An investigation of inter-relationships between personality, cognitive style and language learning strategies: with special reference to a group of adult overseas students using English in their specialist studies in the U.K.

Roger Hawkey

Volume 1 of a thesis submitted for the PhD degree of London University.

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

1982
ii.

ACKNOWLEDGMENTS

I would like to express my sincere thanks to the many friends and colleagues who have helped me with this research. To the 27 students and their tutors, who so willingly allowed me to pursue my inquiries; to the British Council's English Language and Literature Division, especially Jenny Pugsley, John McGovern and the ELTI teachers, for their really active interest and support; to Choochit and Vizes Nakornchai, for their help with aspects of the empirical study; to Cyril Weir, with whom I share a similar research focus; to Chris Brumfit, my supervisor, whose guidance and understanding have been invaluable; to my wife Mavis for her constant professional as well as personal support, and my children, Ian and Emma, for their patience, and for their contributions as (fairly) naive native speaker informants.

Finally, I should like to acknowledge the fact that this work would not have been possible without the financial support and encouragement of the British Council.
A B S T R A C T

In the first part of this inquiry, Chapters One to Four, the main thesis and three broad research hypotheses are formulated. The thesis (see Chapter One) is that detailed, multi-dimensional learner profiles can fruitfully inform the design of training programmes for overseas students in the United Kingdom. This proposition is first examined through the re-assessment of various relevant literatures, with a particular focus on evidence that it is amenable to empirical investigation. Chapter Two explores the perceived relationships of cognitive/affective and social variables with target language learning and use. Positive evidence of such relationships permits the formulation of Research Hypothesis 1. In Chapter Three findings from previous inquiries into overseas students in foreign cultures are used to formulate Research Hypothesis 2, which proposes that learner profiles drawn up early on in the period of residence overseas can predict training outcomes. Chapter Four develops a framework for the elicitation and evaluation of target language data. Research Hypothesis 3, suggesting the importance of such data in the profiling of individual learners, is formulated.

Chapters Five to Eight report on the two phases of an empirical study collecting and analysing longitudinal data on the participant group of 27 Overseas Development Administration study fellows. Phase One of the study focuses on their pre-sessional remedial English programme in the UK. Phase Two follows them through their subsequent specialist training periods at various receiving institutions. Specific hypotheses necessary to the investigation of the three broader research hypotheses are tested.
iv.

using a combination of quantitative and qualitative methods. The main thesis is re-examined in the light of evidence from the empirical study, with conclusions drawn on the future training of students similar to the participant group. Possible areas of further research are suggested.

RAH
January 1982
# CONTENTS

## VOLUME 1

### CHAPTER ONE

**THE EVOLUTION OF A RESEARCH DESIGN**

1. Introduction .................................................. 1
2. The Research Impetus: experience, interest and a 'problem' 3
3. 'Extra-Linguistic' Factors and Language Learning: some views from the applied linguistics paradigm 6
4. Methods of Inquiry: the scientific and alternative paradigms 14
5. Towards a Research Design 22

### CHAPTER TWO

**KEY LEARNER AND LEARNING VARIABLES**

1. Introduction .................................................. 27
2. Potential Variables: nature, variety and inter-relationships 29
3. Cognitive/affective and Social Factors: theoretical discussion and empirical findings 30
4. Language in Study Use 33
5. Research Hypothesis 1 Formulated 80

### CHAPTER THREE

**THE OVERSEAS STUDENT**

1. Introduction .................................................. 87
2. The Overseas Student Question: sources and foci 89
3. The Overseas Student Question: methods of inquiry 90
4. Research Hypothesis 2 formulated 107
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>CHAPTER FOUR</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TARGET LANGUAGE EVALUATION: FROM THEORY TOWARDS PRACTICE</td>
<td>119</td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>121</td>
</tr>
<tr>
<td>2.</td>
<td>Target Language Evaluation: functions and needs</td>
<td>122</td>
</tr>
<tr>
<td>3.</td>
<td>The Notion of Validity</td>
<td>123</td>
</tr>
<tr>
<td>4.</td>
<td>Communicative Competence and Performance: scope and status</td>
<td>128</td>
</tr>
<tr>
<td>5.</td>
<td>Communicative Testing? Some mainly theoretical issues</td>
<td>136</td>
</tr>
<tr>
<td>6.</td>
<td>Communicative Testing: from theory towards practice</td>
<td>143</td>
</tr>
<tr>
<td>7.</td>
<td>General Evaluational Requirements: a framework for performance test construction and assessment</td>
<td>162</td>
</tr>
<tr>
<td>8.</td>
<td>Research Hypothesis 3 formulated</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER FIVE</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>THE DESIGN FOR THE EMPIRICAL STUDY</td>
<td>189</td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>191</td>
</tr>
<tr>
<td>2.</td>
<td>The Design of the Empirical Study</td>
<td>192</td>
</tr>
<tr>
<td>3.</td>
<td>Sampling: theory and practice</td>
<td>196</td>
</tr>
<tr>
<td>4.</td>
<td>Describing the Participant Group</td>
<td>204</td>
</tr>
<tr>
<td></td>
<td><strong>CHAPTER SIX</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td>THE EMPIRICAL STUDY: PHASE ONE - TARGET LANGUAGE EVALUATION</td>
<td>223</td>
</tr>
<tr>
<td>1.</td>
<td>Introduction</td>
<td>225</td>
</tr>
<tr>
<td>2.</td>
<td>Test Construction and Construct Validity</td>
<td>226</td>
</tr>
<tr>
<td>3.</td>
<td>The Tests Described</td>
<td>234</td>
</tr>
<tr>
<td>4.</td>
<td>Statistical Validation 1: reliability</td>
<td>252</td>
</tr>
<tr>
<td>5.</td>
<td>Statistical Validation 2: validity</td>
<td>290</td>
</tr>
<tr>
<td>6.</td>
<td>Individual Profiling: some case studies</td>
<td>300</td>
</tr>
<tr>
<td>CHAPTER SEVEN</td>
<td>THE EMPIRICAL STUDY: PHASE ONE — THE EVALUATION OF COGNITIVE/AFFECTIVE AND SOCIAL FACTORS</td>
<td></td>
</tr>
<tr>
<td>---------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>1. Introduction</td>
<td>319</td>
<td></td>
</tr>
<tr>
<td>2. Personality Factors</td>
<td>319</td>
<td></td>
</tr>
<tr>
<td>3. Cognitive Style</td>
<td>336</td>
<td></td>
</tr>
<tr>
<td>4. Motivation and Attitude</td>
<td>361</td>
<td></td>
</tr>
<tr>
<td>5. The Contribution of Cognitive/Affective and Social Factors to Participant Profiles: case studies</td>
<td>367</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER EIGHT</th>
<th>PHASE TWO AND BEYOND</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Introduction</td>
<td>375</td>
</tr>
<tr>
<td>2. Phase Two Data Collection: an overview</td>
<td>375</td>
</tr>
<tr>
<td>3. Phase Two Target Language Data</td>
<td>377</td>
</tr>
<tr>
<td>4. Phase Two Cognitive/Affective, Academic and Social Data</td>
<td>402</td>
</tr>
<tr>
<td>5. Profile 1 and Profile 2 Connections: towards some conclusions</td>
<td>418</td>
</tr>
<tr>
<td>6. The Research Hypotheses Re-visited: some theoretical, practical and methodological implications</td>
<td>434</td>
</tr>
<tr>
<td>7. The Main Thesis Re-visited</td>
<td>442</td>
</tr>
</tbody>
</table>

| BIBLIOGRAPHY | 447                                      |
APPENDICES

APPENDIX 1: TARGET LANGUAGE TESTS

1.1 Test 1
   Pre-Test
   Post-Test
   Final Test

1.2 Test 2
   Pre-Test/Post-Test
   Follow-up Test/Final Test

1.3 Test 3
   Pre-Test
   Post-Test

1.4 Test 4
   Pre-Test
   Post-Test
   Follow-up Test
   Final Test

1.5 Test 5
   Essay Titles (All versions)
   Marking Scheme

APPENDIX 2: COGNITIVE/AFFECTIVE DATA ELICITATION

2.1 ELTI course feedback sheets

2.2 Experiments 1, 2 and 3

2.3 Motivation/Attitude Questionnaires 1, 2 and 3

2.4 Profile pro-forma

APPENDIX 3: FOLLOW-UP INTERVIEW DATA ELICITATION

3.1 Tutor interview form

3.2 Moller language ability rating form

3.3 Student interview form
FIGURES

1.1 A framework for research foci
12

2.1 Implied relationships between variable types and variable domains
34

2.2 Motivation, attitude and related concepts
47

2.3 Types of cultural interaction in an academic setting
49

2.4 Personality profiles of high and low TL improvers
59

2.5 A continuum classifying and characterising theories of learning and studying
72

2.6 A summary of problems relating to TL-medium study
84

4.1 Relationships of likelihood between test use and type of validity
125

4.2 Schema for education
126

4.3 A logical hierarchy of validity categories (1)
127

4.4 A logical hierarchy of validity categories (2)
150

4.5 Performance - assessment routes summarised
175

4.6 General evaluational requirements
177

5.1 The empirical study: framework, foci and chronology
193

5.2 Summary description of the participant group
206

5.3, Relationships between geographical regions and area
5.4 of specialist training
214
217

6.1(a), 6.1(b), 6.2 Summary statistics on Test 1
253–6

6.3 to 6.15 Summary statistics for Test 2
258–71

6.16 to 6.17 Summary statistics for Test 3
273–5

6.18 to 6.21 Summary statistics for Test 4
277–80

6.22 to 6.25 Summary statistics for Test 5
285–8

6.26 Correlation matrix for pre- and post-test batteries
291

6.27 Scatter diagram of scores on Pre-Test 3 and Pre-Test 4
296
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.28</td>
<td>Correlation Matrix for prognostic validity of Pre-tests for post-test performance</td>
<td>297</td>
</tr>
<tr>
<td>6.29</td>
<td>Scatter diagram of pre- and post-test scores on Test 4</td>
<td>298</td>
</tr>
<tr>
<td>7.1</td>
<td>Scatter diagram relating TL proficiency scores and scores on QI in the Cattell 16 PF inventory</td>
<td>335</td>
</tr>
<tr>
<td>7.2</td>
<td>Scattergram of scores on field dependency and QI</td>
<td>337</td>
</tr>
<tr>
<td>7.3</td>
<td>Summary of group activities during Experiment 3</td>
<td>345</td>
</tr>
<tr>
<td>8.1</td>
<td>Sources and types of Phase Two data and their connections with final participant profiles</td>
<td>376</td>
</tr>
<tr>
<td>8.2, 8.3</td>
<td>Summary statistics for Test 2 (all versions)</td>
<td>379-81</td>
</tr>
<tr>
<td>8.4</td>
<td>Summary statistics for Test 1 (all versions)</td>
<td>382</td>
</tr>
<tr>
<td>8.5</td>
<td>Summary statistics for Test 3 (all versions)</td>
<td>388</td>
</tr>
<tr>
<td>8.6</td>
<td>Summary statistics for Test 4 (all versions)</td>
<td>391</td>
</tr>
<tr>
<td>8.7</td>
<td>Summary statistics for Test 5 (all versions)</td>
<td>394</td>
</tr>
<tr>
<td>8.8</td>
<td>Correlation matrix for prognostic validity of post-tests for final test performance</td>
<td>396</td>
</tr>
<tr>
<td>8.9</td>
<td>Summary statistics for whole test battery</td>
<td>402</td>
</tr>
<tr>
<td>8.10</td>
<td>Graph for U-curve hypothesis</td>
<td>418</td>
</tr>
<tr>
<td>8.11 to 8.15</td>
<td>Phase One presage and programme variables related to Phase Two outcomes for all sub-groups</td>
<td>420-31</td>
</tr>
</tbody>
</table>
CHAPTER ONE

THE EVOLUTION OF A RESEARCH DESIGN
1. **Introduction**

This chapter has the following aims:

1. to explain the impetus for my research: an interest in the role of 'extra-linguistic' factors in language learning.

2. to describe the focus and findings of a preliminary check on the research viability of this interest in the light of recent language educational discussion.

3. to trace the evolution of my research design.

An important assumption in this chapter and throughout my thesis is that explicit comment on the process of arriving at approaches to its concepts and methods has a legitimate and useful place in any research account. I accept, in fact, the educational anthropological model which recognises 'the uncovering and description of the researcher's own assumptions as a fundamental aspect of the investigation itself' (Committee on Anthropology and Education 1978).

2. **The Research Impetus: experience, interest and a 'problem'**

The initial impetus for the research described in this volume came from my previous professional activities in the language teaching field. Whether the main responsibility was teaching, teacher training, syllabus design or the production of materials, it had been my common experience that factors normally described as 'extra-linguistic' played a significant part in the success or otherwise of the language learning and teaching process. Yet the interaction of individual differences of aptitudes, attitudes, personality, motivations and cognitive styles with features of any educational context remained obscure. Consideration of it in the field was most often ad hoc, and theoretical or empirical accounts rarely reached the level of descriptive, let alone
explanatory adequacy. One could agree with statements like Oller and Richards (1973), but not without wondering at their vagueness and feeling that more should be done to investigate their implications:

"Attitudes and resultant motivations may well be more important sources of data for theories of second language learning than are formal characteristics of language" (Oller and Richards 1973 p.233).

Now, my concern really crystallised as a problem when I was working in the area of purpose-specific ELT course design immediately before my research period, analysing language needs and specifying syllabus content for a variety of academic or occupational training institutions. The model I used on most of these assignments was Munby (1978). This approach (discussed in Chapters Two, and Four below) explicitly excludes from consideration at the syllabus definition stage, 'implementational constraints' (p.217) of the socio-political, logistical, administrative, psycho-pedagogic and methodological kind. Yet intuition, experience and evidence of a growing interest in the individuality of people and institutions as well as in the individualisation of learning methods, seemed to question the wisdom of so delaying attention to such real facts of language learning life.

Here was a problem, not only in the practical sense but also in the sense in which the term is used by Popper, for example in Conjectures and Refutations, when he imagines an experienced researcher's advice on starting points for scientific investigation:

"Find out where difficulties arise, and take an interest in disagreements. These are the questions you should take up. In other words, you should study the problem situation of the day. This means that you pick up, and try to continue, a line of inquiry which has the whole background of the earlier development of science behind it" (1968 p.129, original emphasis).

In recommending early recourse to the 'problem situation of the day',
Popper is underlining the importance of the researcher's clarifying the relationship of his research with what has gone before, especially since, in Popper's belief, previous investigation will have left problems resulting from successful refutation of theories, rather than confirmed solutions. The researcher must check the relevant paradigm or paradigms, where this term is used as Kuhn redefines it, as a 'disciplinary matrix':

"... disciplinary because it refers to the common possession of the practitioners of a particular discipline; 'matrix' because it is composed of ordered elements of various sorts, each requiring further specification" (Kuhn 1970 p.182).

My immediate need, therefore, was to check the viability of an interest in 'extra-linguistic' factors as a 'problem' qua research starting point. Key questions requiring early answers were:

1. Was this interest reflected in the focus and tone of the current applied linguistic paradigm?

2. Was it, by definition, likely to lead me into other 'disciplinary matrices'?

3. What early clues were there concerning methods of inquiry for my research?

Whatever answers emerged, the initial paradigm check could help to ensure an investigation that kept its balance:

"A scientist cannot be an isolated rebel who relies on personal observation alone to create fruitful hypotheses. His work will be hopelessly hampered if he is not thoroughly familiar with established facts, existing theories and previous research relating to his problems .... Complete slavery to traditional thought and excessive specialisation in a field, however, may crush creativity" (Van Dalen 1966 p.147).

This early probe into the literature was selective and preliminary to more searching and specific reassessments of previous work that I shall
undertake when they are required for the formulation of the working hypotheses specified in my actual research design (see Section 5 below).

3. 'Extra-Linguistic' Factors and Language Learning: some views from the applied linguistics paradigm

Even a tentative sampling of recent applied linguistic discussion reveals considerable evidence that a focus on learner, learning and educational contextual factors is in tune with the times.

Characterising the 'latest stage of development which we are currently passing through' as 'the learner-centred stage', Strevens (1980) sees it as the re-acceptance, after a break under the sway of audio-lingualism, that language learning:

"entails the interaction of at the very least three elements: the learner; the teacher; the presentational design (.... 'the method'). And the history of language teaching demonstrates a slow, irregular, lurching progress towards ever-greater awareness of the complexity of each of these elements and therefore of the actual and potential interactions among them" (p.19).

Strevens' whole account of the situation is significant for this research in its emphasis on the complexity and range of factors to be taken into account, the divergence in research perspectives and the crucial mutability of key learner variables. Each of these insights is taken up later.

For Oller and Richards (1973):

"The focus is on language learners - their capabilities, attitudes, strategies and, of course, what it is they learn" (1973 p.v).

In their collection, the focusing is done through various lenses by various investigators: cognitive strategists, language acquisitionists, testers, socio-psychologists and methodologists. It was instructive that statements of the perceived influence of extra-linguistic factors
so regularly indicated (though to a varyingly explicit degree) that its study should involve a broad-ranging approach making full use of insights from other disciplines. Answers to my second question were in fact embedded in answers to my first. The problem was widely recognised, the informing paradigms were often identified, yet the tone was still tentative and speculative, as the following selection of views indicates:

"Now, for the first time since Gouin wrote his famous *L'art d'enseigner et d'étudier les langues* (1880), we are attempting to take a genuinely inter-disciplinary approach to second language research. Neither linguistics nor psychology nor any other discipline alone will produce final answers. The next decade should provide us with information about the human person which, when creatively interrelated with our accumulated knowledge in theoretical linguistics, sociolinguistics, psycholinguistics, and the psychology of learning, will possibly enable us to construct a viable theory of second language acquisition" (H D Brown 1973 pp.242/243).

"Affective factors have at least as much influence on language learning as do ability factors" (Chastain 1975 p.153).

".... I would highlight the importance of individual differences .... although the research literature has not yet caught up with and classified the many cognitive variables involved, the task is clearly important. It concerns not only individual differences of pace and concentration, and cognitive variables .... but also very personal reactions to particular instructional activities ...." (S McDonough 1978 pp.147/148).

"Personality tests have existed for a long time yet how many times have they been used to help language learners?" (Pickett 1979 p.5).

".... recent basic research on second language acquisition and use .... shows considerable promise as a bridge between the masses of data available in the foreign language classroom and the consideration of such theoretical questions as the nature of the ability to acquire languages beyond the age at which first language acquisition ordinarily takes place and to use languages so acquired. This recent work thus brings the relationship between foreign language and second language teaching and its basic disciplines .... closer in character than it had previously been ...." (Ritchie 1978 p.1).

One noted here and there indications that a multi-disciplinary approach is not only logical given the 'extra-linguistic' focus but also
necessary to combat the tendency of applied linguists towards over-

generalisation or lack of precision:

"We are often guilty of using rather sweeping terms as if they were
carefully defined" (H D Brown 1973 p.232).

"While most teachers recognise the importance of the social aspect
of language they tend to place all socially oriented language
learning problems into a single category of 'culture'
differentiation" (ibid p.235).

"Contemporary psychological findings indicate that the attitudinal-
motivational variable is probably the single most important factor
in academic success" (Hancock 1972 p.225).

yet:

"Motivation is whatever makes the learner ready and willing to
rearrange his own molecules, but what is that?" (Stevick 1971
p.21).

"The present stage of research into the acquisition, learning and
teaching of languages is so early and so fragmented, and work done
up to now has its origins in such disparate fields of study, that
there is no universal acceptance of what the most important variables
are ...: every new formulation by a different researcher for his
own purposes provides some fresh illumination for others" (Strevens

In such comments the beginnings of answers to my third question seemed
to be emerging, on the type of research approach that might be the most
viable and fruitful means of investigating the factors I was interested
in. A stage of inquiry that was characterised as 'fragmented' and
disparate as regards disciplinary matrices needed further exploration
that was rooted empirically in a real learning and living situation.
Support for such an inference was fairly easy to find though it was by
no means consistent in its specific recommendations:

"That there are great problems for any method of investigation of
anything so large as language learning is obvious; but it is also
obvious that no one method can be sufficient alone. There have been
many experimental investigations, few of them have been capable of
supporting generalisations; survey data of various types exist, often
highly dependent on the original test population and, again,
offering only weak support for theoretical statements; systematic
observation of FL classroom learning has been in progress for some
years, but has been hampered by the lack of analytical systems" (McDonough 1978 p.133).

