Testing: The Construction and Use of Foreign Language Tests. Longman.
- LAING, A. (1971). 'Common Study Difficulties of Overseas Students.' in FURTHER EDUCATION STAFF COLLEGE 1971.
- LARTER, J.K. (1962). 'English Language Problems of Overseas Students in Technical Colleges.' The Vocational Aspect of Further Education 14.
- LAWRENCE, M.S. (1972). Writing as a Thinking Process. Ann Arbor, Michigan: University of Michigan Press.
- LEE, K.C. (1976). 'Trends in the Study of E.S.T.' in RICHARDS (ed.) 1976.
- LEE, M.Y., ABD-ELLA, M. and BURKS, L.A. (1981). Needs of Foreign Students from Developing Nations at U.S. Colleges and Universities. N.A.F.S.A., Washington.
- LEE, Y.P. (1981a). 'Evaluation and Measurement of Communicative Competence: the Case for Direct Language Tests.' in READ (ed.) 1981b.
- LEE, Y.P. (1981b). 'Some Notes on Internal Consistency Reliability Estimation for Tests of Language Use.' in FOK, LORD, LOW and T'SOU (eds.) 1981.

- LEE, Y.P. and LOW, G. (1981). 'Towards a Classification of Tests of Language Use: The Dimension of Test Directness.' Paper delivered at 6th A.I.L.A. World Congress.
- LEECH, G.N. and SVARTVIK, J. (1975). A Communicative Grammar of English. Longman.
- LEEMING, J. and LOW, G.D. (1977). 'Outline Proposals for a Test in English for Overseas Science and Technology Students in Higher Education (Mode 2).' Mimeo. North Trafford College of Further Education.
- LEWIS, C.J. (1981). 'Testing Communicative Competence.' M.A.L.S. Journal 6.
- LEWIS, D.G. (1965). 'Objectives in the Teaching of Science.' Educational Research 7/3.
- LIEBERMAN, P. (1965). 'On the Acoustic Basis of the Perception of Intonation by Linguists.' Word XXI.
- LILLEY, A.D. (1976). 'Towards the Analysis of a Communicative Situation: the Science Lecture.' Mimeo. British Council, E.T.I.C. Archives.
- LOW, G.D. (1982). 'The Direct Testing of Academic Writing in a Second Language.' System 10/3.
- LOW, G.D. and LEE, Y.P. (1982). 'How Shall a Test be Referenced?' Paper given at Language Testing Symposium, December 1982. University of Hong Kong.
- LUCAS, A.M. (1971). 'Multiple Marking of a Matriculation Biology Essay Question.' British Journal of Educational Psychology 41/1.
- LUNZER, E. and GARDNER, K. (eds.) (1979). The Effective Use of Reading. Heinemann Educational Books Ltd. for Schools Council.
- LYONS, J. (ed.) (1970). New Horizons in Linguistics. Penguin.
- LYONS, J. (1977). Semantics. Cambridge University Press.
- MCDONOUGH, J. (1978). Listening to Lectures - Sociology. Oxford University Press.
- MCDONOUGH, S. (1981). Psychology in Foreign Language Teaching. Allen and Unwin.
- MELDOWNEY, P.L. (1974). 'Notes for the Guidance of Teachers Offering Candidates for the University Entrance Test in English (Overseas).' Joint Matriculation Board, Manchester.
- MELDOWNEY, P.L. (1976). 'Test in English (Overseas). The Position after Ten Years.' Occasional Paper 36. Joint Matriculation Board, Manchester.