"The suggestion of a model based on sociocultural distance is clearly in need of supportive empirical research .... Anecdotal evidence, introspective observation of learners of a second language in a second culture is supportive so far. Further support needs to be sought from longitudinal studies of learners with careful measurement of sociocultural change and language success" (H D Brown 1980 p.162).

".... I deem it important to establish what kinds of language abilities can be demonstrated to exist, regardless of their possible relevance outside the testing room or laboratory, because until the existence of such abilities is well established, their possible ecological validity cannot even be determined. It is conceivable, however, that some of the more important language abilities can be established only in realistic, non-testing situations" (J B Carroll 1979 p.22).

"If we accept that both statistics and scaling are useful tools in language research, how can we use them together to best advantage? An analysis of variance could help a researcher determine which variables frequently co-occur and help the researcher to focus on those items that may be most productive for scaling .... We feel that a valid criticism of traditional experimental methods is that they tend to limit the examination of experimental data to whether or not the data confirm or refute a previously formulated hypothesis .... The researcher should be flexible and free to examine data from all angles without preconceived hypotheses as limitations, but all data should be taken into account" (Madden, Bailey, Eisenstein, Anderson 1978 p.122).

It seemed reasonable at this stage to abstract the following answers to the three key questions focusing my paradigm search:

1. The issue of the influence of 'extra-linguistic variables' was indeed receiving considerable attention from applied linguists, but the questions being asked and the answers being so tentatively offered revealed an uncertainty inviting further investigation.

2. The individuality, mutability and the complex inter-relationships of factors suggested as relevant yet often still ill-defined, seemed to call for an even more multi-disciplinary perspective than had already been attempted or advocated.
3. Further inquiry would most fruitfully be made through a variety of methodological channels but with an empirical focus on individuals learning and using another language in a lasting and real social context.

Given these answers, particularly the third one, I decided to find out whether a group of mixed background, mixed speciality overseas students newly arrived in Britain for various types of training would be a suitable and willing group as subjects for empirical research. From July 13 to August 24 1979 the British Council was giving pre-sessional language and orientation to such a group at their English Language Teaching Institute (ELTI) in London. I had already been involved in the needs analysis, syllabus specification and materials production activities carried out in preparation for this programme (see Hawkey, Liesching, Stimson 1981). I now undertook extra teaching and evaluation responsibilities, the results of which provided me with certain procedural insights and pilot data discussed in Chapters Four to Eight below. A similar programme for a parallel group of overseas students was planned for July and August 1980. I was given permission to enlist their help as participants in the main body of my empirical research.

Thus, as a result of preliminary reading and pilot group contact, the following research topic emerged, to be designated according to the design proposed as my pre-thesis, that is a statement of intention as yet uncommitted in terms of methods of inquiry:

'An investigation of inter-relationships between personality factors, cognitive style and language learning with particular reference to adult, non-beginner learners from overseas requiring English in connection with their specialist training in the UK.'
Implications for the training of such learners drawn from empirical research carried out with a group of Overseas Development Administration Study Fellows.

It was clear from this pre-thesis that the research canvas would need to be a broad one. The aim was to discover as much relevant information as possible about a particular group of individuals: information about what they brought to the new learning and socio-cultural situation, information about events, processes and progress during the experience and information about outcomes. Figure 1.1 presents a tentative framework for my research foci. It is informed by insights from Biggs (1978) in the field of educational psychology, Pilliner (1974) on programme evaluation and from the empirical study, The Good Language Learner, (Naiman, Frohlich, Stern and Todesco 1978).

The diagram in Figure 1.1 is not a comprehensive descriptive model, certainly not an explanatory one. At this stage in the evolution of the research design it serves as a summary of potential variables with indications of the chronology of data collection for their investigation.
We have a problem in search of a research design.

4. **Methods of Inquiry: the scientific and alternative paradigms**

The search in Section 3 for corroboration of the significance of my problem was confined to a particular disciplinary matrix and, in the main, to the theories rather than the methods of that paradigm. But a Kuhnian paradigm embraces 'law, theory, application, and instrumentation together' (1970 p.10). Here, the focus must be on 'application and instrumentation' and my need now is to establish a general methodological approach for my study. For that purpose insight will be sought from several different paradigms.
An ambivalent attitude towards the hypothetico-deductive scientific method of inquiry is characteristic of discussions of educational research. Often, the classical route from systematic observation through hypothesis formulation and hypothesis testing to predictive theory is put forward as the ideal model. Its suitability in educational research, however, is then questioned on the grounds, for example, that 'the notion of a human society involves a scheme of concepts which is logically incompatible with the kinds of explanation offered in the natural sciences' (Winch 1958 p.72). The methodological difficulty caused by the 'scheme of concepts' in a human society is presumably the fact that educationalists tend to be dealing with human activity, with phenomena 'which will normally be subject to human purposes and meaning in a manner in which natural phenomena are not' (Bantock 1965 p.155). The difference is not just a quantitative one, that there are more variables around when you are investigating human behaviour. There is, more crucially, a qualitative difference in the nature of the factors involved and in the events influencing their interaction. Yet, surely illogically, discussions of educational research methodology often identify instances where investigations are considered to have failed because they have erred from the path of the scientific method. The confusion is encapsulated in this excerpt from the Open University course in Methods of Educational Enquiry:

"It is, however, noticeable that good illustrations of the hypothetico-deductive method have to be drawn from the natural sciences. Although the behavioural sciences, such as psychology, have attempted to follow the same procedures, their success has been less marked. Indeed .... it has been suggested that this model is inappropriate in much educational research. The strength of a scientific theory is its ability to predict future events. The reasoning is 'a priori' - beforehand. This direction is important. It is too easy to explain away results; reasoning 'ex post facto' - after the event. Much educational research suffers from this weakness .... On the other hand, in the early stages of research,
where it is difficult to put forward definite hypotheses, ex post facto reasoning will be unavoidable. There is still room for such exploratory research in education, but with the recognition that exploration should lay the foundation for more definitive studies. Exploratory studies do not provide hard evidence in themselves" (Entwistle 1973 p.16/17).

Two key questions arise from this debate, questions that have to be answered before I can propose a design suitable for my research purposes:

1. How can the classical scientific method inform the kind of research I am undertaking?

2. Which alternative approaches can be of use?

A fact that may be under-estimated in discussion of the first question is that the crucial unit of currency in the scientific method is the hypothesis, conventionally specified as two or more precisely related variables (H W Smith 1975 p.40), quantitatively operationalised (Wallace 1969 p.ix) and controllable; initially narrow in scope (Willer and Webster 1970 p.755) and empirically testable to falsification (Popper eg 1968 p.228ff). The rigorous process of hypothesis formulation and testing probably has its place in most educational research but, as has already been indicated, there are powerful reasons why it may not fit as a model for the overall shape of a research design or, exclusively, for those parts of a design where specific hypotheses are under investigation. The whole epistemological history of scientific research itself reveals arguments against the postulation of a single method of inquiry. Kuhn (1970) warns of the potential restrictiveness of existing 'normal scientific' paradigms though a balance must be kept and 'novelty ordinarily emerges only for the man who, knowing with precision what he should expect, is able to recognise when something
has gone wrong' (p.65). Popper (1968) stresses the importance of critical thought and contradiction and, while feeling that science should be applied to social problems, warns against the 'pseudo science' sometimes invoked to predict events in modern society even when it is not a 'well-isolated, stationary and recurrent system' (p.339). He seems far away indeed from the world of observable, measurable and predictable phenomena when he sees the main function of social science as tracing 'the unintended social repercussions of intentional human beings' (p.342).

The first message here is that even in the natural sciences and in the brands of social science that are modelled most closely on them, it is only the specific hypothesis-testing stage that is tightly linked to systematic observation, induction, operationalisation and testability. This message will be reflected in my own research design. Before we reach the level of operationalised hypotheses, however, intuition, experience, critical logico-deductive thought are a crucial explicit part of the inquiry. They too must be accounted for in the design.

A second message can also be inferred, namely that it is unacceptable or, at best, merely pseudo-scientific to attempt to operationalise and test what is not susceptible to such treatment. ('How can anxiety be given an operational definition?' Kelly (1955 and see below). No true scientist would wish any researcher to try, but it is not reasonable to invoke the rules of the hypothesis-testing part of scientific methodology to prohibit the investigation of facts and factors that cannot claim full variable status. If allowing such facts and factors into the investigation categorises research as 'exploratory' rather
than 'definitive' (in Entwistle's terms above), the former label need not imply 'second best'. Given the nature and stage of applied linguistic inquiry in the area of 'extra-linguistic' variables, the pattern should be to explore extensively but with an eye for certain spots where 'definitiveness' may be approached.

And there is a third important message. Scientifically rigorous hypothesis testing often demands fully controlled experimental conditions. But the problem is (see, for example, Cole (1975) or Neisser (1976)) that the more controlled, the 'purer' the experiment, the less generalisable to real-life conditions it becomes. Even worse is a danger pointed out by Howmans (1954):

"Hypotheses are often selected merely because of the facility with which they can be given quantitative demonstration" (p.729).

So, mindful of some of the problems associated with the scientific method in educational inquiry (though by no means rejecting its insights), in which other directions do we look for answers to the second key question?

With my commitment to an investigation that will attempt to take into account the clusters of factors identified under presage, programme, community, process and product in Figure 1.1, my research design needs to be informed by paradigms that are traditionally prepared to handle such complexity. For the general insights needed at this stage in the evolution of the design, I shall, therefore, look to the psychology of personal constructs, social anthropology and illuminative evaluation.

Both George Kelly's psychology of personal constructs itself and the
methodological stance it favours are relevant. Kelly's theory (1955) focuses on how people construe events and organise their constructs into systems, constantly using personal experience to modify, refine and enrich these constructs, the better to anticipate and cope with the future. Rejecting the excesses of both the psychometric and the humanists schools (irreverently characterised as interested in 'man-the-biological-organism' and 'man-the-lucky-guy' (Kelly 1955, Volume 1 p.4)), Kelly regards the individual as an active, self-changing 'man-the-scientist', constantly developing and trying out hypotheses prompted by personal reaction to experience. The emphasis of this constructive alternativism on the individuality of the process of trying out constructs and their alternatives 'in an infinite series of successive approximations' (Kelly 1955, Volume 1 p.15) has clear implications for my own investigation. This is especially true where I follow the progress of individuals under the extra pressure to re-construe in a foreign academic and social setting. It also explains my efforts to trace the developing relationship between their expectations and satisfactions (see Chapter Eight below).

And on the subject of research methods, Kelly and his supporters are equally instructive. Kelly's own position is eclectic. He sees a place for hypothetico-deductive, hypothetico-inductive and even for 'statistical dragnet' procedures in research (1955, Volume 1 pp.32-34). All have their part to play, though he is wary of the ingenuities sometimes associated with the latter, reminding us that they do not actually generate new ideas; they merely provide quick and reasonably sure validation of what has already been hypothesised. Perhaps other members of the psychology of personal constructs school are less
forgiving than Kelly about the dominance of the psychometric tradition (perhaps because they are commenting two decades later in a less restrictive era). Georgi (1977) counts the cost of the rule of the scientific inductive method, taken over, he feels, unquestioned from physics then taking over a field in which it is not necessarily appropriate. When scientificness becomes the main prerequisite for validation, qualitative description and analysis may be accepted only in the speculative, pre-empirical stages of research and then only if accompanied 'by an apologia twice as long as the description' (p.12). Phillida Salmon (1977) finds the demand for 'hard-nosed' objectivity, especially when it comes prematurely in the research process, counter-productive:

"The greatest importance is attached to whether or not the applicant can frame his question in terms of a standard research design and specify the form of statistical analysis he would use on his results. The question itself is simply not seen as an issue for discussion" (p.36).

I shall attempt to let the design and statistics serve my question (or problem) not vice-versa.

The second alternative source of methodological insights is social anthropology. Its general influence on the framework I am about to propose is captured by these words from Lévi-Strauss (1960):

"Yet it happens that in anthropology experimentation precedes both observation and hypothesis. One of the peculiarities of the small societies we study is that each constitutes a ready-made experiment because of its relative simplicity and the limited number of variables required to explain its functions .... By comparison with the natural sciences, we enjoy an advantage and suffer from a handicap. We find our experiments already set up but we cannot control them. It is therefore natural that we attempt to substitute for them models, or systems, of symbols which preserve the characteristic properties of the experiment but which (unlike the experiment) we are unable to manipulate" (p.15).
The 'small society' constituted by my group and the 'small societies' in which they subsequently find themselves will be explored in this research in both controlled and uncontrolled ways. Dell Hymes (1979) adds an extra specific dimension, has a warning for both the controllers and non-controllers, but strongly favours anthropological approaches in educational research. The extra dimension is the ethnological, which stresses the:

"examination of the ways in which larger forces for socialisation, institutionalisation, and reproduction of an existing order are expressed and interpreted in a specific setting" (p.5).

His warnings are, on the one hand, that:

"A mode of research that focuses on experimental design, quantitative techniques, and the impersonality of the investigator has its place; but, carried to its perfection, as the exclusive mode, it would tend to divide society into those who know and those who are known" (p.7).

and on the other, against:

"letting the anthropological perspective on education become equated in other minds with just a mode of field work. The result will be dozens of people called 'ethnographers' because they have observed, although with little or no training in cultural analysis; attempts to insert 'ethnographic components' into research designs ...." (p.5).

But his preference is for an 'integrative' approach to educational research with full weight given to the socio-cultural context and for work that is 'cumulative, comparative, co-operative'.

The design to be proposed below uses controlled experiments (see Chapter Seven) quantitative techniques (Chapters Six to Eight) and reports on observation (see Chapter Seven). By its longitudinal definition it is cumulative and comparative and in its real-life course design context (ie Phase One) it is co-operative.

The third alternative paradigm is an offspring of the second, namely
illuminative educational evaluation as described, for example, in Hamilton et al. (1977). Again, there is an explicit reaction against dealing mainly in the currency of the scientific hypothesis (described by Hamilton et al. as from the 'agricultural-botany paradigm' (p.7)). Illuminative evaluation is represented as 'a general research strategy' rather than 'a standard methodological package'. It requires the acceptance of the 'untidy reality' of the learning milieu:

"The learning milieu represents a network or nexus of cultural, social, institutional and psychological variables. These interact in complicated ways to produce, in each class or course, a unique pattern of circumstances, pressures, customs, opinions and work-styles which suffuse the teaching and learning that occur there" (Parlett and Hamilton 1977, p.11).

The researcher is warned that too great a tendency to think in terms of parameters and factors rather than individuals and institutions may divorce a study from the real world. Illuminative evaluation 'stands unambiguously within the alternative anthropological paradigm'. Methodologically it deals in 'description and interpretation rather than measurement and prediction' (Parlett and Hamilton 1977, p.10). Parlett and Hamilton claim that this represents an actual paradigm shift in the sense in which Kuhn defines the term:

"The transition from a paradigm in crisis to a new one from which a new tradition of normal science can emerge is far from a cumulative process, one achieved by an articulation or extension of the old paradigm. Rather it is a reconstruction of the field from new fundamentals, a reconstruction that changes some of the field's most elementary theoretical generalizations as well as many of its paradigm methods and applications" (1970 pp.85/86).

But such fundamental change is not necessarily entailed. Sometimes the implication is of greater eclecticism and flexibility rather than fundamental change. L M Smith's 1971 model qualifies as illuminative evaluation but without being entirely radical:
"The model sought to bring together three research strategies. 
(1) An experimental design with pre- and post-tests of achievement, 
control groups and inferential statistics. (2) A second strategy 
was the social survey with interviews and questionnaires, random 
sampling of programme-relevant individuals (teachers, parents and 
pupils), quantification and cross-tabulation of response. (3) The 
third 'leg' of the model was a participant observer study of the 
programme" (p.203).

So, the researcher may fit his methods to the particular focus of each 
part of his investigation; it is legitimate to combine apparently 
contrasting techniques taken from different paradigms. Where 
quantitative data collection and analysis are appropriate, for example 
to establish an extensive data base before the focus is switched to 
more dynamic, emergent issues, psychometric techniques should be used. 
Where less predictable phenomena need to be explored, the research 
design should allow for more naturally elicited information and ex 
post-facto hypothesising.

Although much of the methodological discussion in the social science 
literature stresses its scientific pedigree, there is, then, a more 
eclectic tradition in the discipline which reflects the more open-ended 
approaches of the alternative paradigms I have examined briefly and 
selectively here. Hovland's classic paper on attitude change studies 
is quoted by H W Smith (1975) in support of his plea for 'a social 
research norm that gives lowest degrees of confirmation to propositions 
confirmed by only one method and higher degrees of confirmation when 
multiple methods are used' (p.292).

"Integration of the two methodologies will require on the part of 
the experimenter a greater awareness of the narrowness of the 
laboratory in interpreting the larger and more comprehensive effects 
of communication. It will require on the part of the survey 
researcher a greater awareness of the limitations of the 
correlational method as a basis for establishing relationships" 
(Hovland 1959 p.13).
My own study will use a combination of methodologies not only because it has to perform a variety of functions but also because (see especially Chapters Six to Eight below) correlational techniques alone are inadequate. The combined methodologies approach is, it seems, always acceptable, usually advisable and sometimes essential.

We can thus abstract the following summarised answer to the two key questions on research methods:

The classic scientific method will inform my design both at the level of overall structure, where it encourages explicit account of the general pattern of logico-deductive and empirical inductive reasoning, and at the level of narrower hypothesis testing, where it can make its most justifiable claims to authority. The alternative paradigms discussed will contribute to the design where the need is for more qualitative description and interpretation of foreseen and unforeseen processes. Quantitative and qualitative inquiry methods will frequently be combined.

5. Towards a Research Design

This chapter, even thus far, is part of my overall research design in that it has established my broad problem area ($P_1$), checked on its theoretical currency, formulated the Pre-Thesis and looked for insights into appropriate methods of inquiry. In its overall shape, the chapter resembles the next three chapters of this research, all of them part of the logico-deductive process which leads from a critical re-assessment of relevant disciplinary areas to the formulation of a 'research hypothesis', Bennett's 1973 term for 'the general problem or idea (put) into a form which enables us to investigate it' (p.19, my emphasis). Re-formulating my Pre-Thesis (see the end of Section 3
above) in the manner Bennett is suggesting, I arrive at my Main Thesis (T): *T = Training for EAP learner/users in a foreign culture (C2) can be fruitfully informed by the systematic, sensitive, multi-dimensional profiling of individuals.*

The task of the remaining three pre-empirical chapters is to re-assess areas of theoretical discussion or empirical research relevant to the formulation of the general research hypotheses requiring investigation if my Thesis is to be supported. Thus Chapter Two re-assesses previous inquiry into the relationships between 'extra-linguistic' factors and target language (TL) learning to inform my decisions on appropriate foci and methods for my own study. This re-assessment leads to the formulation of the first Research Hypothesis (RH₁):

\[
\text{RH₁ = individual and combined cognitive/affective and social factors TL learning and use}
\]

Chapter Three re-assesses another, not entirely discrete, area of inquiry equally necessary to the meaningful exploration of my Thesis, namely factors that have been found relevant to the success or otherwise of students training in a foreign culture. Again previous methods of inquiry as well as their foci will be examined, both in the formulation of RH₂:

\[
\text{RH₂ = detailed learner profiles prediction of C future}
\]

Chapter Four, the last of my pre-empirical chapters, re-assesses previous approaches to the elicitation and assessment of TL data and develops a framework for my own necessary attempts to evaluate TL proficiency and progress. In general hypothetical terms:
RH₃ = TL elicitation/ assessment
learner profile data

Now, these Research Hypotheses are not the parsimonious, operationalised, empirically testable kind discussed in connection with the normal scientific paradigm in Section 4 above. Each of them will require later analysis into narrower component hypotheses where deduced consequences can be tested against factual evidence. Nor are my Research Hypotheses independent. Clearly, for example, RH₂ overlaps RH₁. Having identified key 'extra-linguistic' factors and their supposed inter-relationships with TL for RH₁, I re-examine these inter-related factors in the foreign cultural context for RH₂. And RH₃ differs in kind from both in that it is essentially facilitative, instrumental, methodological rather than phenomenological. After all, unless valid measures of TL proficiency and progress can be established, none of the issues crucial to the Thesis can be explored.