- McELDOWNEY, P.L. (1982). 'A Place for Visuals in Language Testing.' in HEATON (ed.) 1982.
- McLEISH, J. (1968). 'The Lecture Method.' Cambridge Monographs on Teaching Methods No. 1. Cambridge Institute of Education.
- MACFARLANE-SMITH, I., HOWES, R. and SHEPHERD, K. (1976). 'A Study of the Abilities and Interests of Overseas Students.' The Vocational Aspect of Education XXVIII/70.
- MACKAY, R. (1974). 'Teaching the Information Gathering Skills.' R.E.L.C. Journal 5/2.
- MACKAY, R., KLASSEN, B. and GARST, T. (1975). 'Practical Steps Towards the Classification of Reading Comprehension Exercise Types.' Eduotec: Languages for Special Purposes 6.
- MACKAY, R. and MOUNTFORD, A. (1973). 'A Programme in English for Post-Graduate Students in the Faculties of Science, Applied Sciences and Agriculture in the University of Newcastle-upon-Tyne.' in de GREVE et al. (eds.) 1973.
- MACKAY, R. and MOUNTFORD, A. (1974). 'Recent Developments and Research in Reading Relevant to the Teaching and Learning of English in South-east Asia.' S.E.A.M.E.O. Regional English Language Centre, Ninth Regional Seminar.
- MACKAY, R. and MOUNTFORD, A. (eds.) (1978). English for Specific Purposes. Longman.
- MACKENZIE, J.G. (1977). 'Some English Language Problems of Latin-American Post-Graduate Students.' in COWIE and HEATON (eds.) 1977a.
- MANOOK, S. and KING, A.R. (1981). 'Note-Taking in Lectures.' Mimeo. University of Lancaster Fifth International Conference on Higher Education.
- MARTIN, A.V. (1976). 'Teaching Academic Vocabulary to Foreign Graduate Students.' T.E.S.O.L. Quarterly 10/1.
- MASON, C. (1971). 'The Relevance of Intensive Training in English as a Foreign Language for University Students.' Language Learning 21/2 Michigan.
- MASSEY, A.J. (1979). 'Comparing Standards in English Language: A Report of the Cross-Moderation Study Based on the 1978 Ordinary Level Examinations of the Nine G.C.E. Boards.' Southern Universities' Joint Board, Test Development and Research Unit.
- MATTHEWS, A. (1978). 'Lecture Comprehension and Note-Taking.' in BRITISH COUNCIL 1978c.

- MEAD, R. (1980). 'An Analysis of Small Group Discourse and its Application.' University of Birmingham Ph.D. Thesis.
- MEAD, R. (1981). 'Review of J. Munby "Communicative Syllabus Design".' Applied Linguistics III/1.
- MELNIK, A. and MERRITT, J. (eds.) (1972). Reading: Today and Tomorrow. University of London Press.
- MILLER, G.A. (1956). 'The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information.' The Psychological Review 63/2.
- MOLLER, A.D. (1977). 'A Case for a Crude Test Overseas.' in COWIE and HEATON (eds.) 1977a.
- MOLLER, A.D. (1981a). 'Assessing Proficiency in English for Use in Further Study.' in READ (ed.) 1981a.
- MOLLER, A.D. (1981b). 'Reaction to the Morrow paper (2).' in ALDERSON and HUGHES (eds.) 1981.
- MOLLER, A.D. (1982). 'A Study in the Validation of Proficiency Tests of English as a Foreign Language.' University of Edinburgh Ph.D. Thesis.
- MONTGOMERY, M. (1977). 'Some Aspects of Discourse Structure and Cohesion in Selected Science Lectures.' University of Birmingham M.A. Thesis.
- MOODY, K.W. (1976). Written English under Control. Ibadan: Oxford University Press.
- MORRIS, B.S. (1967). International Community? National Union of Students and the Scottish Union of Students.
- MORRISON, J.W. (1974). 'An Investigation of Problems in Listening Comprehension Encountered by Overseas Students in the First Year of Post-Graduate Studies in Sciences in the University of Newcastle-upon-Tyne and the Implications for Teaching.' University of Newcastle-upon-Tyne M.Ed. Thesis.
- MORRISON, R.B. (1968). 'English Language 'O' Level - June 1967 Marking Experiment.' Mimeo. Associated Examining Board.
- MORRISON, R.B. (1969). 'English Language 'O' Level - June 1968 Marking Experiment.' Mimeo. Associated Examining Board.
- MORRISON, R.B. (1970). 'English Language 'O' Level Marking Experiment 1968-1969.' Mimeo. Associated Examining Board.
- MORRISON, R.L. and VERNON, P.E. (1941). 'A New Method of Marking English Compositions.' British Journal of Educational Psychology 11/2.