The process in the first part of my research seems close to Willer and Webster's process of abduction, a process that aims 'to formulate possible .... theories for later testing and to explicate the concepts which are the subject of the theories' (1970 p.754). The second part, the empirical study itself, represents this 'later testing'. In this part of the research design the broad Research Hypotheses are analysed into testable components with recourse to standardised data, experimental group treatments, independent and dependent variables with statistically explored causal relationships. But, as signalled in Section 4 above, the methodological approach of the study is eclectic. Perhaps it would be accurately described as an explanatory descriptive study, making use of survey, case study, causal-comparative,
correlational and developmental data, to use Van Dalen's 1969 attempt at a classification of inquiry methods (pp.285ff). One of the consequences of my mixed-method model is that both a priori and ex post facto hypotheses will be allowed. Another is that the data collected through fundamentally different means (eg standardised cognitive style test observations vs. introspective self-reporting) will be used for mutual corroboration or refutation purposes.

The developmental/longitudinal theme of the empirical study is reflected in its division into two main phases. Phase One covers the investigations carried out during the intensive six-week period when I was in daily contact with the participant group. Phase Two traces events during the following 9 months or so. All three Research Hypotheses depend on data from both phases, as would be expected given the study's emphasis on the prognostic validity of learner profiles. The detailed design for the empirical study appears as Figure 5.1 in Chapter Five below, where foci, methods and chronology are summarised.

At the end of the report on the empirical study, connections are renewed with my Research Hypotheses and main Thesis. The concluding summary of the educational implications of the study signals fresh problems (P₂ in Popper's terms) for further investigation.

This chapter has traced the development of a research design from an initial problem. It has also indicated the methodological approaches that seem most appropriate for my investigation.
26.
CHAPTER TWO

KEY LEARNER AND LEARNING VARIABLES
1. Introduction

The purposes of this chapter are the following:

1. to discuss the distinctions and overlap between factors normally considered to influence language learning

2. to re-assess the foci of previous theoretical and empirical investigations of variables related to learning another language, making connections with my own inquiry

3. to construct and use a descriptive framework for the features of target language use that are to be explored in inter-relationship with these variables

4. to posit the first general Research Hypothesis (RH₁), derived from the re-assessment of relevant variables in the chapter.

As the title and the research design description in Chapter One, Section 5 show, this chapter has a key theoretical role in the overall structure. The variables introduced here are the essence of my initial problem. In research design terms RH₁ is superordinate to RH₂ and RH₃, which exist as the means of testing it.
2. Potential Variables: nature, variety and inter-relationships

Chapter One has already indicated that support for research taking account of the numerous inter-related factors affecting second language learning and acquisition is plentiful in the applied linguistics literature. Key themes, it will be remembered, were that the focus should be on the learner (Oller and Richards 1973), the scope should be 'global' (Schumann 1976), the disciplinary matrices 'disparate' (Neufeld 1979), the variables 'countless' but 'interacting' (Brown 1973) and the perspective longitudinal (passim). Although not every study to be considered here in support of my major research hypothesis in fact takes a really global view of learners, it is interesting to note that the complex links between the variables themselves often makes a multi-dimensional approach inevitable as well as, in my view, desirable. It is almost impossible to isolate the variables even if you want to. Some researchers find this inevitability a fruitful fact of language learning life:

"... we have been exploring the motivational system and the part it plays in second-language learning. But as we examine the many ramifications of the language learner's motivation, we see that it can affect and be affected by the other essential components of learning as well" (Gardner and Lambert 1972 p.134).

Others see theoretical and methodological dangers:

"(Carroll and Pimsleur's) data base inevitably entailed many variables outside language learning ability, including social pressures and norms, individual psychological traits, cultural biases, student-teacher relationships and so forth. Since these researchers made no attempt to isolate these characteristics, they could not possibly have separated language learning ability from other classroom learning variables" (Neufeld 1979 p.232).

No one seems to deny the difficulty of handling either the desired or the inevitable overlaps among variables:
"Of course the possibility that one's theoretical constructs may not be universally understood or correctly interpreted is an occupational hazard in any scientific endeavour. In applied, interdisciplinary, transpositional research such as ours it is even more problematic" (Guiora and Acton 1979 p.194).

In the statement of my initial problem in Chapter One the factors emerging from language learning and teaching experience as worthy of investigation were tentatively described as 'extra-linguistic'. Now the use of this term is problematic at several levels. The problem at the philosophical level is summed up by Stern in Oller and Richards (1973) as one of

"the intricate relationship between language and meaning, between language and thought and emotion, between language and culture. The questions that have puzzled the scholars here are whether language is shaped by these various factors, or merely reflects them, or, whether language, in turn, exercises a shaping influence on them. The general tenor of the discussion suggests the close interaction between language and many other factors" (p.25).

There can be very little that could accurately be labelled 'extra-linguistic' when language is assigned its fullest thought- and event-shaping role. Methodologically, too, Stern's implication that it is hard to be sure of the direction of causal relationships between language and the other variables will be taken seriously.

At a second level, with a Chomskyan limitation of the scope of the term 'linguistic' to 'structural descriptions assigned by (the) grammar', it is possible to refer to 'personality, beliefs and countless other extra-linguistic factors' (Chomsky 1970 p.10, my emphasis). But this restricted view of competence is not in line with the broad communicative competence construct developed in Chapter Four and assumed in the way I collect and evaluate my data. When your potential hypotheses, however, are about relationships between variables such as those to
be discussed in Section 3 here and TL features described in the broad communicative way exemplified in Section 4, there is obviously a danger of having at least parts of the same phenomenon at both ends of a hypothesized relationship, of overlap between independent and dependent variables. Although this is in practice a problem to be faced at the hypothesis-testing rather than formulation stage, some of its implications will emerge in this Chapter, for example when I attempt to classify learning processes, strategies, skills, habits in 3.6 and language in study use in Section 4.

And at a third level, of very practical relevance to the researcher, the complex relationship between language and the events of real life again poses problems. Oller and Perkins (1978) claim that language proficiency is a source of 'non-random but quite extraneous variance in measures of affective variables' (p.85). In some cases, language, especially the target language when it is used as a medium for the elicitation of affective data, actually hinders the investigation of what Oller himself (1979;1) somewhat paradoxically calls the 'extra-linguistic context', that is:

"...states of affairs constituted by things, events, people, ideas, relationships, feelings, perceptions, memories and so forth" (p.19).

Oller and Perkins' warning of the interference of the target language in the investigation of affective factors is noted and will be examined in its empirical context in Chapters Six to Eight. But clearly his broad interpretation of the term 'extra-linguistic' is inappropriate in the predominant communicative competence paradigm. But then Oller, like others, is using the term as a convenient shorthand. More accurate, though as we shall see below, not entirely non-tautologous, are
descriptions such as 'affective, cognitive and social factors' (Tucker, Hamayan and Genesee 1976 p.216), 'attitudinal/motivational characteristics' and the 'socio-cultural context' (Gardner, Smythe and Clément 1979 p.319), Schumann's 'initiating factors' (social and psychological) and 'cognitive process' (1976) and Chastain's 'affective and ability factors' (1975).

The label 'cognitive/affective' is useful since it covers the domains of cognition and emotion and suggests the kind of inter-relationships borne out in research. The label 'social' speaks for itself; it is applicable whether the reference is to the narrowest aspects of inter-subjectivity or to the broadest influences of whole cultures. In Section 3, then, I attempt to re-assess theoretical and empirical insights of a cognitive/affective nature, in Section 4 a way of describing language in study use. This re-assessment should establish the viability of a major hypothesis seeking to relate the two.

3. **Cognitive/affective and Social Factors: theoretical discussion and empirical findings**

The inquiry assessed in this section is explored under headings relevant to my own problem and proposed methodology. It will quickly become obvious that the 'global' scope and longitudinal perspective emphasise the inter-relatedness of variables. Figure 2.1 below gives an impression of this with the headings I have selected for this section plotted against the three main variable domains:
Figure 2.1: Implied relationships between variable types and variable domains

The diagram thus suggests that cognitive style has affective and social connections, is not purely a matter of cognition (see 3.5 below) while personality (3.4) has social and cognitive as well as affective implications. The positions plotted on the diagram are impressionistic and would vary from study to study. The message that nothing is clear-cut in people-centred research, however, is accurate enough.

3.1 Time
As might be expected in a study which is longitudinal but not over a period spanning the movement of individuals from one critical age to another, time is more important here than age. Whereas I am interested in the points in time at which various learning experiences take place, in changes in communicative performance or attitudes over
the period spent by participants in the UK, I do not have evidence on the role of age as a process factor. However, there is more applied linguistic research focusing explicitly on age than on time and since much of it has something to say about adult learners like my group, it repays a brief re-assessment for possible insight into age as a presage factor.

Inquiries into the optimal age hypothesis often find their impetus in neurophysiological studies such as Penfield and Roberts (1959) and Lenneberg (1967) which suggest that the inevitable, irreversible decrease in left-hemisphere cerebral plasticity is a biological reason why adults cannot achieve phonological ambilingualism whereas children can. Findings from studies testing this hypothesis, which tend logically enough to compare adult and child language acquisition, are of interest for what they suggest adults can or cannot do. In a useful survey of eight such studies Krashen, Long and Scarcella (1979) find the following adult-versus-child differences:

1. learner/users arriving in C2 as adults do not achieve such high levels of proficiency in TL as learner/users arriving as children.
2. adults acquire TL more quickly and efficiently than young children but not always than 12 to 15 year-olds in the 'early stages of morphology and syntax' (p.576).

From other sources additional relevant points about adult learner/users may be made:

3. that they already have more abstract concepts and lexical labels in L1 (Ausubel 1964, Butterworth and Hatch 1977)
that they may well, because of their greater conceptual range, be more concerned to express a wider variety of meanings in L2 with a possible resultant downgrading of formal accuracy (Ervin-Tripp 1974, Ramirez and Politzer 1975, Hatch 1978).

that they already have fully developed cognitive and memory heuristics and are thus helpfully predisposed to analyse (Butterworth and Hatch 1977, Ervin-Tripp 1974) with a higher chance of attaining 'cognitive/academic language proficiency' (CALP) (Cummins 1980).

And on the question of time independent of age it is worth remembering that in mastery learning theory (eg Bloom 1971) time is the key variable:

"... for most of the tasks in the regular school curriculum, it can be expected that every student will reach (a required) criterion if given enough time - a time within reason" (J B Carroll 1970 p.73).

The more, broadly speaking, the better, with intensity a vital interacting factor:

"... as intensity goes up the learning-effectiveness of each hour of teaching goes up more than proportionately. In other words, at 25 hours per week, a 100-hour course engenders more learning than does the same course given at 5 hours per week" (Strevens 1977 p.29).

But as Walberg, Hase Hecho and Pinzur Rasher (1978) point out, a law of diminishing returns operates in relation to time and stage of TL learning. You cannot go on gaining new competence as quickly later in the process as you can at the beginning.

It is significant in the light of the comments prefacing this section that many of the age or time-related findings in the studies surveyed inter-connect with other factors. The 33 year-old Alberto has not
fossilized at his restricted level of pidgin because of his age but rather, it is suggested, because of a whole complex of factors of social and psychological distance (Schumann 1977). There is no simple answer to the optimal age question:

"The question of when foreign languages are to be taught .... is a complex problem that involves political, social, philosophical considerations and should not be reduced to a matter of neuro-physiology as it has become fashionable to do in recent years" (Jakobovits 1970 p.73).

There are thus several interesting time-related language learning and acquisition issues on which my empirical study may throw some light. I am certainly interested in variation in rates and curves of TL progress. The six-week remedial programme which forms Phase One of my study should be revealing here, especially in comparison with the longer period of Phase Two when my participants' main focus is switched from TL preparation to TL as the medium of their actual specialist training. The two phases will also be informative as regards the time:learning:acquisition question in that most of their TL experience during the second phase is 'naturalistic' rather than 'formal'. And since the C2 durations of participants range from five months to a year, there is scope for an evaluation of the possible TL advantages of longer periods of residence, perhaps with evidence one way or the other on the law of diminishing returns. Also, the age variations among my participants may just reveal differences that go beyond the rare references to the inter-adult age-based factors mentioned in the literature, for example Snow and Hoefnagel-Hohle's 1975 comment that 'learning how to learn abilities' are strong until 40+.

But my design (in both senses of the term) accepts the complexity of inter-variable relationships noted by other researchers. The time
factor will not be forced out of its context as a variable central to any specific hypothesis. It will be discussed if and when it appears to offer useful insight.

3.2 **Aptitude**

There would seem at first glance to be at least three good reasons why it is unnecessary to consider language aptitude in its conventional sense in connection with this research.

1. The concept of language aptitude is associated with non-TL test instruments designed for the purpose of 'screening a student for language study' (Pimsleur 1966 p.185). Since all my participants are already experienced learners of the target language with no alternative but to continue with it, a concern with their potential for language learning might seem belated.

2. The foci of standard language aptitude tests such as the Carroll-Sapon Battery (1959) with its components of artificial number learning, phonetic script, spelling clues, words in sentences and paired lexical associates seem too rooted in the structuralist tradition, too restricted to linguistic rather than communicative competence to suit the broad construct of my inquiry.

3. The concept of aptitude, when it is allowed the rather broader scope that Pimsleur's *Language Aptitude Battery* (1964) gives it (i.e. grade point average + interest + vocabulary + language analysis) so immediately and clearly overlaps into the domains of the other variables to be considered in this section as to become redundant as a discrete heading.
But a study of the recent literature suggests that the issue of aptitude does warrant separate exploration especially if we are concerned with individual differences. Gardner and Lambert put a key question tellingly:

"How is it that some people can learn a second or foreign language so easily and so well while others, given what seem to be the same opportunities to learn, find it almost impossible? .... Perhaps the knack for languages lies in a profile of abilities or aptitudes that develop differently from person to person, some profiles favouring the language-learning process more than others. This idea makes good sense, but there is likely to be more to it than aptitudes" (1972 p.131).

Now the profile notion is in line with my main thesis and plays a crucial part in my empirical design (see Chapter Five). There is certainly more to it than aptitude but aptitude, as a fairly independent concept or, more likely, a constellation of factors, is at issue. The fact that my participants are almost by definition not where they are (as specialists in science, technology or social studies on a pre-sessional remedial English programme) because they have demonstrated a particular aptitude for English is intriguing in this respect. I have the chance to find out about abilities that are or are not present and to what degree, in relation, of course, to other aspects of communicative competence. The Good Language Learner (Naiman et al. 1978) the report on a study which, like mine, attempts to explore and relate a number of different cognitive/affective and social variables but, unlike mine, selects subjects because they are apparently good at language learning, regrets that it did not include a language aptitude test instrument to elicit data that could be correlated with measures on other variables. But this is to assume the existence of language aptitude in the narrow traditional sense, an assumption now increasingly questioned.
Chastain (1971) cites Spolsky (1966) and Osgood and Sebeok (1965) in an early attempt to broaden the aptitude concept. The latter (p.95) identify three levels, the representational, the integrational and the skill level, the latter two areas of aptitude representing the move towards a more cognitive interpretation. Neufeld (1979) finds the aptitude concept ambiguous as well as (as we have already seen) under-researched. The Carroll and Pimsleur argument for the existence of L2 aptitude assumes that L2 learning differs in kind from L1 because everyone masters a first language whereas much L2 learning is unsuccessful. Neufeld is unconvinced, feeling that the disparity more probably resides 'in social and psychological factors which function independently of actual language learning ability' (p.231). His hypothesis that aptitude 'does not vary significantly from individual to individual or culture to culture' is tested (and corroborated) in studies showing that adult L2 learners can achieve phonological ambilingualism, a finding that again seems rather narrow in its implications.

It is interesting to note how John Carroll (1979) puts his updated view of the relevance of the aptitude concept:

"To the extent that sociolinguistic studies involve observation of speakers interacting in communication situations, it would be desirable to consider the possible relevance of various kinds of language abilities. For example, relative dominance of persons in dyadic interchange may interact with verbal fluency factors. Persuasive speech styles are possibly a function of organising verbal expression in retrieving appropriate verbal memories" (p.23).

Aptitude has entered the communicative age. I shall be looking at participants in a variety of 'communication situations' and studying their performance in terms of various kinds of abilities. Perhaps something not explicable in terms of any of the variables or evaluation criteria used may still emerge and be attributable to 'aptitude'. 
3.3 **Attitude and Motivation**

Theoretical question marks left by the treatment of the complicated concepts of attitude and motivation in applied linguistics make it necessary to go back into the parent paradigms for clarification if variables under these headings are to be investigated in my study. The following examples illustrate some of the potential confusion.

"It has been argued notably by Gardner and Lambert and their associates, that the successful acquisition of a foreign language depends, in the main, on specific motivational factors.... According to this view, foreign-language learning is less likely to meet with success if the student's underlying motivation is 'instrumental' rather than 'integrative'" (Burstall 1978 p.2).

In fact, though, if you look at the various elicitation devices used by these researchers (eg Gardner and Lambert 1972, Gardner, Smythe and Clément 1979), the term 'motivation' appears only once in their 43 (1972) or 24 (1979) data collection instruments (in the 'motivational intensity scale' which asks about 'the amount of effort students felt they expended in learning French' (1979 pp.309/10)). More frequent categories are 'orientation' and 'attitude'. Others are 'desire', 'preference' and 'inquisitiveness'.

Savignon (1976) is uncompromising when she states:

"**Attitude is the single most important factor in second-language learning**" (p.295 my emphasis).

She is also less sweeping than Burstall in her interpretation of Gardner, Lambert and colleagues, noting that they 'have woven a rich texture of the motivational and attitudinal strands in learning to speak, and hence to be like another' (p.295 original emphasis).

Naiman et al. (1978) suggest that attitudes to the language learning situation may be more important than integrative/instrumental
motivation and have attitude rather than motivation as a theme in their student interview questionnaire, though the terms appear separately and with equal status in their initial 'model' of the second language learner and language learning (p.2).

In fact, it is in the various inventories, classifications and 'models' that the conceptual and categorising confusion emerges most vividly. Yorio (1976) has integrative and instrumental motivation as one major category under 'affective domain', separate from 'attitudes towards L2 culture/people' a sub-category of his 'socio-cultural factors'. Schumann (1978) has attitude as a 'social factor' on a par, for example, with 'size' (of L2 learning group) whereas his 'motivation' is an 'affective factor' alongside 'culture shock'. By the time he has been adapted by Edwards (1980 p.482) 'attitude' has disappeared but 'motivation' remains, teamed with intelligence and language aptitude as the sources of individual difference in a model of second language acquisition. Altman's 1980 'dimensions of individual difference in L2 learning' (pp.5/6) contains 'attitudes and motivation' as one dimension with subdivisions:

"(a) motivational orientation
   (i) integrative
   (ii) instrumental
   (iii)'linguistic hobby'
(b) intensity of motivation
(c) source of motivation

(d) attitudes toward:
   (i) the target language
   (ii) the target culture or people
   (iii)language learning in general
   (iv) the target language teacher
   (v) the environment for learning" (pp.5/6).
It is not at all clear where his (a) (i) ends or where his (d) begins. Perhaps it would be helpful to look to the parent disciplines for help in the clarification of the constructs where applied linguistics seems uncertain and inconsistent.

The task is not going to be easy. The initial mood of social-psychological investigations of the problems is nearly always pessimistic:

"Chapter headings in psychology, emotion, motivation, perception, learning, and the like, designate neither distinctly separate sets of behavioural phenomena nor unique psychological processes .... Chapter headings change as our ideas about the nature of these processes change." (Bindra and Stewart 1966 p.9 original emphasis).

"For several decades now there has been in the literature on attitudes a continuous undercurrent of controversy over the theoretical and the operational definition of the term" (Jahoda and Warren 1966 p.7).

But if I am to collect meaningful data on these variables I need a model offering definitions that take reasonably systematic account of inter-relationships between the variables. Fishbein and Ajzen's (1975 p.29) is intuitively attractive as a starting point in that it sorts out three overlapping concepts: belief, value and attitude. If beliefs are one's mainly cognitive inferences about the world, one's conception of it in terms of degrees of probability that things are so, and values are one's mainly affective feelings about what one would like the world to be like, attitudes, say Fishbein and Ajzen, are a cognitive/affective blend of beliefs and values. Formulaically: \( A_0 = b_1 e_1 \).

Thus someone's attitude to highly specialised education in Britain could be measured in terms of the strength of his belief that such education existed (eg a figure of 0.8 certainty) \( x \) his estimate of its efficiency (his value) (eg at +2 on a scale from -3 to +3) to produce a positive attitude (at 1.6 out of a maximum of 3). I shall not actually use
the formula but it is helpful in sorting out the constructs. It also suggests attitudes as enduring though not unalterable by experience in new situations, for example the kinds of changes I shall want to trace and measure (see Chapter Eight). The functional approach to attitudes (eg Smith, Bruner and White 1956, Katz 1960 and Lane 1969) is useful, too. Their categories, for example, of social adjustment, personal integration, value expressive attitudes are relevant. Abelson (1972) considers the concept of attitudes acceptable as describing mediators of personal feeling but not initiators of behaviour. Perhaps this view is a clue as to where the concept of motivation might belong in a potential model.

As usual in research, we face the problem of how far back to go into the parent paradigm. It does not seem particularly relevant, given my focus, to re-investigate the two initial psychological approaches to motivation, the psycho-analytic (Freud 1917-1949) or the behavioural/associationist (Hull 1943) which, though very different in many respects share the assumption that 'all behaviour is basically carried out in an effort to reduce internal tension or stimulation and rests on a limited set of supposedly primary drives such as those for food, water and sex' (Deci and Ryan 1980 p.40). More appropriate are the approaches of later theorists from areas such as cognitive development (eg Piaget 1952), cognitive psychology (eg Hunt 1965), social motivation (eg McClelland, Atkinson, Clark and Lowell 1953), humanist psychology (eg Maslow 1954), n Ach (or need for achievement) and expectancy theory (Atkinson 1957 and 1964) or the psychology of personal constructs (eg Kelly 1955). All of these concentrate on something 'independent of the primary tissue-based drives' (Deci and Ryan P.40) namely intrinsic motivation, defined
by the latter two social psychologists as follows:

"Intrinsically motivated behaviours are those behaviours that are motivated by the underlying need for competence and self-determination" (p.42).

I find a particularly interesting connection between this view of motivation and Kelly's 'fundamental postulate' (1955 p.46). This puts the emphasis on anticipation, with the individual testing various hypotheses about his world so that he can get better at anticipating events. It is this process of 'constructive alternativism', that explains, Kelly thinks, his 'pushes and pulls', his motivation. He is motivated to reach a point where he can anticipate things accurately; his attitudes at any particular time are affected by his felt success at anticipating.

Now if the full implications of these various views of motivation are considered it becomes clear that motivation affects the attitude = belief x value equation all along the line. It can operate as a trigger, as a 'secondary drive' (see Hunt 1965) such as the need for money or an intrinsic one such as the need for competence or self-determination, getting someone thinking about their beliefs and values, their attitudes towards, say, a particular training programme in Britain. As these attitudes function as mediators of behaviour during the training period, they will, in interaction with intervening events, influence the intensities (plural because of the mutability of an individual's 'profile of variables' (Strevens 1980)) of motivation which now dictates the degree of persistence or perseverance applied. And such a model also shows motivation interpreted in terms of goals as it seems to be in discussions on reinforcement (eg Skinner 1954).
Figure 2.2 (on the following page) attempts to illustrate the main concepts explored here in a context relevant to my study.

Whatever conceptual stance is taken on attitude and motivation, the problem of their measurement is recognised throughout the socio-psychological literature. The researcher is inevitably confronted with the problem of measuring underlying traits which may not be translated or translatable into observable behaviour. Whether self-reports (free or through scales), observation, 'objective' tasks or physiological reactions are used as the techniques for data collection, validity and reliability seem to be more serious problems than in other areas of measurement. As usual, the researcher is enjoined to combine a variety of data elicitation devices (see for example Cook and Selltiz 1966 pp.325-353).

We can now return to the applied linguistics paradigm with a little more confidence about the concepts. The implications of the discussion will influence my study in the following ways:

1. The insights and experience of Gardner, Lambert and associates are considered a useful starting point. It is now easier to distinguish between their 'orientation' and 'behavioural intensity', Motivation 1 and Motivation 2 in my diagram. The measurement of the latter in my study will be informed by other similarly directed instruments, for example the Aberdeen Academic Motivation Inventory (Hartley, Holt and Hogarth 1971).

2. My longitudinal design offers the chance to trace changes in attitudes and motivation. This is important given their essential mutability and the need to compare initial 'orientating' motivations
Figure 2.2: Motivation, Attitude and Related Concepts
with the eventual reality of Motivation 3. Obviously I would agree with McDonough (1981), at the same time accepting his warning on the complexities of causality:

".... in discussing an association between attitudes or orientations and learning, and achievement, it is only possible to make causal inferences, that is, state which affected which, in a longitudinal study" (p.153).

3. The question of expectations, a complex combination of attitude and motivation, will receive special attention. The level and nature of what different individuals expect from their UK training period must influence the satisfaction they feel during and after it. The expectation: satisfaction relationship will be regularly evaluated in my study. James (1980) makes some interesting generalisations in this connection. His neat summary is reproduced here as Figure 2.3. It will be revealing to see to what extent my individual participants, studied in depth, correspond to his 'types'.
TYPES OF CULTURAL INTERACTION IN AN ACADEMIC SETTING

Figure 2.3: from James, K. 1980.
4. Oller and Perkins' theoretical and methodological attacks on attitude and motivation research in L2 learning studies (e.g., Oller 1979; 1, Oller and Perkins 1980) are taken seriously:

"The trouble with attitudes is that they are so out of reach, and at the same time they are subject to a kind of fluidity that allows them to change (or perhaps to be created on the spot) in response to social situations. Typically it is the effects of attitudes that we are interested in rather than the attitudes per se, or it is the social situations that give rise to both the attitudes and their effects that are the objects of interest" (Oller 1979; 1 p.112).

True. But in my study the fluidity and change are seen as potentially revealing, as is their relationship with the social situation; the 'trouble' is considered worth taking. Oller's colourful description of the methodological problems in collecting valid data, especially those caused by response set, socially acceptable answer-spotting, self-ignorance and deception, is salutary but not pre-emptive, particularly for a study where the participants are quite well-known to the case-study researcher. Besides, it is not logical to dismiss attitudinal/motivational factors as important variables in L2 learning because 'the largest amount of variance in language proficiency scores that was predicted (in a review of 33 surveys) on the average by attitude measures was never greater than 8½%' (Oller 1979; 1 p.125) when you have just condemned most such measures as invalid and unreliable.

As the design in Chapter Five Figure 5.1 shows, attitude and motivation receive a fair amount of attention in my study for what they can contribute to the detailed profiles of participants. Findings on them should also be useful in the necessary reinvestigation of the concepts themselves.
3.4 **Personality**

Under this heading the pattern of preliminary re-assessment will be similar to that in the previous section, given the kind of overlap between the notions of attitude and personality encapsulated in the following definitions:

"an attitude can be defined as an enduring organization of motivational, emotional, perceptual and cognitive processes with respect to some aspect of the individual's world" (Krech and Crutchfield 1948 p.152).

Personality is "the dynamic organization within the individual of those psychological systems that determine his characteristic behaviour and thought" (Allport 1963).

Again, this overlap is reflected in inconsistencies in various inventories and 'models' of language learning variables. Naiman et al. (1978) give personality equal status with their two other separate learner variables, motivation and attitude (p.2). Yorio's lengthy classification does not use the term personality at all but has, in his 'affective domain' the category 'egocentric factors' including 'depression, anxiety, homesickness, ego permeability, rejection, self-consciousness etc'. (1976 p.61 my emphasis). For him, it may be remembered, motivation and attitude also belong in the affective domain, but at different, unconnected levels. Schumann (1978) includes four 'personality factors' (tolerance for ambiguity, sensitivity to rejection, introversion/extroversion and self-esteem) distinguished, though not transparently, from his 'personal factors' (nesting patterns, transition anxiety, reaction to teaching methods, choice of learning strategies, the latter not, it seems, belonging among his 'cognitive factors'). Altman 1980 has 'personality factors such as: introvert-extrovert, capacity for empathy with strangers, goal-oriented vs role-oriented, competitive vs withdrawing'; again it is surprising that
the latter two are not included in his quite comprehensive attitude and motivation category (p.5/6). 'Personality' appears as an input to the 'socio-affective filter' and 'cognitive organiser' in Dulay and Burt's 'working model for some aspects of creative constructs in second language acquisition' (1978 p.70), a development of Krashen's monitor model. In Edwards' adaptation of Gardner et al. and Schumann (1980), there is no explicit place for personality as a source of individual difference in second language learning, intelligence, language aptitude and various kinds of motivation apparently filling the gap.

So back into the parent paradigm for clarification. Practical application of psycho-analytic theory is not, of course, a possibility for the non-psychologist researcher. An alternative approach is the use of factor-analysis-developed personality questionnaires. Participants select limited-choice responses to questionnaire items and these responses are aggregated to produce scores on each of the identified personality dimensions. The two main contenders here are Cattell (eg 1970) and Eysenck (eg 1964) both of whom, in spite of significant differences that the latter at least would claim exist between their models, derive their classifications of personality factors from similar correlational techniques. As Bynner and Whitehead (1972) point out, with an extended combination of Cattell's narrower factors 'one would finally end up with two or three factors which are virtually the same as Eysenck's' (p.21). However:

"The real difference between Cattell's and Eysenck's models of personality structure is not one of empirical findings, but is more a question of the theoretical properties of the factors 'identified'. Cattell favours the linking of his factors to concepts put forward by other personality theorists such as the psycho-analysts; Eysenck links his factors back to neurophysiological concepts and learning theory" (p.21).
Bynner and Whitehead also distinguish Eysenck's 'main interest level' from Cattell's, seeing the two general factors identified by the former (extroversion/introversion and neuroticism/stability) as operating at the level of personality type whereas the latter's interest is in personality traits. Thus, Cattell's sixteen 'first-order' factors are claimed to be 'dimensions of personality, functionally unitary aspects of personality which show themselves in behaviour' (Cattell 1969 p.20). The factors, using his 'popular labels' and presented in Independent Assessment and Research Centre computer-scored format, are:

### 16PF
Personality Profile

<table>
<thead>
<tr>
<th>Score</th>
<th>Trait</th>
<th>Low Meaning</th>
<th>High meaning</th>
<th>Percentile</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw Stem</td>
<td></td>
<td></td>
<td>1 2 3 4 5 6 7 8 9 10</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>cool, reserved</td>
<td>warm, easygoing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>dull</td>
<td>bright</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>easily upset</td>
<td>calm, stable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>not assertive</td>
<td>dominant</td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>sober, serious</td>
<td>happy-go-lucky</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>expedient</td>
<td>conscientious</td>
<td></td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>shy, timid</td>
<td>venturesome</td>
<td></td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>tough-minded</td>
<td>tender-minded</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>trusting</td>
<td>suspicious</td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>practical</td>
<td>imaginative</td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>forthright</td>
<td>shrewd</td>
<td></td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>self-assured</td>
<td>apprehensive</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q₁</td>
<td>conservative</td>
<td>experimenting</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q₂</td>
<td>group-oriented</td>
<td>self-sufficient</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q₃</td>
<td>undisciplined</td>
<td>self-disciplined</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Q₄</td>
<td>relaxed</td>
<td>tense, driven</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
I chose to use Cattell's 16PF instrument in my empirical study because of its more explicitly broader coverage of personality traits, (cf Adcock 1970), because I was familiar with its use on an experimental basis in the British Council's recruitment procedures and because of some apparently interesting findings when I administered it to the 1979 pilot group (see Chapter Seven, Section 2 below).

But a survey of the psychological literature on Cattell offered warning signals. At the theoretical level, there are always doubts about whether traits arrived at by the factor analysis of observable responses can in fact claim to represent underlying psychological realities (see Lubin 1970, Wittenhorn 1970, Bouchard 1975). And Allport (1963) feels that such instruments may measure elements of behaviour but produce 'uncemented mosaics of elements' ignoring the varied inter-relationships of traits in individuals. This latter criticism is not altogether fair since the profile of measures on Cattell's 16PF can be translated into a discursive 'portrait' by trained experts and this certainly allows for interaction between traits. (Part of my research into this point was to have my own 16PF profile so interpreted.)

At the empirical level, some of the criticisms of the Cattell instrument are more worrying. Harsh (1970) is concerned at the lack of information on the derivation of the items in Cattell's questionnaire, finds individual items of an 'objectionable ambiguity', not 'subtle or penetrating' (p.560) yet sees the 16PF's potential as a research instrument. Wittenhorn (1970) is not convinced of the independence of the sixteen factors and, like Harsh and Lubin, is surprised at the test's apparently low reliability. Lorr (1970) has the most disturbing findings on this front, claiming higher inter-item correlations among
items not representing the same trait than among those that are, a fault, however, not supported by Adcock's checks (1970). Bouchard (1975) worries about the lack of consistency between different forms of the 16PF (Cattell offers five versions, two pairs of parallel tests and one version for less literate candidates). And although the 16PF Handbook (1970) gives norm profiles for a variety of groups (by profession and culture, for example), Rorer (1975) complains that the figures are given without sufficient data on sampling techniques or validation procedures. Like other critics, however, including even Lorr, Rorer considers the Cattell instrument the best personality inventory available. Perhaps Bynner and Whitehead (1972) sound the most appropriate warning note for my purposes:

"Although Eysenck's two dimensions, extraversion and neuroticism, give only a very generalised and perhaps over-simplified description of personality, they can be measured very reliably (Eysenck 1965) .... With Cattell's more specific dimensions, however, there are less grounds for optimism. Although Cattell's factors provide a more detailed description of personality, they cannot be measured as reliably as Eysenck's factors .... It is a case of what you gain on the swings you lose on the roundabouts" (p.34).

Given my own case study approach, my need for detailed learner profiles and a design that builds in alternative means of personality data collection, I choose to ride the swings. As the empirical study report in Chapters Six to Eight shows, the question marks over the Cattell 16PF are taken seriously and investigated fairly thoroughly.

A useful direction-indicating preliminary to the study itself is to summarise some of the more interesting findings on personality as a factor in general educational research. The main focus here has been on personality and academic achievement. Furneaux (1957, 1962) found that students who scored higher on Eysenck's neurotic and introvert
scales were academically more successful at tertiary level, Lynn and Gordon (1961) and Kline (1966) that high introversion, on its own, correlated positively with university achievement. Entwistle and Wilson (1970) claim that introverted students have better study methods but discover no correlation between neuroticism and degree class. There is also evidence in the research literature of connections between personality and choice of subject area. Eysenck and Cookson (1969) claim introversion: science and extraversion: languages relationships. Savile and Blinkhorn (1976) using the Cattell 16PF in a major study of British and American students, find that arts students are generally more outgoing, socially bold, tender-minded and experimenting while science students are higher on stability and self-control.

We should note, however, that a common conclusion in most of these studies is that, while personality factors do contribute to the prediction of academic achievement, this contribution has to be explored in interaction with other variables. Kline (1976) goes a little further but still catches the mood of most personality/educational research:

".... we should turn to the study of the learning situation itself, the interaction between teacher and pupil, and use our knowledge of personality measurement to further our studies in that direction" (p.60).

Research such as mine is in that direction, though with insights from personality measurement used in wider social contexts than teacher-pupil interaction.

I can now return to the field of applied linguistics with a usable though realistically evaluated personality measure and evidence that personality dimensions are a relevant focus of interest. If they are
to be used in profiling English language learner/users in a C2 context, not too much must be claimed for them without support from other kinds of personality data or without putting them into their full cognitive/affective and social context.

Most of the studies associating TL learning and personality seek narrower, more specific a priori connections than I shall. Naiman et al. (1978) pre-select four different personality scales rather than one broad instrument such as Cattell's. Their choice is Budner's intolerance of ambiguity scale (1962), Mehrabian's sensitivity to rejection measure (1971), Hogan's empathy scale and Eysenck's introversion/extraversion (1963). In general they find no systematic relationships between these and their criterion measures, the IEA French tests (picture recognition + sentence recognition + listening comprehension + conversation, all these with multiple choice answers, and an elicited sentence imitation test) though a low tolerance of ambiguity seemed a bad sign for early L2 learners. It is interesting that Naiman et al. felt that their semi-structured student interviews were more revealing of personality than the tests they used. Perhaps this is not surprising given the limited range of factors tested. Perhaps also they should have designed into their study more explicitly the use of subjective personality data in support of the objective.

Tucker, Hamayan and Genesee (1976) look at small groups (N=17, N=28) of early and late immersion course students using Cattell's High School Personality Questionnaire (HSPQ), a younger people's variant of the 16PF, to identify predictor variables of French language learning alongside IQ, aptitude, cognitive style, attitude/motivation and language
background. Their whole constellation of factors is found to be predictive of second language proficiency and of the Cattell factors, the shy-venturesome continuum seems most significant, apparently suggesting better oral proficiency and a greater likelihood of C2 contacts. This probably ties in with H Brown's 1980 emphasis of empathy and lack of inhibition as helpful contributors to L2 success. It may also relate to Cummins' 1980 concept of 'basic interpersonal communication skills' (BICS), which he sees as less automatically transferrable by an individual from L1 to L2 than the complementary notion of 'cognitive/academic language proficiency' (CALP). Chastain's 1975 investigation of anxiety and extraversion in TL learning again found the reserved/outgoing dimension significant, an empirical finding he acknowledges most language teachers would intuitively confirm.

An interestingly different approach to the personality:language learning issue is represented by Guiora and Acton of the Personality and Language Behavior Research Group at the University of Michigan. This group started from an interest in clinical psychology then moved towards a 'systematic study of the inter-relationship between personality parameters and language behaviour' (Guiora and Acton 1979 pp.193/194). This pedigree probably explains why their useful constructs of language ego, language ego boundaries, the permeability of language ego boundaries and empathy are measured in such objective experimental ways, through their micromomentary expression test, for example, or through the controlled use of hypnosis, alcohol and drugs. I am in no position to employ such techniques, of course, but the constructs are important and probably accessible through less rigorously experimental means of
elicitation. Certainly, with my focus on individuals learning and living with TL in C2, any characteristics that may influence their receptiveness to both must be crucial.