- MORROW, K.E. (1977). Techniques of Evaluation for a Notional Syllabus. The Royal Society of Arts.
- MORROW, K.E. (1979). 'Communicative Language Testing: Revolution or Evolution.' in BRUMFIT and JOHNSON (eds.) 1979.
- MORROW, K.E. (1980a). Skills for Reading. Oxford University Press.
- MORROW, K.E. (1980b). 'Measuring Communicative Ability.' Paper presented at the English Speaking Union Conference on English for International Communication, London.
- MORROW, K.E. and JOHNSON, K. (1977). 'Meeting some Social Language Needs of Overseas Students.' in COWIE and HEATON (eds.) 1977a.
- MORSE, H.T. and McCUNE, G.H. (1971). 'Selected Items for the Testing of Study Skills and Critical Thinking.' National Council for the Social Studies Bulletin 15, 5th Edition.
- MOST, R.L. (1972). 'Testing Aural Comprehension.' Rapport D'Activités de L'Institut Phonétique (Brussels) 6.
- MULLEN, K.A. (1980). 'Evaluating Writing Proficiency in E.S.L.' in OLLER and PERKINS (eds.) 1980.
- MUNBY, J.L. (1977). 'Designing a Processing Model for Specifying Communicative Competence in a Foreign Language: A Study of the Relationship between Communication Needs and the English Required for Specific Purposes.' University of Essex Ph.D. Thesis.
- MUNBY, J.L. (1978). Communicative Syllabus Design. Cambridge University Press.
- MURPHY, R.J.L. (1978a). 'Sex Differences in Objective Test Performance.' Mimeo. Associated Examining Board Research Report RAC/56.
- MURPHY, R.J.L. (1978b). 'Reliability of Marking in Eight G.C.E. Examinations.' British Journal of Educational Psychology 48/2.
- MURPHY, R.J.L. (1979). 'Mode 1 Examining for the General Certificate of Education: A General Guide to some Principles and Practices.' Mimeo. Associated Examining Board.
- MURPHY, R.J.L. (1980). 'Sex Differences in G.C.E. Examination Entry Statistics and Success Rates.' Educational Studies 6/2.
- MURPHY, R.J.L. (1982). 'A Further Report of Investigations into the Reliability of Marking of G.C.E. Examinations.' British Journal of Educational Psychology 52/1.
- NACKE, P.L. (ed.) (1974). Interaction: Research and Practice for College-Adult Reading. 23rd Yearbook of the National Reading Conference.

- NATIONAL ASSOCIATION FOR FOREIGN STUDENTS' ADVISERS (N.A.F.S.A.), (1961). 'Research in Programs for Foreign Students.' N.A.F.S.A. Studies and Papers: Research Series 2.
- NG'OMBE, J.L. (1981). 'An Investigation into the Communication Syllabus for Second Language (Tertiary) Students.' University of London Ph.D. Thesis.
- NYE, P. (1978). 'Student Variables in Relation to Note-Taking During a Lecture.' Programmed Learning and Educational Technology 15/3.
- OLLER, J.W. (1971). 'Dictation as a Device for Testing Foreign Language Proficiency.' English Language Teaching 25/3.
- OLLER, J.W. (1972a). 'Assessing Competence in E.S.L. Reading.' T.E.S.O.L. Quarterly 6.
- OLLER, J.W. (1972b). 'Scoring Methods and Difficulty Levels for Cloze Tests of E.S.L. Proficiency.' Modern Language Journal 56.
- OLLER, J.W. (1976). 'Evidence for a General Language Proficiency Factor: An Expectancy Grammar.' Die Neuren Sprachen 2.
- OLLER, J.W. (1979). Language Tests at School. Longman.
- OLLER, J.W. (1980). 'A Comment on Specific Variance Versus Global Variance in Certain E.F.L. Tests.' T.E.S.O.L. Quarterly 14/4.
- OLLER, J.W. and CONRAD, C. (1971). 'The Cloze Technique and E.S.L. Proficiency.' Language Learning 21/2.
- OLLER, J.W. and HINOFOTIS, F.B. (1980). 'Two Mutually Exclusive Hypotheses about Second Language Ability: Indivisible or Partially Divisible Competence.' in OLLER and PERKINS (eds.) 1980.
- OLLER, J.W. and PERKINS, K. (eds.) (1978). Language in Education: Testing the Tests. Newbury House.
- OLLER, J.W. and PERKINS, K. (eds.) (1980). Research in Language Testing. Newbury House.
- OLLER, J.W. and STREIFF, V. (1975). 'Dictation: A Test of Grammar-Based Expectancies.' English Language Teaching Journal 30/1.
- OSTLER, S.E. (1980). 'A Survey of Academic Needs for Advanced E.S.L.' T.E.S.O.L. Quarterly 14/4.
- PALMER, A.S. (1972). 'Testing Communication.' I.R.A.L. 10/1.
- PALMER, A.S. (1981). 'Measurements of Reliability and Validity of Two Picture-Description Tests of Oral Communication.' in PALMER, GROOT and TROSPER (eds.) 1981.