Willis, Doble et al. (1978) make use of the Cattell 16PF to compare, among other things, 'high improvers' with 'low improvers' on a speaking test designed to measure 'ability to deal with sustained discourse' (P.89). Their results are reproduced as Figure 2.4 here:

![Figure 2.4: A comparison of the personality profiles of high and low improvers on a speaking test and re-test. (Willis, Doble et al. 1978 p.78).](image)

Given the doubts about the reliability of Cattell's test and the fact that most of the profile scores here fall in the 'normal' 5 to 6 sten
score range (which means, for instance, that neither high nor low improvers are particularly reserved or outgoing) the findings, though interesting, are not as clearcut as they first appear. I shall not be able to use Cattell so straightforwardly in my study. Increased reliability will be sought through repeated administration of more than one form of the test; attempts to surmount the TL problem mentioned by Oller will be made by the use of translation and checks on the validity of the 16PF for my purposes will be carried out using data collected in other ways. In essence I shall be using insights from Kelly's psychology of personal constructs (1955) (see Chapter One, Section 4) to close factor-analytic loopholes. Participants will be invited to speak for themselves about personal constructs in corroboration or refutation of what an objective personality test may have said about them.

3.5 Cognitive Style

Any educational researcher, teacher or learner will admit to the desirability of a valid, objective, consistent criterion to help understand and predict behaviour. Doubtless this feeling added impetus to the development of intelligence tests, but these were soon dogged by definitional problems, the unacceptable face of the nature/nurture debate, and a credence that was allowed to affect people's futures too much, too soon. More acceptable is a means of evaluating 'information-processing habits .... characteristic modes of operation which, although not necessarily completely independent of content, tend to function across a variety of areas', to quote Messick's general description of the term cognitive style (1970) cited in McDonough (1981 p.130). This is not based, as IQ is perceived to be, on continua where one end represents 'good' and the other 'poor' but allows measurement of
individual differences that in the main cut across levels of 'intelligence', even if not always across 'socially or academically valued' characteristics (McDonough 1981 p.133). The location of this section between those on personality and learning is logical even if, as ever, the category 'cognitive style' is not totally discrete. It is conventional in the educational research literature for cognitive style to be considered under headings such as 'personality and learning' (cf Open University Course E201), a lead frequently followed in applied linguistics, where H D Brown's point is fair:

"It is difficult to argue that cognitive style is a strictly affective factor; it is more a combination of affect and cognition" (Brown 1973 p.238).

In fact most of the means developed to investigate cognitive style seem to have started at the level of perception or cognition then been broadened inductively into the affective domain.

Floyd (1976) makes a useful distinction between two general methods of attempting to get at underlying cognitive processes through performance. The first sets tasks during which the overt behaviour of individuals can be examined and measured; the second focuses on the outcome of performance on tasks. Both then make inferences about covert styles. Important cognitive style types identified through the first method, that is by 'externalizing the thinking process' (Floyd 1976 p.10) are Bruner's focusers and scanners (1956) and Pask and Scott's serialists and holists (1972).

Bruner sets a task requiring the classification of categories. A focuser (or wholist (sic, and not Pask and Scott's holist)) prefers a step-by-step approach to problem-solving, concentrating on one
definitely relevant attribute at a time and wanting to test his hypothesis about it before progressing to the next logical step. A scanner (or partist) is more prepared to take risks, juggling several ideas at once (some of which may have to be discarded) without expecting each to crystallize quickly or neatly as a certain and direct logical advance. Pask and Scott also use classification tasks to examine differing cognitive styles or, as they add, 'mental character' or 'cognitive competence' (Pask and Scott 1972 p.257).

"Serialists learn, remember and recapitulate a body of information in terms of string-like cognitive structures where items are related by simple data links: formally, by 'low order relations'. Since serialists habitually assimilate lengthy sequences of data, they are intolerant of irrelevant information unless, as individuals, they are equipped with an unusually large memory capacity. Holists, on the other hand, learn, remember and recapitulate as a whole: formally, in terms of 'high order relations'" (Pask and Scott 1972 p.258).

Again, then, there is a general distinction between the logical, step-by-step build up (by the serialists) and a broader-ranging approach looking for new angles on things and a wider variety of ways of learning (from the holists). An interesting source of insight used by Pask and Scott is the different ways in which learners handle 'teachback', an essential part of 'conversation theory' involving the learners' own accounts of how they actually performed the tasks. Serialists tend to teach back with events sequenced as they were learnt whereas holists are more likely to give personally restructured accounts.

The less direct route, seeking to get at the processes constituting a cognitive style through the products of people's thinking is, however, the one more frequently taken. Kagan et al. (1966) use the Matching Familiar Figures Test (MFFT) to discriminate between impulsive and reflective thinkers. The impulsive child, for example, makes
decisions quickly and seems 'minimally concerned about making mistakes' (Kagan, Pearson and Welch 1966 p.315). The reflective child delays response, 'considers the differential validity of alternative answers' and 'inhibits potentially incorrect hypotheses'. The scope of the basic cognitive distinction has been broadened to suggest that while reflectivity is an advantage in tasks requiring inductive reasoning, a reflective child might be 'left dithering, unable to make up his mind' (Floyd 1976 p.42) in decision-demanding situations or find the creative thinking and mastery of principle required in non-science subjects inhibited by an over-reflective orientation (Kagan 1965).

Pettigrew (1958) posits the cognitive style of category breadth. Broad categorisers (as assessed on a test asking for responses to questions on the probable maximum/minimum sizes of things) accept a wide scope for concepts, taking the risk of including in a hypothesis more than it really encompasses. Narrow categorisers go by tightly confined rules that may miss potential generalisability.

The two product-evaluated cognitive styles discussed so far involve the elicitation of data through problems that have only one correct or significant answer. In that sense they are dealing with convergent thinking, that is the intake of information and the selection of a pre-processed response. But the intellectual operation of convergent thinking presupposes a divergent-thinking alternative (Guilford 1959), the kind demanded by problems where there is no 'correct' answer, where we have to create our own responses. Liam Hudson (eg 1966, 1968) has investigated the differences in people's facility in handling the two kinds of problem, identifying as a result convergence and divergence as cognitive styles. Convergent thinkers are the probabilistic 30%
of a population who perform better in closed-response tests, divergent thinkers the 30% with better performance on open-ended tasks. Hudson and others have taken the investigation further, with the predictable result that they are soon talking about more than the interaction between problem and problem-solving preference. In summary, the extended scope of convergent and divergent styles becomes the following:

<table>
<thead>
<tr>
<th>Convergent</th>
<th>Divergent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Better at science subjects.</td>
<td>Better at arts subjects.</td>
</tr>
<tr>
<td>More popular with teachers.</td>
<td>Less popular with teachers.</td>
</tr>
<tr>
<td>More likely to succeed at school.</td>
<td>More likely to succeed at higher</td>
</tr>
<tr>
<td></td>
<td>tertiary level.</td>
</tr>
<tr>
<td>[+ authoritarian]</td>
<td>[+ liberal]</td>
</tr>
<tr>
<td>[+ masculine]</td>
<td>[+ feminine]</td>
</tr>
<tr>
<td>[+ unemotional]</td>
<td>[+ emotional]</td>
</tr>
<tr>
<td>[+ sober]</td>
<td>[+ humorous]</td>
</tr>
</tbody>
</table>

These cognitive styles are still fairly exploratory, not completely 'secure' (see Hasan and Butcher 1966) and often based on small-sample research (for example Hudson's 30 divergers followed through three years at Cambridge (Hudson 1968)). But the constructs are intuitively attractive and likely to be relevant in a research design eliciting data through a variety of means, both closed- and open-ended.

The cognitive style the investigation of which is actually designed into my study is field-dependency as measured by the Group Embedded Figures Test (GEFT) (Witkin, Oltman, Raskin and Karp 1971). This 'characteristic mode of functioning that we show throughout our perceptual and intellectual activities in a highly consistent and pervasive way' (Witkin 1973 P.2) is about "the extent to which a person is able to deal with part of a field separately from the field as a whole .... how analytic he is" (p.5). Field dependency was originally identified
as exclusively perceptual (evaluated according to people's ability to line up vertically a chair in a moving room or a rod with a moving frame). After considerable research and generally convincing extrapolation, however, the implications of field dependency have been discovered to extend into personality and social domains. The current scope of the concept may be summarised as follows:

<table>
<thead>
<tr>
<th>Field-dependent</th>
<th>Field-independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple form not easily identified in the embedding complex figure (in the CEFT). Mean scores higher in women.</td>
<td>Easy identification of embedded simple form in the complex figure (in the GEFT). Mean scores higher in men.</td>
</tr>
<tr>
<td>Diffuse, global; overall organisation of the prevailing field dominates; parts of field experienced as fused with background.</td>
<td>Experience is delineated, articulated; parts of the field experienced as discrete, the field as a whole, as structured.</td>
</tr>
<tr>
<td>Less analytic; benefiting from discovery method.</td>
<td>Analytic; better at problem-solving tasks.</td>
</tr>
<tr>
<td>Dependent on others for definition of self-view.</td>
<td>Sense of separate identity; less need for the approval of others; independent.</td>
</tr>
<tr>
<td>More constant in rating other people.</td>
<td>More variability in rating people.</td>
</tr>
<tr>
<td>Fluctuations in specialist preferences.</td>
<td>Stability in specialist preference.</td>
</tr>
<tr>
<td>Attentive to human content of the environment; socially sensitive.</td>
<td>Less socially aware.</td>
</tr>
<tr>
<td>Preference for people-oriented jobs eg humanities, social sciences, teaching, selling; more conventional jobs.</td>
<td>Preference for object-oriented jobs eg sciences, maths, engineering and technology; less conventional jobs.</td>
</tr>
<tr>
<td>[+ conforming] [- self-directed] [- self-esteem]</td>
<td>[- conforming] [+ self-directed] [+ self-esteem]</td>
</tr>
</tbody>
</table>

There is obviously considerable scope here for the exploration of inter-relationships between the more cognitive, the more affective and the
more social dimensions of this style. And the range and tone of previous research with field-dependency is reassuring with some of the recent studies having clear implications for TL inquiry. Satterly and Telfer (1979) investigating field dependency and advance organisers in learning and retention, find a significant positive connection between field independence and the successful transfer of learnt knowledge, with or without the use of an initial general introduction to key concepts by the teacher. Annis (1979) presents evidence that field-independent university students are better at analysing organised discourse and at imposing a structure when the text is disorganised. It was also found, in the same study, that field-dependent learners were less efficient at distinguishing between sentences that were 'structurally important' in a text and those that were not. Some of these findings are re-investigated in my empirical study.

There are, in fact, quite a few examples of applied linguistic investigations making use of the field dependency cognitive style. In The Good Language Learner (Naiman et al. 1978) field dependency is hypothesised as relevant to the language learner's need to reject irrelevant 'clues'. In the main, however, the study did not find strong relationships between cognitive style and its criterion measures (see 3.4 above) though there were indications that field-independent learners wanted greater use of L2 in the classroom and were more likely to have an instrumental than an integrative orientation towards the language. There is also the suggestion that field dependency becomes a more significant predictor variable at the later stages of L2 learning, a relevant point, perhaps, given the age and stage of my participants, but, as McDonough (1981) points out, again perhaps a causality question
mark as older students 'might have been encouraged to adopt an analytic
turn of mind as part of their preparation for tertiary education' (p.132).

Genesee (1978) uses the GEFT as one of his measures of 'various predictor
variables' (p.494). He finds it associated with high achievement on the
Test de Rendement, his criterion measure of French phonology,
vocabulary, grammar and reading comprehension.

Genesee also makes some interesting methodological points in his
research report. Particularly relevant to my study is his emphasis on
the value of the case study method focusing 'directly on the individual
second-language learner' and, thereby, perhaps more than any other
method 'attempting' to look at 'processes during second-language
learning', and 'ethnography', 'the study of language usage patterns - who
says what to whom in what language' (p.492). As is suggested elsewhere
in this thesis, this kind of methodological mix may reveal important
variable relations that do not emerge from quantitative, high-N, cross-
sectional studies. In Chapter Seven, for example I report on experi-
mental work where small groups, selected with field dependency as the
key independent variable, work at study-related tasks in TL with their
performance monitored quantitatively (in terms of outcome and transfer)
and 'ethnographically' (in terms of who does and says what during the
performance).

My choice of field-dependency as my main cognitive style focus is based
firstly on its developmental pedigree:

".... indeed one of the most attractive features of this test is
its firm anchoring in a systematic context of theory and empirical
evidence" (Gough 1970 p.1015).

Then there is the appeal of its broad scope, proven relevance in the
kind of learning and language using activities I am interested in, the objective, language-independent nature of the GEFT itself and (a significant virtue when research is carried out in the context of real-life teaching programmes) its administrative convenience. The focusing/scanning dimension would have to be assessed through a complicated performance task which aims at revealing processes similar to those operating in more authentic guise in some of the language performance tests described in Chapter Six. The Pask and Scott model seems too complex for use on a pre-sessional language programme, requiring a mini-course in itself to elicit its data, and anyway remains at a stage of development where its use is restricted to science students at College of Technology level. Category width almost certainly overlaps field dependency but cannot match it in terms of previous research interest. Hudson's convergent/divergent insights certainly are of interest but, as characterised above, hardly seem to need his tests to reveal them. In any case, tests such as his 'meanings of words' and 'uses of objects' (Hudson 1966) are particularly language bound.

But indications of revealing overlap between the cognitive/affective and social dimensions of field dependency and other cognitive styles will be considered when they seem to emerge from my study. I have not selected one style to the exclusion of others. Certainly, cognitive style insights seem likely to contribute valuably to the kind of detailed participant profiles that I need.

3.6 Learning Processes, Strategies, Skills, Habits

Even the title here is immediately controversial. The discussion of relevant language learning variables via the field of cognition and learning theory is especially difficult because of inherent conceptual
complexity as well as common categorial inconsistencies. Bruner 1978 sets the scene:

".... the sheer range of studies that qualify for inclusion under the general label 'thinking' is staggering not only in its diversity of content .... but in the method and presuppositions that underlie the choice of materials and tasks (where the range is so dauntingly multi-dimensional as to defy description)" (p.vii).

And Schumann's 1976 warning of the confusion that may be caused by the blurring of the cognitive processes involved in using a language with the cognitive processes involved in learning a language show how matters are made even worse with the kind of research design that I am following.

I am interested in both language learning and language use. Research Hypothesis 1 seeks relationships between various cognitive/affective and social variables and both of these, with the suggestion (see the chapter framework in Chapter One) that language learning and use are the dependent variables ie Cognitive/affective social factors -----------\rightarrow TL learning and use

Independent Variables ----------- Dependent Variables

But since language learning is learning, we begin to see the problem of overlapping independent and dependent variables foreshadowed in Section 2 above. As suggested there, the brunt of this design difficulty will be borne when the various hypotheses are tested, but even in this present hypothesis-formulation discussion, there will be some rather untidy blurring. The area of study skills, for example, is conventionally regarded as a part of the learning theory domain. It is also, logically enough, a key aspect of language learning and, particularly with the EAP students I am interested in, of language
Thus, study skills will appear in this section, where my primary focus is on learning variables and their previous investigation with reference to L2 learning and acquisition, and in Section 4 where the primary focus is a description of the features and problems of the target language in study use.

First of all, though, an attempt to sort out some of the cognition-related concepts whose diversity and presuppositions Bruner is concerned about. Figure 2.5 overleaf is a continuum suggesting that there is covariation between the degree of overtness and the degree of universality of the concepts connected with learning. The concepts plotted on the continuum are characterised briefly in terms of their focus, the theories or approaches with which they are associated, and finally their conventional units of currency.

The main theoretical and methodological problems are at the 'deep' (left-hand) end of the continuum. Applied linguists have nevertheless attempted to make inferences about internal processes. Kennedy (1973) talks of 'the various cognitive processes occurring as the individual learner works on reorganising and classifying the primary linguistic data' (p.11). Corder's 1978 interlanguage models involve the processes of accommodation and assimilation. He also makes the point that information processing models may be misleading in a language learning context because they do not seem to account for the common gap between input and intake (p.81). The construction of certain of my performance tests and experimental tasks (see Chapters Six and Seven) accept the necessity of giving learners the 'real' learning opportunities of reorganisation, classification, accommodation and assimilation. The language acquisitionists also investigate covert processes, often at
the morpheme level. Given the communicative construct of my investigation and the level and experience of my TL learner/users, there does not seem to be much applicable insight to be gained from speculation on 'universal language processing strategies' such as 'creative construction' when it is based on morpheme-level findings (eg Dulay and Burt 1974). Like Hatch (1978) or Wagner-Gough (1975) I find this exclusive emphasis on linguistic forms unpromising. Hatch posits a vertical rather than a horizontal acquisition structure with the focus on such processes as identifying, predicting and matching utterances in full discoursal context.

At the second point on the continuum (distinguished admittedly in a somewhat blurred way from the first), learning models 'have attempted to reveal the anatomy of the solving process, mainly by making the solver's covert decisions, back trackings and intermediate solutions public' (McDonough 1981).

Perhaps the steps selected by learners as they find their way through one of Pask's 'domain maps' represent actual covert, unconscious learning mechanisms. If they do we can discover something valuable about learning and try to foster it by matching task structure to learning process and preference.

Language learning research and teaching has usually been quick to take up externalised learning process models. The audio-lingual era saw a direct transfer of S—R theory into language teaching. The communicative approach can be seen as an attempt to put into practice insights from discovery heuristics (eg Bruner 1975, Goodman 1967) along with the message of the humanists' self-actualisation through authentic
| Level and Scope          | + covert | - | - | - | - | - | - | - | - | - | - | - | - | - | + overt | + unconscious | + universal | + conscious | + individual |
|-------------------------|----------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|----------|-------------|-------------|-------------|-------------|
| Method of Inquiry       | theoretical/ | observation and evaluation | covert | overt | learning | overt | study | habits |
|                         | neurophysiological | process | of product | cognitive | style | strategies | habits |
| Focus                   | psychological reality | externalisation | cognitive style | intellectual | skills | (Gagné) | psychology of study (Mace) |
| Theories                | developmental (Piaget) | S-R (Skinner) | cognitive style | Witkin | intellectual | skills | (Gagné) | study methods (Entwistle) |
|                         | cognitive (Bruner) | Conversation | cognitive | strategies | (Gagné) | study skills | (Buzan) | study habits (Brown & Holtzman) |
|                         | info-processing | theory (Pask) | cognitive | strategies | (Gagné) | study skills | (Buzan) | study habits |
|                         | (Lindsay, Norman) | Humanist | (Maslow) | |
| Example Categories      | assimilation, accommodation; problem-solving, H-testing; decoding, storing, retrieving etc. | chaining; comprehension; learning; self-actualising etc. | disembedding; monitoring; problem definition; etc. | multiplication; patterned; note-taking; skimming; etc. | work planning; delay avoidance; revising |

Figure 2.5: A continuum classifying and characterising theories of learning and studying.
learning experience. For the researcher seeking information on variables affecting language learning and use, such models encourage:

1. the design of tests, tasks and other elicitation devices where the emphasis is on what participants are doing as well as what they produce after doing it

2. designs ensuring the maximum authenticity of task and context (see my 'general evaluational requirements' in Chapter Four)

3. the establishment of methods and criteria for evaluating the evidence received which fit the process-focused construct.

At the third point on the continuum in Figure 2.4, the move is to approaches that seek insight into covert processes from the evidence of product (and again the line between this category and the previous one is not clearcut). These would include the second kind of cognitive style test (eg Kagan, Witkin) discussed in 3.5 above.

It will be noticed from Figure 2.4 that monitoring is given the same status as a cognitive style dimension such as field dependency.

Although the monitor in Dulay and Burt's working model of L2 acquisition (1977) and in Krashen's Monitor Model itself (1975, 1976) is represented as an 'internal mechanism' (Dulay and Burt 1977 p.70), Krashen's later description (1977) stresses individual variation in monitor use, and implies that monitoring is a more conscious strategy:

"Monitor users show an overt concern with "correct" language, and regard their unmonitored speech and writing as "careless"" (p.177).

Thus there are monitor over-users and under-users just as there are field-dependent and field-independent learners. And as with the cognitive style of field dependency, there are social and affective covariants:
### Individual Variation in Monitor Use

<table>
<thead>
<tr>
<th>Monitor user</th>
<th>Spoken style</th>
<th>Uses conscious rules?</th>
<th>Personality type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal</td>
<td>- Hesitant</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Overuser</td>
<td>+ Hesitant</td>
<td>Yes</td>
<td>Self-conscious</td>
</tr>
<tr>
<td>Underuser</td>
<td>- Hesitant</td>
<td>No*</td>
<td>Outgoing</td>
</tr>
</tbody>
</table>

* May pay lip service to value of rules.

(Krashen 1977 p.182)

At the same level we may include **avoidance strategies** (see Schachter 1974, Tarone, Frauenfelder and Selinker 1975, Kleinmann 1978) such as synonym use, paraphrase, circumlocution as well as Corder's **risk-taking strategies** eg borrowing, guessing, inventing (Corder 1977). Both types of strategy show individual variability 'with personality and speech situation' (Corder 1977). There is plenty of scope for the elicitation of evidence on the role and inter-relationships of these strategies in my empirical study.

A further related point of clarification is in order here in connection with the terms 'process' and 'product'. Recent discussion of syllabus design in ELT has distinguished between product (or goal) -oriented and process-oriented approaches to what learners need:

"The expression 'learner needs' is open to two interpretations. On the one hand it can refer to what the learner needs to do with the language once he has learnt it. This is a goal-oriented definition of needs and relates to terminal behaviour, the ends of learning. On the other hand, the expression can refer to what the learner needs to do in order to actually acquire the language. This is a process-oriented definition of needs and relates to transitional behaviour, the means of learning." (Widdowson 1979 p.1).

The syllabus discussion and the learning theory discussion are making different though interestingly related process/product distinctions.
the learning theory context, the key distinction is about whether you try to externalise covert operations as they are activated or in terms of their cumulative results (eg Pask vs Witkin). In syllabus defining/designing discussion, the contrast is between the targets and the transitional behaviour, covert as well as overt, en route. What both discussions have in common is a concern that too much emphasis on product may have a negative effect on the quality of learning:

".... primary attention on products of thought .... must inevitably mean educational procedures and psychological research which are inappropriate, superficial and, in all likelihood, incorrect" (Bloom and Broder 1950 p.31).

In my empirical study the primary interest is in underlying processes even when the evidence is from their product.

At points 4 and 5 on my continuum the categories are more overt, conscious and individually variable. The term 'study' tends now to replace the term 'learning', often combined with 'skills' at 4, 'habits' at 5. In both cases we are talking about elements that can be explicitly trained. The concepts are surveyed briefly here in their general learning theory context. But as we see from the descriptive matrix in Section 4 below, which exemplifies the concepts in their TL research interpretation, they belong for my purpose mainly in the description of target language in study use. Study skills seem to belong in Gagné's 'intellectual skills' domain (Gagné 1975). As examples of these Gagné offers: 'the ability to perform mixed arithmetic operations' or to 'respond appropriately to the compact and complicated sentences of text (a learner) encounters in his reading' (p.23). These are skills that a learner concurrently deploys and develops in coping with particular tasks. Schumann's 1976 learning/use 'blurr' is
encapsulated here, but then the overlap is conveniently accepted in learning theory anyway:

"We indicated that within psychology the term learning may not be used in quite the same (lay) way, and we tried to arrive at a more technical working definition ... more or less permanent changes of behaviour that are the result of experience" (Borger and Seaborne 1966 p.15. My emphasis).

The 'techniques' at point 5 on the continuum are more about general approaches and habits. Their combined cognitive/attitudinal nature locates them in Gagné's 'attitude' rather than his 'cognitive strategies' or 'intellectual skills' domain. Such techniques are exemplified by Entwistle N and D (1970) as 'study methods' like work planning, thinking ahead or recognising the importance of finding conditions suitable for efficient studying.

Inevitably there are labeling inconsistencies, however. Biggs' 1978 Study Process Questionnaire has a sub-category called study 'skills' including features such as work scheduling or regular revision that belong at point 5, as study techniques, on my continuum. So do the kind of L2L behavioural strategies identified by Stern (1974), Rubin (1975) and Fillmore (1976) eg 'willingness to practise' (Stern) 'join a group' (Fillmore).

In the investigation of my participants' target language learning development and use (both elements of the 'process factors' section of the framework for my study shown in Figure 1.1 Chapter One) insights into all five categories on the learning theory continuum will be sought.

3.7 Environments

Figure 1.1 also shows the learning, living and attitude development
factors investigated during the longitudinal study symbolically framed by 'programme' and 'community' considerations. The symbolism is appropriate. The fact that participants are all living their real C2 lives means that it would be hard to 'neglect' their 'situation' (Goffmann 1964) even if I wanted to. So given my intention to investigate variables in context, it is necessary in this hypothesis-formulating chapter to check insights into the influence of environmental factors on learning.

My contexts are the relatively structured setting of an intensive pre-sessional English course during Phase One and the more naturalistic settings of the participants' specialist training institutions in Phase Two. Both phases, of course, take place in C2, so the UK cultural setting is a continuous influence. In Tucker's 1978 terms, I have 'instructional settings' and 'socio-cultural contexts' to account for.

The two main general sources of information from previous applied linguistic inquiry are those looking at the formal language learning setting, particularly the classroom in comparison with natural acquisition contexts, and those looking at the question of bilingualism.

Most recent discussion of the foreign language learning classroom as an environment emphasises its unhelpful atypicality. It is (see, for example, Holmes 1978) an over-formalised context, 'superposing' an inappropriate language variety which negatively affects role relationships, topic, medium, functions, turn-taking strategies. D'Anglejan (1978) cites empirical evidence of the inferiority of classroom L2 learning compared with acquisition in naturalistic settings and Kennedy (1973) claims quite simply that 'it is not learning second languages
which is difficult, but learning them in classrooms' (p.77). Yet, as Schumann (1976) points out, a lot of L2 learning takes place in classrooms. Research certainly cannot afford to ignore actual teaching programmes, especially since learners on them are useful informants on the subject of what is learnt versus what is acquired.

The current tendency is to attempt (with a varying degree of urgency) to recreate natural acquisition conditions in taught course contexts (see, for example, Savignon (1972), D'Anglejan's survey article (1978) and Allwright's minimal language teaching strategy (1977), but without sacrificing classroom-specific advantages such as its potential for accuracy work, where accuracy is seen as 'a short cut, as a way of enshrining the central truths of the target language, so that subsequent modifications take place as experience is gained of a wider and wider variety of situations' (Brumfit 1979 p.187). Bowers (1980) also suggests that the language classroom can still play an important role, that there are procedures allowing the teacher 'as initiator if not controller of the proceedings' the chance of 'ensuring that the place of the individual in planning and participating in the learning process is not suppressed by the predetermination of a 'syllabus' or goal-oriented approach, or by the built-in constraints of the group context' (p.72).

Schumann's 1977 analysis of social distance characteristics of language acquisition situations is the neatest collection of variables in the broader community context. Social distance depends on the perceptions of L1 and L2 groups of who is dominating whom; the degree of cultural integration desired by both groups (in terms of assimilation, mere acculturation or preservation); the degree of enclosure (ie social
separation of institutions) wanted by each group; group size; cultural congruence; typical inter-group attitudes and the intended length of residence of the L2 group in the host culture. Acton's 1979 addition of the notion of perceived social distance (that it is how individuals see the gap between themselves and the host culture that matters) is a useful insight. It is very much in tune with my own intention of letting participants speak for their own constructs and with the issue of expectation stressed in 3.3 above. H.D. Brown (1980) also develops Schumann's ideas in his 'optimal distance model of second language acquisition', suggesting that it is at the third stage of acculturation, a stage characterised by culture stress and anomie that 'an optimal cognitive and affective tension' is actually conducive to progress in L2 (p.161). This will be an interesting point to investigate, especially as participants are seen in their initial close-knit group, then in their separate wider worlds. There is a clear link here, too, with the U-curve hypothesis mentioned in Chapter Three below.

Now it is important to note that Schumann's list is of social factors influencing group relations. The 27 participants in my study are not a speech community in Gumperz's 1968 sense of:

"any human aggregate characterised by regular and frequent interaction by means of a shared body of verbal signs and set off from similar aggregates by significant differences in language usage" (p.219).

Even in Phase One of my study when participants are together as the ELTI 'class of 1980', they do not really qualify, though aspects of their dynamics and cohesion will be examined. So while factors like Schumann's will inform my empirical inquiry, the focus will be on how they affect individuals in a foreign community. Many of the studies
mentioned under the various headings in this section are from sources where the environment is a bilingual community and the residence of the learners is permanent. Social factors that have been identified as particular to overseas students, especially those temporarily resident in the UK, are covered in Chapter Three. There is, of course, considerable overlap between the situations of a permanent entrant to a bilingual community and a temporary EAP visitor to a monolingual one.

Another social dimension, which is of indirect rather than direct interest in my study, is the cross-cultural dimension. My primary interest is in the characteristics and experiences of a group of individuals from various cultural backgrounds in a particular set of UK environments. It is not primarily in cultural thought patterns (Kaplan 1966) or values (Welte 1977) or cultural profiles (eg Hawkey and Nakornchai 1980), seen as typical of whole groups from the same background. Cross-cultural observations made by the participants themselves or by those in contact with them, will be referred to in my study, however.

In both phases of the study, the learning and living context of the participants will be explored in some detail. They will be seen interacting in performance tests, in class, in experimental and discussion groups. They will also describe their own perceptions of programme and community features through both controlled and free elicitation procedures at various stages throughout their UK residence.

4. **Language in Study Use**

The foregoing re-assessment of cognitive/affective and social factors reveals them as at least potentially suitable elements in Research
Hypothesis 1, which, as diagrammatically represented in Chapter One, has them as individual or combined variables on the independent variable end of the hypothesised relationship:

\[
\begin{align*}
\text{cognitive/affective} & \quad \text{--- target language} \\
\text{and social factors} & \quad \text{learning and use}
\end{align*}
\]

Their significance and inter-relationships will be tested in specific hypotheses in the empirical study. But what about the other side of the relationship? We have already seen that the learning/language learning overlap will make the formulation of testable hypotheses rather complicated, but we still need some kind of framework for specifying the features of TL use that can be assumed to be relevant for participants using it in C2 student life. Such a framework will need:

1. a full communicative scope in line with the competence construct developed in Chapter Four
2. a particular reference and relevance to the type of overseas EAP students who are the focus of this study (see Chapters Three and Five)
3. a suitability for use in the construction and evaluation of the kind of tests and activities required by my empirical methodology (see Chapters Four and Six).

The original informing source must be the events of real life. Descriptions of relevant features of these are of the following overlapping kinds:

1. 'needs analysis' models (eg Munby 1978) specifying language features from which syllabus items appropriate for stereotype students faced with specific EAP training may be selected
2. 'text-pragmatic' approaches (eg Tadros 1978) and 'text-semantic' approaches (eg T. Johns 1978) where the discourse functions and information structure of text from particular subject areas are examined

3. analyses of situations in which students have actually experienced difficulties (eg Jordan 1977)

4. test instruments developed for use with the kind of EAP students we are concerned with (eg B Carroll 1978).

The message from such sources is that those whose interest is in describing what is important for EAP learners in C2, for example, teachers, testers, course designers in departments or units helping overseas students with their English, describe their needs and problems in terms that are as broad in scope as the continuum plotting learning processes, skills, habits etc. in 3.6 above. The problem, therefore, is not finding evidence of needs and problems but of organising them as a useful source of reference for my empirical purposes.

Candlin, Kirkwood and Moore (1978) relate 'study skills' and 'linguistic skills' through a framework having study modes as the superordinate with macro skills such as listening comprehension, note-taking as hyponyms, their model thus cutting across the traditional four skills or 'epiphenomena' (Corder 1973). This is a useful means of description though it involves the repetition of the same skills (eg 'oral delivery') under more than one mode. Also, their description does not cover narrower, more purely linguistic elements such as 'time, tense, aspect' (Leech and Svartvik 1975). In fact Leech and Svartvik's view of form and meaning as seen below, is another useful insight for my descriptive framework:
My framework in Figure 2.6 overleaf learns most from Candlin, Kirkwood and Moore and from my own continuum in Figure 2.5. Helpful influences on my 'descriptive levels' are acknowledged on the diagram itself. The actual items specified on the grid come from the predict, discourse and problem analysis sources described above. The fourth source, namely the field of language testing, is more focal in Chapter Four below.

The horizontal axis on the matrix in Figure 2.6 uses the study modes conventionally identified as part of most academic training (though see Chapters Three and Seven for evidence of blurred distinctions between various modes of study). The vertical axis attempts to categorise linguistic and study-related levels of description for what has been previously identified as important for the kind of learners involved in my investigation. The examples actually plotted on the matrix are located under the study mode where the acknowledged researchers originally identified them. The horizontal lines indicate their potential 'spread' into other modes. Examples of perceived areas of

**FIGURE 2.4: A SUMMARY OF PROBLEMS RELATING TO COMMUNICATION IN TL-MEDIUM STUDY**
difficulty may also problematically overlap the vertical levels of description. When they do, vertical dotted lines suggest the fact. The intrinsic subsumption of lower levels of descriptive categories by higher ones (eg phonological features by morphological) is implicit in the structure of the matrix (as well as, of course, in communication itself) and does not thus need to be made explicit in this way. So, to Figure 2.6 itself, a distillation of received wisdom on the what of EAP problems, a prerequisite for study attempting to relate the what with the who and the why.

Since the formal, functional, study-skill and attitudinal examples plotted on the grid have all been identified as actual sources of difficulty for overseas EAP students, they will serve as a key reference source for my empirical study in the following ways:

1. as indicators of the kind of study-related activities and problems that should inform the construction and evaluation of the competence and performance tasks to be used in my study

2. as indicators of the kind of study-related points on which, along with the broader community factors explored in Chapter Three, I should seek further evidence in my follow-up visits to participants at their receiving institutions

3. as indicators of potential relationships between target language learning (eg by participants on the pre-sessional course) and target language use (eg on their subsequent specialist training programmes).

Even a glance at Figure 2.6 reveals (through the obvious concentration of examples at the second four levels of the matrix) that interested parties see higher-order language and study features as the main required
focus of attention. This, of course, is in line with my own broad communicative construct. It may make the formulation and testing of narrow hypotheses more problematic, but it is clearly the way things are currently viewed.

5. Research Hypothesis 1 Formulated

The quality and quantity of the theoretical discussion and empirical research re-assessed in this chapter is considered sufficient to support the further pursuit of corroboration of my thesis (T) that training for EAP learner/users can be fruitfully informed by the systematic, sensitive, multi-dimensional profiling of individuals. It is also considered as sufficient evidence that my main Research Hypothesis (RH1) is a valid area of investigation. Thus I can, as suggested in the summary in Chapter One, Section 5, formulate the 'range of convenience' (Kelly 1955) of RH1:

\begin{align*}
\text{individual} & \quad \text{and combined} \quad \text{target} \\
\text{cognitive/affective} & \quad \text{language} \\
\text{and social factors} & \quad \text{learning} \\
& \quad \text{and use}
\end{align*}

The investigation of the many specific hypotheses that implicitly constitute the general relationship suggested here is the task of my empirical study (see Chapters Five to Eight). Again, there is ample evidence in the research just re-assessed that the direction and directness of the causal relationships between variables involved in my hypothesis will be key issues. Is variable A causally related to variable B? B to A? A to B, which then causally relates back to A? Or is there a causal A:B relationship with the intervention of variable C? The hypothesis-testing task of the empirical study will be to put names to the variables involved in RH1 and to put arrow-heads on the lines connecting them.
CHAPTER THREE

THE OVERSEAS STUDENT
1. **Introduction**

This chapter has the following main purposes:

1. to re-assess previous investigations into the influences of the foreign culture \((C_2)\) on overseas students' progress and problems.
2. to learn from the methodologies of such investigations.
3. to formulate Research Hypothesis 2.

There is inevitable overlap with Chapter Two above, where many of the insights gained also came from surveys or studies (the distinction is quite often unclear) of overseas students. However, most of the Chapter Two sources and especially those cited in Figure 2.6 have the TL as their main focus, those connected with its teaching as the main focusers. Most of the inquiries considered here do not. They tend to come mainly from non-ELT professional sources and/or to be investigating the TL issue among many others. I should be leaving a vital source of topic and methodological insight untapped if I did not make at least a selective survey of such surveys. The task is essential to the formulation of Research Hypothesis 2, which posits a relationship between multi-dimensional learner profiles and degree of success in a foreign culture.

2. **The Overseas Student Question: sources and foci**

One way of classifying the multifarious inquiries into overseas students (defined here as those with a 'permanent home residence' (Overseas Students Trust (OST) 1979 p.2) in a country other than the one they are studying in at the time they are questioned) is according to the sources of those inquiries. I shall use the following categories: aid donor agencies (eg the Overseas Development Administration (ODA) or the British Council); associations of interested parties (eg the
Overseas Students Trust (OST) or the National Association for Foreign Student Affairs (NAFSA); individual receiving institutions and individual researchers. Relationships between the source of an inquiry and its priorities will quickly become clear.

2.1 Aid Donor Agencies

Training and Development (British Council 1981;1) is an evaluation report that is:

"an ex-post study is one carried out after the event .... Firstly to assess the effectiveness of its aid activities and secondly, to learn lessons for improving the effectiveness of future aid activities" (ODA 1979).

Like most such reports commissioned by the ODA or the Council, it thus collects data from trainees after they have returned home and had time to judge the value of their UK training. Quite logically the questions to which answers are sought are on the choice of fields of study; the selection of awardees; UK placement; pre-departure briefing; living conditions in Britain; TL and specialist subject difficulties; the personal benefit and relevance of the training on return to C1. The 1981 report covers 238 returnees in 10 countries, summarising its data in a mainly qualitative, case study way and reaching a conclusion relevant to its remit, that 'training was a good British investment and valuable to development in the 10 countries' (p.1).

Recent bibliographies of ODA and Council inquiries (eg in Jacobs (ed) 1978, Technical Cooperation and Training Department (TCTD), British Council (1981;3)) indicate that they mostly share the foci of Training and Development, though usually dealing with the training programme of a single country at a time (eg Evaluation of Training Programme
While not all aid donor agency studies are carried out in the trainees' countries of origin, their UK-based investigations also underline the relationship between the source of a survey and its focus. Chapman (1976) reports on the living conditions of Study Fellows in the UK and Courtenay and Makinson (1979) on problems of separation from the home environment. The emphasis and tone of both the elicitation techniques and the analysis of data are quantitative and administrative rather than qualitative and personal. For Chapman:

"The point at issue was the absence of any marriage or family allowances for TA (Technical Aid) Study fellows, although marriage allowances had been paid for Commonwealth scholars since 1959" (p.1).

Of the 78 Tables in her statistical analysis, only 11 are about social or linguistic matters, none about academic ones. Courtenay and Makinson stress that:

"The main focus of this study is a consideration of ways in which Study Fellows have coped with the problems of separation from their home countries, and their views on accompaniment and home leave."

And while they acknowledge that 'coping successfully with separation difficulties is linked with the ability to adapt to the new environment', they 'could not probe deeply into issues of adjusting to life in the UK' (p.64). I do have to probe into such issues, of course, and with students from just the kind of populations that the ODA and the Council have learnt so much about. The approaches of the kind of reports referred to here (as well as the valuable advice received from those involved in them) constantly inform my own study.

None more so than Factors affecting the Performance of ODA-Sponsored Study Fellows (British Council 1981;2). This investigation, carried
out by the Council's English Language Division, looked into the following variables: background, specialist subject knowledge, sex, age, TL level, motivation, health and accompaniment. The sample was 61 Study Fellows from Egypt, Thailand, Indonesia and the Yemen Arab Republic, data for analysis based on documentary evidence, pre-departure TL test results, student questionnaires and semi-structured interviews with students and tutors at receiving institutions. The main findings are that the 'unsatisfactory' performance of 30% of the sample was caused mainly by 'inadequate language proficiency and motivation and a failure to adapt successfully to the new academic and social environment' (British Council 1981;2 p.1). This project is of particular interest to my study since I took part in its planning and implementation; discussion in Section 3 of this chapter will show how I was fortunate enough to gain 'piloting' experience from it. The relevance of the project's population, foci and general methodology are already clear.