- PALMER, A.S. and BACHMAN, L.F. (1981a). 'Basic Concerns in Test Validation.' in ALDERSON and HUGHES (eds.) 1981.
- PALMER, A.S., GROOT, P.J.M. and TROSPER, G.A. (eds.) (1981b). The Construct Validation of Tests of Communicative Competence. T.E.S.O.L.: Washington D.C., U.S.A.
- PARISH, C.M. (1977). Teaching and Learning Methods and Materials in Selected University Courses for Overseas Students. University of Reading, Agricultural Extension and Rural Development Centre.
- PARKINSON, J.A. (1971). 'Overseas Students' Language Needs.' in FURTHER EDUCATION STAFF COLLEGE 1971.
- PAULSTON, C.B. (1974). 'Linguistic and Communicative Competence.' T.E.S.O.L. Quarterly 8/4.
- PENFOLD, E.D.M., (1956). 'Essay Marking Experiments: Shorter and Longer Essays.' British Journal of Educational Psychology 26/2.
- PERERA, K. (1980). 'The Assessment of Linguistic Difficulty in Reading Material.' Educational Review 32/2.
- PERKINS, K. (1980). 'Using Objective Methods of Attained Writing Proficiency to Discriminate among Holistic Evaluations.' T.E.S.O.L. Quarterly 14/1.
- PERKINS, K. (1982). 'The Test of Ability to Subordinate: Predictive and Concurrent Validity for Attained E.S.L. Composition.' in CULHANE, KLEIN-BRALEY and STEVENSON (eds.) 1982.
- PERREN, G.E. (1963). 'Linguistic Problems of Overseas Students in Britain.' E.T.I.C. Occasional Paper No. 3. British Council.
- PERREN, G.E. (1967a). 'Testing Ability in English as a Second Language: 1. Problems.' English Language Teaching 21/2.
- PERREN, G.E. (1967b). 'Testing Ability in English as a Second Language: 2. Techniques.' English Language Teaching 21/3.
- PERREN, G.E. (1967c). 'Testing Ability in English as a Second Language: 3. Spoken Language.' English Language Teaching 22/1.
- PERREN, G.E. (1968). 'Testing Spoken Language: Some Unsolved Problems.' in DAVIES (ed.) 1968.
- PERREN, G.E. (ed.) (1971). Science and Technology in a Second Language. C.I.L.T. Reports and Papers No.7.
- PERREN, G.E. and TRIM, J.L.M. (eds.) (1971). Applications of Linguistics. Selected papers of the Second International Congress of Applied Linguistics, Cambridge 1969.

- PHAM VAN DANG, P. (1977). 'E.S.P. Course Design with Special Reference to E.S.T.' University of London M.A. Thesis.
- PICKETT, D. (1980). 'Aspects of the Cambridge First Certificate in English Examination, 1975-1980.' University of London M.A. Thesis.
- PILLINER, A.E.G. (1968). 'Subjective and Objective Testing.' in DAVIES (ed.) 1968.
- PILLINER, A.E.G. (1969). 'Multiple Marking: Wiseman or Cox.' British Journal of Educational Psychology 39.
- PLAISTER, T.H. (1968). 'Testing Aural Comprehension.' in UPSHUR and FATA (eds.) 1968.
- PLAISTER, T.H. (1976). Developing Listening Comprehension for E.S.L. Students: the Kingdom of Kochen. Prentice-Hall.
- POLITICAL AND ECONOMIC PLANNING (P.E.P.) (1965). New Commonwealth Students in Britain: with Special Reference to Students from East Africa. Allen and Unwin.
- POPHAM, W.J. (1978). Criterion Referenced Measurement. Prentice-Hall.
- POPHAM, W.J. and HUSEK, T.R. (1969). 'Implications of Criterion-Referenced Measurement.' Journal of Educational Measurement 61.
- PORTER, D. (1983). 'Assessing Communicative Proficiency: The Search for Validity.' in JOHNSON and PORTER (eds.) 1983.
- PRICE, J.E. (1977a). 'Information on Pre-sessional English Language Courses in Britain Today.' Paper prepared for a 1977 S.E.L.M.O.U.S. Seminar.
- PRICE, J.E. (1977b). 'Study Skills - with Special Reference to Seminar Strategies and One Aspect of Academic Writing.' in HOLDEN (ed.) 1977.
- PRICE, J.E. (1980). 'The Structure of Post-Graduate Taught Courses in Engineering: Some Implications for Language Course Design.' in GREENALL and PRICE (eds.) 1980.
- PRIDE, J.B. and HOLMES, J. (eds.) (1972). Sociolinguistics. Penguin.
- QUIRK, R., GREENBAUM, J., LEECH, G. and SVARTVIK, J. (1972). A Grammar of Contemporary English. Longman.
- RAATZ, U. (1981). 'The C-Test. A Modification of the Cloze Procedure.' Paper delivered at the Fourth International Language Testing Symposium of the I.U.S., Colchester, England.