Attitudes and Social Relations of Foreign Students in the United States (Sellitz et al. 1963) is another example of establishment-sponsored research, reporting on two overlapping studies, one backed by the Committee on Cross-Cultural Education of the Social Science Research Council, the second by the International Education Exchange Service of the State Department. Apart from their obvious interest as data on a different host culture, these studies again offer a relevant set of foci, mainly factors arising from the search for answers to three key questions:

"(1) What factors - and especially what environmental conditions - influence the development of social relations between foreign students and members of the host country? (2) What effects do these social relations have on foreign students' attitudes toward the host
country and its people? (3) What are the effects of a preliminary orientation period designed to ease the students' transition to their new situation?" (Selltiz et al p.6).

As the tone of these questions suggests, the emphasis here is on socio-cultural aspects of the foreign environment rather than on the characteristics of the guests. And there is comparatively little on language or specialist study problems and progress. Still, the research will prove informative to the present study for its thorough examination of cross-cultural contact and adjustment as well as some of its honestly admitted half-findings:

eg "Yet students who differed in their command of English did not differ, on the average, in the extent of difficulty they reported with academic work. .... Had we considered the students' self-ratings of their competence in English, we might have found - as we found concerning the development of social relations - that confidence in one's ability to communicate had a greater influence on adjustment than actual ability to communicate as judged by another " (ibid p.259) (my emphasis).

It also provides a useful re-assessment of well-established theories such as the U-curve hypothesis (Sewell and Davidsen, 1954), briefly, the tendency towards a high-low-high pattern of morale during a C2 study period, or the National Status Hypothesis (Morris, 1960), the effect the C2 view of a student's own country has on his feelings about the host culture.

2.2 Associations of Interested Parties

One of the most important sources under this heading has of course already been tapped in Chapter Two, especially in Figure 2.6, where key EAP foci as identified by receiving institution organisations such as SELMOUS are described. Here, therefore, I shall again concentrate on a selection of broader inquiries from associations of interested parties, starting with a study from the less than recent
past, Commonwealth bursars: problems of adjustment (Burns, 1965).

In this, the emphasis is on English language background and use, responses to different teaching methods used on specialist courses, and study problems, techniques and habits. Such an emphasis clearly reflects the interests of the 'committee of tutors' responsible for the inquiry. A rather surprising finding, however, is the following:

"None of these students admitted to any sense of handicap (except for a small number of difficulties arising from local usage) either in the use of English in their studies or in communication; and though a closer and more objective assessment would no doubt have revealed many inaccuracies in expression and understanding, it is at least reassuring that these obtruded so little in the students' daily use of the language, whether in writing or speaking" (p.40).

This may of course be because of the preponderantly English-medium background of the bursars concerned, but it seems more likely that the open questionnaire plus sample-interview research design did not adequately tease out the linguistic and the non-linguistic threads of the learners' communicative situations.

The finding is certainly not typical of many overseas student tutors' views. In their Coombe Lodge Report (1971), Further Education College tutors give a more characteristic reaction. They underline the complex way in which linguistic and study demands interact and consider how they may be analysed, assessed and met. Collins points out how well some overseas students can perform on their academic courses even though their second language competence is apparently low. This is partly because other learner and learning factors override target language competence, partly because the tests used to assess that competence are invalid. Laing claims to have tried the Davies, Ingram, and the Northern Universities JMB tests with his overseas students at Leeds University, but found that none of them 'adequately assesses the
diverse skills required for university study' (p.12). What is needed is a measure that takes account of language level, required study activities and individual learning strategies and styles. Chapters Five to Eight of this volume may have a contribution to make.

The OST has as its 'broad purpose' 'to improve the provision made in the UK for the increasing number of students from overseas and thereby help to ensure they (make) the best use of their stay here' (see note in Williams (ed) 1981). Work to this end commissioned by the Trust should thus be especially relevant to my study; Freedom to Study (Reed, Hutton and Bazalgette 1978) certainly is. This inquiry seeks to let the students speak for themselves and looks for 'hidden needs rather than the expressed wants' (p.20). In terms of the range of its foci the report has as broad a coverage as any and in a number of areas, for example, the reasons why particular facts of foreign culture life are problematic for particular groups of students, the students' view of what social adaptation is really about, and the distinction between welfare and faring well, Freedom to Study appears also to probe deepest. Its use of the role concept as a framework for the interpretation of its data is neat and persuasive. With role defined in the terms of George Kelly (1955) as 'an ongoing pattern of behaviour that follows from a person's understanding of how others who are associated with him in the task think' (Vol I, p.8), the report adopts the following position:

".... to reinforce the significance and importance of the student role, but at the same time to maintain that only as the client and visitor roles are respectively identified and accepted by the individual from overseas, is that individual likely to be able to experience the necessary authority and freedom to pursue his studies" (p.142).
It will prove interesting in the context of my research, with its emphasis on individual differences, to see how the interaction of perceived and expected roles as it affects particular members of my group at particular times may contribute to a re-evaluation of some of the broader hypotheses already mentioned, the U-curve and the national status hypotheses for example.

The OST also commissioned The Overseas Student Question (Williams (ed) 1981), which is not a survey, but a collection of papers putting the situation of those whose studies are hosted by Britain into an up-to-date historical, political, economic and international perspective. The emphasis is on 'the multi-dimensional nature of the overseas student problem' (p.223) and developing a coherent, lasting policy to help solve it. The Survey of Overseas Students in British Higher Education 1980 (an Appendix to the book), is a useful source of presage data for my study (see Chapter Five below). These two OST-sponsored reports can, as suggested in the 1981 volume, be seen as complementary:

"Taken together, we hope that these two books can chart a course that will be beneficial to all the interests engaged in this important question, not least the overseas students themselves" (Foreword to Williams (ed) 1981 p.xii).

For my study, the earlier of the two inquiries has extra methodological implications (see Section 3 below) but if my findings by the end of Phase Two are at all conclusive, they must relate to some of the broader issues broached in the later book.

Klineberg and Hull's ambitious work, At a Foreign University - an International Study of Adaptation and Coping (1979), is carried out in the interests of a large number of receiving institutions. The initiative for this investigation came from the International Committee
for the Study of Educational Exchange, which is made up of senior academics from 14 countries. What is unique about the survey is its truly international scope as it looks at the situation of students of more than 100 different nationalities during their studies in no fewer than 11 host countries. Nor does the size and spread of the sample population limit its range of foci, though its emphasis in depth is on the importance of the ability of overseas students to establish social relationships within the host culture. This 'modified culture contact hypothesis' is found to override both the U-curve and the national status hypotheses:

"those foreign students who are satisfied and comfortable with their interactions with local people and the local culture during their sojourn would report broader and more general satisfaction with their total sojourn experience, not only non-academically but also academically" (p.53).

But, as the authors admit, 'contact is a complex variable'. The study is not fully convincing in unravelling it or in identifying the subtly inter-related characteristics which may differentiate individuals' perception of and need for 'contact'. With my much smaller sample I can go into this issue somewhat more deeply, particularly in terms of its implications for academic study activities, which, like second language factors, are dealt with only fairly cursorily by Klineberg and Hull.

The NAFSA study Needs of foreign students from developing nations at US colleges and universities (Lee, Abd-Ella and Burks 1981) has an aid donor agency connection since the contract for it came from the Agency for International Development (AID). NAFSA itself, however, is an essentially interested party association like UKCOSA (the United Kingdom Council for Overseas Student Affairs), so the report can logically be evaluated in this section. With a nationwide sample of
nearly 2,000 overseas students (a response rate of 29%), the study tests hypotheses derived from categories of 'needs' of the following kinds: information, degree programme, training relevance, extra curricular professional activity, academic life, finance, community life, housing, family, interpersonal relationships, future work and TL. Data collection is entirely by postal questionnaire, with responses submitted to factor analysis to formulate 46 'composite' need categories and univariate analyses to establish the prominence of the needs or goals covered. Obviously, with more than 200 questions to respond to in terms of the importance of particular needs and the level of their satisfaction, the study is useful as a check on coverage and for the student priorities it reveals. (The needs for enough money, a degree and specialist skills; satisfaction with degree, breadth of education and specialist knowledge obtained, emerge as the top three in the two categories (Lee et al. 1981, pp.49-50)). But I am not sure that the study takes into account the influence of perceived strength of need on perceived level of satisfaction, the extent to which it is more or less likely that a strongly or weakly felt need will be more or less easily satisfied. With my case-study sized group and my variety of elicitation approaches, I should have more chance of doing this (see Chapters Five to Eight below).

Also from official interested party association sources, yet as essentially 'subjective' as the Lee et al. study is 'objective', are my final two selections in this section, which probably come closest of all to expressing the feelings of the students themselves. Disappointed Guests (Tajfel and Dawson (eds) 1965) and Suffering for Success (UKCOSA 1979) are collections of essays written in response
to invitations by the Institute of Race Relations and UKCOSA to overseas students to write about their experiences in the UK. From my point of view they are relevant as the most revealing sources of insight into the question of discrimination. In particular, they corroborate an inference which one began to draw even early on in this present study, namely that direct (A \rightarrow\leftarrow B) causal relationships are comparatively rare. A Kenyan student writing in Suffering for Success, for example, appears to face in the foreign culture almost all the experiences that 'should' cause 'failure'; the experiences described would also certainly seem to indicate a serious failure on the part of the host culture. Yet she does not fail academically or in terms of adjustment; in fact, she considers staying on here to help. The intervening variable, what really makes the difference and prevents the expected causal relationship, is the attitude and constructs of the student herself. My research has to examine thoroughly and interpret sensitively both sides of any posited relationships and whatever intervenes.

2.3 Individual Receiving Institutions

Many of the investigations by receiving institutions into overseas student problems have, as we have already noted, been covered in Chapter Two, especially those with a particular focus on the TL and its EAP connections. Here, I re-assess a selection of other institutional studies where the emphasis is different but nevertheless relevant to my own inquiry.

In the report on Commonwealth courses at the University of Birmingham (Dalton, Project Director, 1979) the focus is mainly on the content and methods of courses and the students' evaluation of them on their return home. This interesting example of 'educational illumination and
criticism' surveys a population similar to Burns' (see 2.2 above) as regards national background, and who had attended Birmingham Commonwealth courses between 1960 and 1979. But this time problems with English are not found to be insignificant. Between 8% and 19% of the respondents (and the response rate was 31%) admitted to difficulties with the specified activities: 'everyday conversation', 'understanding lectures', 'writing essays and other papers', and 'reading the recommended course books'. Equally significantly, perhaps, the 'other (please specify)' option was also taken up:

"Other aspects of using English .... are cited as having been especially difficult; technical and educational terms; the correct form for essays and other papers; lectures which were given to classes combined with English students doing the three-year course; the use of the library, especially in obtaining reference books for writing of the dissertation; curriculum projects; language laboratories using earphones, and understanding sales assistants in shops" (p.48).

The surveyors and the surveyed have combined to portray rather vividly the complexities of the communicative picture. A possible criticism of the focus of the Birmingham survey in the light of these responses might be that whereas ex-students were asked to rate a reasonably comprehensive list of teaching approaches in terms of 'importance' and 'value', they were not required to rate them in terms of difficulty. My own follow-up investigations at receiving institutions (in Chapter Eight) will learn from this.

Parrish (1977) feels that the question of methodology may have been somewhat neglected in accounts of how overseas students fare. Thus, one of her main aims in her Reading University sponsored review study of relevant courses at five UK universities is to categorise and classify the whole range of 'methods and materials' that may form part
of the learners' academic experience. The result of her investigation based on questionnaires, observation, interviews and informal contacts (aimed at, according to focus, Course Directors, members of staff and students) was a 'scheme', derived from Verner (1962), of the 'methods, techniques and aids observed, or said to be used in the courses' (p.21). 

Parrish's review contains interesting evidence of the effect of the categories in this particular teaching/learning classification to remind any researcher attempting to investigate what students face and how they cope, of the importance of a comprehensive inventory of the methods, techniques and aids that are actually used. A set of categories such as Parrish's is vital if we are to allow for the many
combinations of methods, techniques and aids encountered.

It is also obvious that on the majority of academic and training courses the combinations and emphases change significantly at different stages; a single cross-sectional sampling of the methodology of a course is not normally adequate. Nor is it safe to make too many assumptions about relationships between level of studies and methods and difficulties. My Phase Two interview pro formae (Appendix 3) try to cover the whole range of inter-related main course activities, learning a lot from Parrish and from pilot work on the British Council project (1981;2).

The Report of the Working Party on Foreign Students' Command of English (Southampton University Students' Union (SUSU) 1979) is broad-ranging and student-oriented, coming as it does from a survey conducted as a result of pressure from a body whose function is to protect student interests. The main message of this small-scale but long-term survey (it covers overseas students from 1975 to 1979) is, firstly that investigations of the problem are becoming more urgent now, as the proportion of Commonwealth students with a second as opposed to a foreign language background in English decreases, and secondly a conviction that problems with English are a more significant factor in academic and social adjustment than people on all sides are prepared to acknowledge:

"Poor command of English is likely to make for a poor social life, and a good social life is of great value in improving students' English. Most importantly, students who are unhappy about being socially isolated are unlikely to work as well as they otherwise might" (p.20).

This statement may be regarded as further confirmation that you cannot
meaningfully investigate cognitive/affective factors and their inter-relationships with second language learning without taking the fullest possible account of the academic and social environment.

I shall conclude this section of my selective re-assessment of relevant inquiries with two brief reminders that not all overseas students are from the developing, nor all host institutions in the developed, worlds.

In their study, *Residence Abroad and the Student of Modern Languages* (Willis, Doble, Umasankar and Smithers, 1978), investigate how a period overseas affects foreign language proficiency rather than how foreign language proficiency influences the stay overseas. And the foreign languages here (French and German for British students) are the academic ends rather than, as is the case with my group, the means to other academic ends. Nevertheless, this University of Bradford study has considerable relevance. The tests used to measure the students' pre- and post-sojourn foreign language competence are an interesting combination of standard proficiency and communicative testing and where the testing methodology is more open-ended, serious efforts are made to achieve reliable assessment through analytic scoring criteria (compare Chapter Six below). Other independent variables objectively measured are attitude, motivation, personality (using Cattell's 16 PF Test) and type of placement (i.e. 'study' or 'work'); these variables are also foci in my study. It is hoped that, in general, this present research can go beyond the Willis et al. study in its attempt to identify and trace the inter-relationships between variables and to benefit from the extra insights gained from the more qualitative field study data collection techniques used (see Chapter Eight).
The final single receiving institution source concerns students from developing countries studying through English in another developing country. *The Asian Institute of Technology: Survey of ELT Needs* (Frankel and Dunlop, 1980) has proved quite useful, especially in the preparation of the data collection instruments for my interviews of tutors and students at their receiving institutions. The questionnaires designed by Frankel and Dunlop are particularly instructive in the way they combine broad coverage and economy in probing the prominence, value and difficulty of the various 'academic uses for English' encountered on the AIT campus in Thailand. The close matching of questions to students and questions to staff makes for convenient comparisons of the two points of view; as is often the case, there are significant differences between the two sides over what the main language problems and demands really are. My own findings on this appear in Chapter Eight below.

### 2.4 Individual Researchers

The opportunity for the personal selection of a research topic enjoyed by most individual researchers means that there is a less consistent connection between source and focus when you look at single-investigator studies on overseas students.

The study of Indian students in Britain (Singh, 1963) is appropriately evaluated at this point as it may be categorised as representing a guest culture and may serve to redress what has hitherto been a balance rather heavily weighted on the host culture side, at least in terms of research source. And indeed two of Singh's three main hypotheses focus on presage characteristics of visiting members of a guest culture, namely the social background and personality factors of the
Indian students he sampled and their effect on how the students survived here. Singh's study is one of the most comprehensive if judged by the number of different factors he covers. Perhaps this is because he, like his subjects, 'observed and judged the English against their own background'; the implication that where the researcher tries hard to look at things from the participant's point of view he tends to have to take account of a very broad range of factors, is an important one. It would seem to support my own multi-dimensional approach.

The well-known investigation by Sen (1970) into problems of overseas students and nurses also has a broad canvas. It covers problems under main headings such as 'methods of study', 'financial difficulties', 'problems of adjustment' ('educational' and 'social'), and 'inadequate preparation'. Sen's own statement of the purposes of the study underlines how ambitious it is:

"to derive an overall picture, based on information received from students, of their academic and social experiences in this country and the ways in which these factors appear to be related to each other; and .... in the light of the evidence yielded by these inquiries, to suggest ways in which the adjustment of overseas students might be facilitated" (p.xix),

Of all the factors considered by Sen in what amounts to a search for a predictor model for overseas student success, much the most attention is paid to English language proficiency as measured by the ELPT (Davies, 1964). Scores by Sen's sample on the 'short' version of this test are investigated in relation to a number of academic and social factors. Her conclusion that 'The English proficiency Test did not provide a practical guide to the final performance of students' (p.159) has been discussed elsewhere (Davies, 1977; Moller, 1977) and will be taken up again in this research. My own study learns from Sen's attempt to take
the broad approach to her topic. But it will probe especially carefully the 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 might be more suitable in research where diagnosis and prediction are at issue. In his report on Sen's study, Yates (1970) is concerned about the lack of 'any significant predictive value' for competence in English but his belief that 'the inevitable crudity of the (final main course pass/fail) criterion has served to obscure its relationship to the students' competence in English' may be missing the point. There could be 'crudity' at both ends of Sen's hypothesised relationship. Like Moller (1977 and forthcoming) and Ryan (1979), I see the need for a finer-grained diagnostic and profiling test battery administered after the arrival of students in the UK for use in the prediction of $C_2$ performance. Chapters Four, then Six to Eight, trace the development and use of such a battery.

Edwards (1978), like Sen, has overseas nurses as her target population. She, too, pays special attention to the TL question, and sees no reason to suppose that the overall picture of severe problems of communication (in its broadest sense) 'is necessarily confined to nurse-learners, to basic education and training or to England and Wales' (p.346). Of particular interest for my study are the following two points made by Edwards:

".... studies show that a speaker's language is a major influence on the impression formed of his or her personality. People judge a speaker's intelligence, character and personal worth on the basis of the language used" (p.342 and cf Stubbs, 1976 p.21).

and

".... it was evident from the comments of many overseas learners
that the period of adjustment was long and it was suggested that some overseas learners never recover from culture shock. Many of their problems of communication, which this research demonstrates to be only partially language-linked, could be overcome if proper counselling were available" (Edwards p.345).

Both the cognitive/affective aspects of communicative competence and the nature of culture shock are, of course, key foci of my empirical study.

3. The Overseas Student Question: methods of inquiry

All the studies in Section 2 above (and others mentioned in Chapter Two or not quoted at all) have informed my study to a greater or lesser extent as regards what needs to be taken into account when overseas students are the focus of inquiry. It is now time to see what can be learnt from how the investigations were carried out. Appropriate, though by no means mutually exclusive, headings for this are: sampling; the approach to hypotheses; quantitative or qualitative orientation and the role of the researcher.