- RACE, P. (1981). 'An Objective-Based Introduction to Report Writing.' Mimeo. Polytechnic of Wales.
- RANKIN, E. (1974). 'The Cloze Procedure Revisited.' in NAKLE (ed.) 1974.
- REA, P.M. (1978). 'Assessing Language as Communication.' M.A.L.S. Journal New Series No. 3. University of Birmingham.
- READ, J.A.S. (ed.) (1981a). 'Directions in Language Testing.' S.E.A.M.E.O. Regional Language Centre. Anthology Series 9.
- READ, J.A.S. (ed.) (1981b). 'Papers on Language Testing.' S.E.A.M.E.O. Regional Language Centre. Occasional Papers No.18.
- REED, B., HUTTON, J. and BAZALGETTE, J. (1978). Freedom to Study. Overseas Student Trust, London.
- REID, S. (1976). 'Testing the Limits of Their English Competence: A Diagnostic Programme Developed at Middlesex Polytechnic.' T.H.E.S. 246.
- RICHARDS, J. (eds.) (1976). Teaching English for Science and Technology. Singapore University Press.
- RICHTERICH, R. (1973). 'Modèle pour la Definition des Besoins Langagiers des Adultes Apprenant Une Langue Vivante.' Modern Language Teaching to Adults: Language for Special Purposes. A.I.M.A.V./DIDIÉR.
- RICHTERICH, R. (1974). 'Language Needs and Types of Adults.' E.L.T. Documents 74/3. E.T.I.C., British Council.
- RIDDLE, M.F. (1978). Test in English Language Performance. Mimeo. Middlesex Polytechnic.
- RIVERS, W.M. (1966). 'Listening Comprehension.' Modern Language Journal (April).
- RIVERS, W.M. (1968). Teaching Foreign Language Skills. University of Chicago Press.
- ROBINSON, H.A. (ed.) (1966). Reading: Seventy Five Years of Progress. Supplementary Education Monograph No.96. University of Chicago Press.
- ROBINSON, H.M. (ed.) (1960). Sequential Development of Reading Abilities. Supplementary Educational Monograph No.90. University of Chicago Press.
- ROBINSON, H.M. (1966). 'The Major Aspects of Reading.' in ROBINSON (ed.) 1966.

- ROGERS, S. (1977). 'The Communicative Needs of some Overseas Post-Graduate Students.' in COWIE and HEATON (eds.) 1977.
- ROYAL SOCIETY OF ARTS (R.S.A.) (1980). Examinations in the Communicative Use of English as a Foreign Language. Specifications and Specimen Papers. Royal Society of Arts.
- RYAN, M. (1979). 'English for Academic Purposes.' University of Wales Institute of Science and Technology (U.W.I.S.T.) M.A. Thesis.
- ST. GEORGE, R. (1981). 'The Language Achievement Tests for Overseas Students - L.A.T.O.S.: A Kiwi Controversy.' New Zealand Journal of Educational Studies 16/2.
- SALIMBENE, S. (1982). Strengthening Your Study Skills: A Guide for Overseas Students. University of London Institute of Education.
- SAVIGNON, S.J. (1972). 'Teaching for Communicative Competence: A Research Report.' Audio-Visual Language Journal 10/3.
- SAVIGNON, S.J. (1982). 'Dictation as a Measure of Communicative Competence in French as a Second Language.' Language Learning 32/1.
- SCHULZ, R.A. (1977). 'Discrete-Point Versus Simulated Communication Testing in Foreign Languages.' Modern Language Journal 61/3.
- SCOTT, M. (1981). 'Some Thoughts on Testing Reading Comprehension in English for Academic Purposes.' Projeto Nacional Ensino de Inglês Instrumental em Universidades Brasileiras. Pontifícia Universidade Católica de São Paulo Working Papers No.1.
- SEARLE, J.R. (1965). 'What is a Speech Act.' in GIGLIOLI (ed.) 1976.
- SEARLE, J.R. (1971). The Philosophy of Language. Oxford University Press.
- SEATON, I. (1981). 'Background to the Specifications for an English Language Testing Service and Subsequent Developments.' in ALDERSON and HUGHES (eds.) 1981.
- SELINKER, L., TARONE, E. and HANZELI, V. (1981). English for Academic and Technical Purposes. Studies in Honour of Louis Trimble. Newbury House.
- SELINKER, L., TODD-TRIMBLE, R.M. and TRIMBLE, L. (1976a). 'Presuppositional Rhetorical Information in E.S.T. Discourse.' T.E.S.O.L. Quarterly 10/3.
- SELINKER, L., TODD-TRIMBLE, R.M. and TRIMBLE, L. (1978). 'Rhetorical Function Shifts in E.S.T. Discourse.' T.E.S.O.L. Quarterly 12/3.
- SELINKER, L. and TRIMBLE, L. (1977). 'Formal Written Communication and E.S.L.' Indian Journal of Applied Linguistics III/2.

- SELINKER, L. and TRIMBLE, L. (1976b). 'Scientific and Technical Writing: the Choice of Tense.' English Teaching Forum 14/4.
- SELINKER, L., TRIMBLE, L. and WROMAN, R. (1974). 'Presupposition and Technical Rhetoric.' E.L.T. Journal 29/1.
- SEN, A. (1970). Problems of Overseas Students and Nurses. National Foundation for Educational Research (N.F.E.R.).
- SHANNON, C.E. and WEAVER, W. (1959). The Mathematical Theory of Communication. Urbana: University of Illinois.
- SHARWOOD-SMITH, M. (1973). 'Teaching Written English for Communication.' R.E.L.C. Journal 4/2.
- SHAW, A.M. (1976). 'Approaches to a Communicative Syllabus in Foreign Language Curriculum Development.' University of Essex Ph.D. Thesis.
- SHIPMAN, M. (1972). The Limitations of Social Research. Longman.
- SIM, D.D. (1974). 'Grammatical Cohesion in English and Advanced Reading Comprehension for Overseas Students.' University of Manchester M.Ed. Thesis.
- SINCLAIR, J. Mc.H. and COULTHARD, R.M. (1975). Towards an Analysis of Discourse. Oxford University Press.
- SINGH, A.K. (1963). Indian Students in Britain. New York. Asia Publishing House.
- SKEHAN, P. (1983). 'Issues in the Testing of English for Specific Purposes.' To appear in the M.A.L.S. Journal 1983. Midland Association for Linguistic Studies.
- SOUTHAMPTON UNIVERSITY STUDENTS' UNION (S.U.S.U.) (1979). Report of the Working Party on Foreign Students' Command of English. Education and Welfare Department, Southampton University Students' Union.
- SPEARRITT, D. (1962). 'Listening Comprehension - A Factorial Analysis.' A.C.E.R. Research Series 76.
- SPENCER, D.H. (1967). Guided Composition Exercises. Longman.
- SPOLSKY, B. (1968). 'Language Testing: The Problem of Validation.' T.E.S.O.L. Quarterly 2.
- SPOLSKY, B. (1976). 'Language Testing: Art or Science.' Paper read at the 4th International Congress of Applied Linguistics. Stuttgart: Hoch Schul Verlag, Germany.
- SPOLSKY, B. (ed.) (1978). Approaches in Language Testing. Advances in Language Testing, Series: 2. Center for Applied Linguistics: Arlington.

- SPOLSKY, B. (ed.) (1979). Some Major Tests. Advances in Language Testing, Series: 1. Center for Applied Linguistics: Arlington.
- STEEL, J.H. and TALMAN, J. (1936). The Marking of English Compositions. Nisbet.
- STEVENSON, D.K. (1974). 'A Preliminary Investigation of Construct Validity and the Test of English as a Foreign Language.' University of New Mexico, Albuquerque Ph.D. Thesis.
- STEVENSON, D.K. (1981). 'Language Testing and Academic Accountability: on Re-defining the Role of Language Testing in Language Teaching.' I.R.A.L. XIX/1.
- STEVENSON, R.W. (1974). Welfare Problems Facing Overseas Students. Edinburgh: University Students' Association.
- STRAKER-COOK, R.H. (1975). 'A Communicative Approach to the Analysis of Extended Monologue Discourse and its Relevance to the Development of Teaching Materials for English for Special Purposes.' University of Edinburgh M.Litt. Thesis.
- STRAKER-COOK, R.H. (1977). 'A Rhetorical-Communicative Approach to Syllabus Design in E.S.P.' in COWIE and HEATON (eds.) 1977a.
- STREVEVS, P. (1971a). 'Alternatives to Daffodils, or - Scientist Thou Never Wert.' in PERREN (ed.) 1971.
- STREVEVS, P. (1971b). 'The Medium of Instruction (Mother-Tongue/ Second Language) and the Formation of Scientific Concepts.' I.R.A.L. 9/3.
- STREVEVS, P. (1978). 'Special-Purpose Language Learning: A Perspective.' in KINSELLA (ed.) 1978.
- STUMP, T.A. (1978). 'Cloze and Dictation Tasks as Predictors of Intelligence and Achievement Scores.' in OLLER and PERKINS (eds.) 1978.
- SWALES, J. (1971). Writing Scientific English: A Text-Book of English as a Foreign Language for Students of Physical and Engineering Sciences. Nelson.
- SWALES, J. (1976). 'Verb Frequencies in Scientific English.' E.S.P.M.E.N.A. Bulletin (Sudan) No.4.
- SWAN, H.A. (1976). New English Proficiency Course. Hulton.
- SWINTON, S.S. and POWERS, D.E. (1980). 'Factor Analysis of the Test of English as a Foreign Language for Several Language Groups.' T.O.E.F.L. Research Reports 6.
- TAJFEL, H. and DAWSON, J. (eds.) (1965). Disappointed Guests. Oxford University Press.

- TAYLOR, C. (1979). 'An Assessment of the University of Cambridge Certificate of Proficiency English.' University of London M.A. Thesis.
- TAYLOR, W.L. (1953). 'Cloze Procedure: A New Tool for Measuring Readability.' Journalism Quarterly 30.
- TEMPLETON, H.R. (1973). 'Cloze Procedure in Aural Proficiency Tests for Foreign Students Studying in English.' University of Manchester M.Ed. Thesis.
- THORNDIKE, R.L. (ed.) (1971). Educational Measurement (2nd Edition). American Council on Education, Washington D.C.
- TINKLER, T. (1973). 'The Use of the Passive in Certain Social Science Lectures and Implications for Teaching English to Non-Native Speakers.' University of Manchester M.Ed. Thesis.
- TORBE, M., FEELEY, M. and PRITCHARD, D. (1981). 'The Uses of English.' Times Educational Supplement (T.E.S.) 3384
- TURNER, V. (1975). Cognitive Development. Methuen.
- U.K.C.O.S.A. (1974). Guidance Leaflet 14. United Kingdom Council on Overseas Student Affairs, London.
- UNDERHILL, N. (1981). 'Objectivity in Oral Testing - Can We Have It and Do We Want It Anyway?' Spoken English 14/3.
- UPSHUR, J.A. and FATA, J. (eds.) (1968). Problems in Foreign Language Testing. Language Learning Special Issue Number 3.
- VALETTE, R.M. (1967). Modern Language Testing - a Handbook. New York: Harcourt, Brace and World.
- VALETTE, R.M. (1977). Modern Language Testing (2nd Edition). New York: Harcourt, Brace, Jovanovich.
- VANCE, S. (1977). 'A Placement Test for an English for Science Program.' in HOLDEN (ed.) 1977.
- VAN PASSEL, F. (1973). 'Sense and Nonsense in Defining F.L.T. Objectives.' in de GREVE et al. (eds.) 1973.
- VERNON, P.E. and MILLICAN, G.D. (1954). 'A Further Study of the Reliability of English Essays.' British Journal of Statistical Psychology 7/2.
- VOLLMER, H.J. (1979). 'Why Are We Interested in General Language Proficiency?' in KLEIN-BRALEY and STEVENSON (eds.) 1979.
- VOLLMER, H.J. (1981a). 'Why Are We Interested in General Language Proficiency?' in ALDERSON and HUGHES (eds.) 1981.

- VOLLMER, H.J. (1981b). 'Response: Issue or Non-Issue General Language Proficiency Revisited?' in ALDERSON and HUGHES (eds.) 1981.
- WALL, D. (1982). 'A Study of the Predictive Validity of the Michigan Battery, with Special Reference to the Composition Component.' in CULHANE, KLEIN-BRALEY and STEVENSON (eds.) 1982.
- WALKER, D. (1978). 'The Integration of Overseas Students into a College of Education.' University of Sheffield M.Ed. Thesis.
- WALKER, R. and ADELMAN, C. (1975). A Guide to Classroom Observation. Methuen.
- WALLACE, M.J. (1980). Study Skills in English. Cambridge University Press.
- WALSH, V. (1982). 'Reading Scientific Texts in English.' System 10/3.
- WANNER, E. (1973). 'Do We Understand Sentences from the Outside-in or from the Inside-out?' Daedalus 102.
- WATERS, A. (1977). 'Some Characteristics of English for Science Programmes.' in HOLDEN (ed.) 1977.
- WATT, I. (1964). 'The Seminar.' University Quarterly 18.
- WATT, J.C. (1980). 'Performance of Overseas Post-Graduate Students: A Management Teacher's View.' in GREENALL and PRICE (eds.) 1980.
- WARD, J.M. (1976). Proficiency in the Use of English. Longman.
- WEBB, J. (1977). 'Reflections of Practical Experience in Designing and Mounting E.S.P. Courses at the Colchester English Study Centre.' I.U.T. Bulletin Pedagogique.
- WEIR, C.J. and ORMISTON, J. (1978). 'A Comparison of the Rankings of Students on a Cloze Test as Against a Composition Test.' University of Edinburgh M.Sc. App.Ling. Assignment.
- WIDDOWSON, H.G. (1971). 'The Teaching of Rhetoric to Students of Science and Technology.' in PERREN (ed.) 1971.
- WIDDOWSON, H.G. (1975). 'E.S.T. in Theory and Practice.' in BRITISH COUNCIL 1975.
- WIDDOWSON, H.G. (1977). 'The Description of Scientific Language.' Indian Journal of Applied Linguistics III/2.
- WIDDOWSON, H.G. (1978). Teaching Language as Communication. Oxford University Press.

- WIDDOWSON, H.G. (1979). 'Directions in the Teaching of Discourse.' in BRUMFIT and JOHNSON (eds.) 1979.
- WIDDOWSON, H.G. (ed.) (1980). Reading and Thinking in English. Discourse in Action. (Teachers' Edition). Oxford University Press.
- WIDDOWSON, H.G. (1981). 'Les Fins et Les Moyens d'un Enseignement de l'Anglais en vue d'Objectifs Specifiques.' Etudes de Linguistique Appliquee 43.
- WIJASURIYA, B.S. (1971). 'The Occurrence of Discourse Markers and Inter-Sentence Connectives in University Lectures and Their Place in the Testing and Teaching of Listening Comprehension in English as a Foreign Language.' University of Manchester M.Ed. Thesis.
- WILKINS, D.A. (1972). Linguistics in Language Teaching. Edward Arnold.
- WILKINS, D.A. (1973). 'Grammatical, Situational and Notional Syllabuses.' E.L.T. Documents 6.
- WILKINS, D.A. (1976). Notional Syllabuses. Oxford University Press.
- WILKINS, D.A. (1980). 'Review of "Communicative Syllabus Design" by J. Munby.' British Journal of Educational Studies 28.
- WILLIAMS, P. (ed.) (1981). The Overseas Student Question. Studies for a Policy. Heinemann.
- WILLMOTT, A.S. and NUTTALL, D.L. (1975). The Reliability of Examinations at 16+. Macmillan.
- WINGARD, P.G. (1971). 'English for Scientists at the University of Zambia.' in PERREN (ed.) 1971.
- WINTER, E.O. (1971). 'Connection in Science Material. A Proposition about the Semantics of Clause Relations.' in PERREN (ed.) 1971.
- WISEMAN, S. (1949). 'The Marking of English Composition in Grammar School Selection.' British Journal of Educational Psychology 19/3.
- WOOD, D.N. (1967). 'The Foreign-Language Problem Facing Scientists and Technologists in the United Kingdom - Report of a Recent Survey.' Journal of Documentation 23/2.
- WOOD, R. and QUINN, B. (1976). 'Double Impression Marking of English Language Essay and Summary Questions.' Educational Review 28/3.
- WOOD, R. and WILSON, D. (1974). 'Evidence for Differential Marking Discrimination Among Examiners of English.' The Irish Journal of Education 8/1.

YORKEY, R.C. (1976). Study Skills for Students of English as a Second Language. McGraw-Hill.

ZIEKY, M.J. and LIVINGSTON, S.A. (1977). Manual for Setting Standards on the Basic Skills Assessment Tests. Princeton, N.J.: Educational Testing Service.