On the first issue, that of sampling, the studies surveyed constitute a strong case for delaying detailed discussion of the question until it arises in one's actual empirical inquiry, in my case in Chapter Five, where the status of the group I work with is described. None of the investigations assessed here or in Chapter Two satisfy the strict demands of random sampling (see Chapter Five, Section 3 below) whether their 'sample' size is small (eg Holes (1974) N=37) or large (eg Sen (1970) N=2918). The reasons for this will also be discussed below. It is relevant here, however, to distinguish between surveys in the narrowest sense, that is, investigations "where a carefully selected sample of people are asked questions .... we know exactly what the figures mean and to what percentages refer" (Hancock, 1964 p.9).
and studies where it is not just a matter of asking straight questions and percentaging answers. In the first category we can include Chapman (1976) or Courtenay and Makinson (1979) where the focus is primarily administrative, where the purpose is to elicit and analyse answers to questions on a specific issue so that the sponsoring agency can decide whether or not to make administrative adjustments. In cases like this, the basic survey method (questionnaire with or without a double-checking follow-up interview of a sub-sample) is used; we are really talking about a question to be answered rather than hypotheses to be formulated and tested. Lee et al. (1981), on the other hand, exemplifies the basic survey methodology, all data collected by means of a postal questionnaire sent to a carefully selected sample N=6523 (though, as usual, with a low response rate (29.3%), but with hypotheses (in this case 26 of them) tested on the basis of the responses. It is tempting to assume that large-sample based studies belong at the classical scientific end of the methodological continuum, small-sample research at the ethnographic end. This would be an over-simplification. If the sampling frame is carefully constructed, the actual sample large but depleted and (as with Lee et al. 1981) with the effects of depletion not taken into account, claims to scientificness are doubtful. If the sample is small but carefully described with its potential representativeness explicitly evaluated, (as with Reed et al. (1978)) a study may still be reasonably scientific. Sample size or representativeness alone do not guarantee scientific respectability.

A brief look at the approach to hypotheses is relevant both to the status of various inquiries in terms of the methodological paradigms discussed in Chapter One above and to my own design. The Klineberg
and Hull study (1979) has as one (among several) of its a priori objectives the examination of one new and two existing hypotheses (see 2.2 above). Their research design, however, is only partly based on their interest in these; other data, both quantitative and qualitative, are freely allowed to influence their hypothesis-testing objective.

Many of the studies surveyed do not formulate or test explicit hypotheses at all; like Sen (1970) or Willis et al. (1978), they set out to collect and evaluate data then draw conclusions or make recommendations but without hypothesising in the scientific sense. And my distinction (see Chapter One above and the empirical study itself below) between general or 'research' hypotheses and their narrow, specific constituent hypotheses is very rarely made. It is commoner in research reports (eg Neufeld 1979 in Chapter Two, Section 3.2) than in theses or large-scale projects.

Reed et al. (1978) come closest to the anthropological paradigm in their approach to hypotheses in that theirs are mainly a posteriori. This is not to suggest, however, that they collect data without a framework; they had in fact already 'classified' their target factors in advance and tried them out at 'initial interviews' (p.17). If this is borne in mind, the reformulation of their hypothetical framework on the basis of their group discussions and in-depth interviews can be seen as an example of the anthropological brand of hypothesis testing.

A warning at this stage from Malinowski himself:

"The greatest source of all the inadequacies and gaps in my own field work has resulted from the dire methodological fallacy: get as many 'facts' as you can while in the field, and let the construction and organisation of your evidence wait till you write up your material" (Coral Gardens and their Magic, 1935 p.467).
The design of my study builds in both quantitative and qualitative elicitation procedures for the investigation of the narrower hypotheses (for example Chapter Seven, Section 3) and the broader hypotheses of which they form part, formulated at the end of each of the first four theoretical chapters.

The data elicitation and interpretation procedures I use for Phase Two of my study (see Chapter Eight) are probably the best illustration of where the study stands on the question of quantitative or qualitative orientation. This is because they include (like Willis et al. 1978) the use of standardised or standardisable tests as well as (like Reed et al.) group discussions and open-ended interviews. But it is unhelpful to assume too clearcut a distinction. The majority of the studies surveyed here might be categorised as quantitative rather than qualitative, with their concern with prevalence rather than range, use of identical questionnaires with limited choice responses as their main elicitation instruments and oral interviewing chiefly as a check on questionnaire data rather than a source of new insights. But most committed scientific sociologists would see quantifiability as implying full statistical validity. None of the studies mentioned here would qualify on this basis, what with their biased samples and widespread use of arbitrary scales and aggregating. And at the other end of the continuum one notes Becker's (1958) point that even 'qualitative' field-study interpretative accounts are ultimately dealing with 'implicitly-numerical' 'quasi-statistics' (p.656). Reed et al. use no statistics when they are analysing their descriptive reports; but they are making quantitative decisions throughout. For example before claiming:
".... Malaysians accepted that it was logical for them to come to Britain .... but there was some regret that they were tied by the language .... and there was a suspicion that the dependence has been exaggerated" (p.96) (my emphases), some kind of quantitative inferencing must have been done.

The data sheets for student interviewing on the 1980 British Council Project (British Council 1981;2), which I helped to design and used in twelve in-depth interviews, served as 'pilot' orientation for my research, since they were for use with a similar sample and had a similar focus. (In fact, the overlap and sharing of experience between these two investigations may be seen as a practical example of the cooperation and comparison advocated by Hymes (1979) Chapter One, Section 3 above.)

In general, the interviews and the reporting procedures served their purpose, but there were problems. The open-endedness and lack of specificity of the data sheets and the interviewer report sheets made subsequent data analyses difficult, especially as a number of different interviewers were involved. Learning from this experience, I tried to ensure that my own pro-formae for Phase Two student and staff interviews at receiving institutions (see Chapter Eight and Appendix 3) were more amenable to quantitative analysis (eg through the use of pre-coded numerical values for responses) yet flexible. Probes are more specific in their focus on, for example, the modes, methods and content of the training programme, language and study problems. On questions of language in social use, however, it is for the participant to describe what he does, with whom, when and with what reactions. The desired free talk in this area might be hindered rather than helped by specific suggestions from the interviewer; they are not, therefore, made. When
it comes to possible personal problems, changes in attitude and personal
development, there are areas that it is necessary to explore to test
various hypotheses. Key topics are thus specified but with room for
free-ranging talk. In fact, wherever the interviewer's form is quantit-
atively pre-coded, there is also provision for 'qualitative comment'.

Now although the in-depth interview data sheets for both the students,
and staff at the receiving institutions have been informed to a greater
or lesser extent by all the studies discussed in this chapter (and
others not actually referred to) my own elicitation instruments do not
look very much like any of the fifteen interview schedules or postal
questionnaires I collected and studied and the main difference would
seem to be that my interviews are based on general and specific topics
for discussion, not on questions and responses.

If this results, as it sometimes did with the similar though still
'pilot' pro-formae used on the earlier project (British Council
1981;2), in the kind of revealing, spontaneous interaction that are
the virtues of the approach, then this part of my follow-up study will
actually have been 'qualitative', in the sense that insights emerge
that the constraints of quantifiability might have left hidden. If
the coding and quantification provisions built into the design can be
shown to have been practicable and prove valid in the context of the
research as a whole, then some of the typical problems of analysis and
interpretation will have been overcome.

The final question that may help categorise my own and other studies
in terms of the dominating paradigms is the question of observer
status. Now it is interesting that there is comparatively little
explicit discussion of this issue, perhaps because most of the researchers concerned would prefer to claim the non-status of the 'objective' observer, the 'fly on the wall' (Hamilton and Delamont 1974), constrained by the precept that any independent scientist viewing the same reality with the same techniques should end up with the same observations. This is probably the case when interviews are used only with a sub-set of the larger non-personally contacted population (eg Sen (1970), Frankel and Dunlop (1980)). Singh and Klineberg also interview sub-sets, but their interviews are to add to the data. Singh is rather vague about the methodology and status of his interviews:

"The final report is based mainly on the analysis of data from four hundred cases of the extensive survey and sixty-seven cases of the intensive survey. In addition, materials have been drawn from interviews, case-histories and participant observation" (p.12) (my emphases),

but he does mention spending time in places where he could meet Indian students, 'long, unstructured interviews with individual students' and that 'on several occasions the problems of Indian students were made the topic of group discussions' (p.127). Klineberg and Hull, too, use interviews to do more than run a check on questionnaire data:

"Our purpose (in the beginning, mid- and end-of-year interviews of 20 students in 7 countries) was to look at the foreign sojourn as a miniature life history which would enable us to see more clearly the gradual progress of adaptation to the new environment and the success (or occasional failure) in coping with the difficulties that arise" (p.5).

Although Klineberg and Hull worry that their interviewers did not all take the same approach or achieve 'temporal' consistency, they seem pleased that some interviewers became friends of the interviewees since the resultant personal contact and rapport encouraged 'freer expression of emotional reactions both with regard to the negative and positive aspects of the experience' (p.101). Parrish, who admits from the out-
set that her investigation is not 'conceived in research terms to test hypotheses or even produce data within a well-defined structure' (p.2), spent ten days in each of the departments involved in her study. Her role was clearly not neutral and interviews of one kind or another were clearly central to her methodology:

"The periods of observation, interviews with participants, the completed questionnaires, and informal discussions served to illuminate the significant features, issues, and problems associated with the teaching of these courses. It was also possible to draw upon and mobilise the knowledge and understanding already existing in each Department, to make it generally available to all concerned" (p.9).

In Freedom to Study the qualitative data that can emerge from in-depth, observer-participatory interaction is even more crucial. In this study, it will be remembered, there was no administration of questionnaires, the belief being that 'the essential difference between data obtained from a questionnaire and from work with this ('subjective') model is that it enables us to have direct access to underlying attitudes' (p.21). So, interviews and, even more important for their higher 'degree of inter-subjectivity' (Cohen and Taylor 1976), group discussions, provide the data if the observer or observers participate appropriately:

"Because of the complexity of the feelings and attitudes of the students, the methodology called for the full involvement of the field workers with the overseas students so that they could be caught up through that full involvement with what it felt like to be in the place of the student" (p.21).

The 'full involvement of the (three) field workers' is shown vividly in the descriptions of the group discussions and interviews, both of which methods used only broad key questions (eg 'What has been your experience since you came to the UK to study?' (p.44)) with prompts and probes formulated and used as the dynamics of the interaction seemed
to warrant. The bold, subjective, interpretative comment reflects the constructs of the observer as well as the informant:

"By the end of this session, especially since it was my own first session on the project, I found myself feeling guilty about the various difficulties these students had experienced. I was most struck by their loneliness and the British insensitivity which they recounted, including the apparent conflict between the British student's view of how to behave and their own. However, I did reflect on my own experience of working away from one's own home ...." (p.48).

But from such qualitative data collection and interpretation, Reed et al. go on to formulate their main hypotheses and in fact make more sweeping and radical recommendations on overseas students than any of the other studies covered here.

In inquiries where the whole of a largish sample population has been interviewed, for example Courtenay and Makinson (1979), Selltiz (1963), the assumption can usually be made that the interviews were structured fairly tightly round a quantitative questionnaire designed for maximum inter-interviewer uniformity. Observer role here, neutral.

Before I attempt to compare my own observer role with the way it was played in some of these studies, it might be timely to look briefly at the participant observer ideal according to social anthropological orthodoxy. Kluckhohn (1940) defines it thus:

"conscientious and systematic sharing, insofar as circumstances permit, in the life activities, and on occasions, in the interests and affects of a group of persons. Its purpose is to obtain data through direct contact and in terms of specific situations in which the distortion that results from the investigator's being an outside agent is reduced to a minimum" (p.331).

It is interesting that proponents of both the contrasting paradigms worry about the 'distortion' that may be caused by the observer presence. But for the scientists the solution is to neutralise the
observer, for the anthropologists it is to involve him. Lutz and Iannaccone (1969) suggest this continuum:

<table>
<thead>
<tr>
<th>Observer Role:</th>
<th>role</th>
<th>natural role in</th>
</tr>
</thead>
<tbody>
<tr>
<td>foreign</td>
<td></td>
<td>the target society</td>
</tr>
<tr>
<td>to the target</td>
<td></td>
<td></td>
</tr>
<tr>
<td>society</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

They also provide a useful grid indicating some of the relationships between observer role and data collection methodology:

<table>
<thead>
<tr>
<th>TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES</th>
</tr>
</thead>
</table>

Lutz and Iannaccone (1969 modified)

Their middle category, observer as participant, is a useful one; it would seem to be the appropriate label for the role played by Reed, Hutton and Bazalgette; it also seems typical of investigations where the sample is small and where most or all of the data collection is done by the individual researcher, as is the case, for example, with
Parrish. The number and placing of the asterisks on the grid above give a general picture of my data collection methods. My Phase Two interviews probably qualify as semi-structured interviews with observer as participant. Another feature revealed by the grid is that I expect to play both the role of observer as participant and as non-participant, in some cases both roles for the same data collection method.

A full description of the methodology and observer roles included in my research design appears in Chapters Five to Eight. At this stage it would at least seem that the design employs the kind of mutually corroborative and refining variety of data collection approaches that is often advocated by adherents of research paradigms from all along the continuum. One accepts that, as Becker (1958) reminds us, participant observation is 'something more than immersing oneself in data and "having insights"' (p.660); also that it is probably much wiser if the logical structure of quantitative research at least is kept in mind to give general warnings and direction to the qualitative observer (Lazarsfeld and Barton, 1955). The best way to keep the logic of quantitative research in mind is to use it alongside qualitative approaches when the particular focus warrants it.

4. Research Hypothesis 2 Formulated

Like Chapter Two above, this chapter has attempted to re-assess theoretical discussion and empirical research in search of evidence that a Research Hypothesis is at least a valid area of investigation. The studies surveyed here certainly reveal an extremely broad range of factors that are potential influences on the success or failure of the period of study spent by overseas students in a foreign culture. And the fact that so many of these factors are found to fulfil this
potential confirms me in the belief that the kind of learner profiles hypothesised in Chapter Two can actually predict C₂ outcomes. It is thus feasible to formulate Research Hypothesis 2 (RH₂) ie

| Detailed Learner Profiles | Prediction of C₂ Future |

In Chapter Four the means of appropriately profiling participants' TL level and potential will be discussed. In the longitudinal study (Chapters Five to Eight below) the profiling itself will be carried out and validated against subsequent real events. It will be only when these events are known, of course, that RH₂ can be finally tested.
CHAPTER FOUR

TARGET LANGUAGE EVALUATION: FROM THEORY TOWARDS PRACTICE
120.
1. **Introduction**

This chapter has the following main aims:

1. to describe the purposes of my TL evaluation procedures
2. to establish the theoretical status of the notion of validity
3. to examine the scope and status of the concepts of competence and performance
4. to investigate the theoretical implications of communicative testing
5. to re-examine key testing issues in the light of these implications
6. to specify sets of general evaluational requirements for the design and construct validation of performance tests
7. to posit the third general Research Hypothesis (RH₃), derived from the issues re-assessed and ideas suggested in the chapter.

It is clear that my study of the inter-relationships between TL learning and use and other variables has as a prerequisite suitable means of evaluating that learning and use. Since I am assuming the responsibility of designing my own TL elicitation and evaluation system, the theoretical rationale and practical implications discussed in this chapter are crucial to my thesis.
2. **Target Language Evaluation: functions and needs**

A reasonable starting point for my rationale for target language evaluation is to consider what, in the research context already established, my language tests and other elicitation devices need to be able to do. The required characteristics listed here may not yet be comprehensive or logically discrete but they should serve to focus the discussion and relate it to key general evaluation issues as they affect this research.

One implication of the model explored in Chapter One (see Figure 1.1) is that I need to know about the target language variable at the presage stage and to keep a check on it right through to the time when there may be a 'product' to compare with it, whether at the end of a period of language training or at some other significant developmental stage. The system of evaluation must, therefore, contain elements of repeatability and direct comparability.

The descriptive model also commits me to some investigation of language learning processes. It will thus be necessary to design into the empirical study means of exploring and evaluating what seems to be happening as the learners learn. The dynamic has to be sampled as well as the static.

The issues discussed and general hypotheses formulated in Chapters One to Three make it clear that whatever is discovered about the proficiency, processes and progress of language learning will need to be related to what is discovered about what the learners are like and about what else is happening to them. This has to be taken account of in the design of the TL assessment system; it is likely to lead to
a search for some measures that are quantitatively compatible with my cognitive/affective data, some where scope is given for data elicitation on specific linguistic and extra-linguistic fronts, and to the inclusion of more qualitative evidence from less controlled situations when the inter-relationships under scrutiny may emerge naturally. The evaluation is going to need to be comprehensive and flexible.

But of course, none of these more specific requirements absolve the researcher from the normal constraints of formative or summative evaluation. Data must still be collected by means that are as acceptable, economical, practicable, reliable and valid as possible. And the most important of these, the superordinate concept, is validity, both in the field of evaluation and in research as a whole, since research itself can in a sense be seen as a continuous process of validation, in pursuit of evidence that hypotheses are 'well-founded and fully applicable to the particular matter or circumstance' (Shorter Oxford English Dictionary 1972).

3. The Notion of Validity

There is considerable uniformity in the language testing literature as regards general definitions of validity. It is all about whether a test measures 'what it claims to measure' (Lado 1961 p.321); 'what it purports to measure' (Pimsleur 1966 p.181); 'the extent to which it does what it is intended to do' (Pilliner 1968 p.30); 'when a test measures what it is supposed to measure and nothing else' (Ingram 1977 p.18); 'what a test actually measures in relation to what it is supposed to measure' (Oiler 1979;1 p.4). But, as Davies (1977;1) points out, there is more to it than this. What is common to all the definitions quoted
(and what is common sense) is that they all relate tests to their pre-determined purposes. 'The only problem here', says Davies (p.58) 'is that it assumes that the exact purpose of a test can be known'.

Certainly such definitions do seem to make assumptions that omit a major part of the validation problem; Morrow (1979 p.146/7) seems to pinpoint this when he suggests that validity is a matter of testers asking 'themselves whether they are actually testing what they think they are testing, and whether what they think they are testing is what they ought to be testing' (my emphasis). There are historical and methodological reasons for both the similarities and differences evident in these definitions of validity and they are germane to the present research.

Validity definitions such as Lado's, Pimsleur's or Pilliner's reflect, perhaps, the approach and ethos of their era, neatly labelled by Spolsky (1975) as 'psychometric-structuralist'. Here, insights from associationist learning theory, structuralist linguistics, contrastive analysis and psychometrics converge. 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 $L_1$ and thus more difficult, are 'drilled in' until $L_1$ 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. A knowledge of a language was a knowledge of the sum of its parts. The tone of confidence in the perceived neatness of the guiding paradigm of the time is sounded by Lado (1961):
"With the use of linguistic analysis and comparison of languages
we are able to locate and describe the significant elements that
will be most troublesome to a particular group of students. We are
thus able to discuss content validity on more solid ground than
previously" (p.322/3).

Now if you are sure of your theoretical framework and if that framework
seems easily translatable into methods and tests, there is less likely
to be so much soul-searching over the match between theory and tests.
This is probably why the issue of validity seems to be less focal and
to be discussed at a less theoretical level in the testing literature
of the 60s when compared with the late 70s onwards. And the real loser
in the discussion is construct validity, presumably because of the
confident theoretical assumptions that the psychometric-structuralists
seemed able to make. Neither Lado nor Valette (1967) mention construct
validity at all and they are fairly typical of their time in not going
beyond the concept of content validity, discussed as one of a list
of types of validity often linked, as they are by Davies (1968 p.10),
with particular types of test use:

<table>
<thead>
<tr>
<th>Use</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aptitude</td>
<td>Construct</td>
</tr>
<tr>
<td>Achievement</td>
<td>Content</td>
</tr>
<tr>
<td>Diagnosis</td>
<td>Content</td>
</tr>
<tr>
<td>Proficiency</td>
<td>Concurrent</td>
</tr>
<tr>
<td>Proficiency</td>
<td>Predictive</td>
</tr>
</tbody>
</table>

Figure 4.1: Relationships of likelihood between test use and type of
validity

But this kind of statement of loose relationships is inadequate,
especially when tests are being developed for research purposes with
the need to investigate matters 'from scratch'. Like Cronbach's 1961
table (p.106) summarising 'types of validity' in terms of the functions,

1. An informal 'reference count' comparing the attention given to
validity in Lado (1961) and Valette (1967) with the treatment of the
same subject in Allen and Davies (eds) (1977) and Oller (1979;1) reveals
a very significantly stronger emphasis in the later works.
procedures, typical use and exemplification of four validation exercises (predictive, concurrent, content and construct), it is a useful check-list for the post facto investigation of test results rather than a source of guidance to the test constructor. Again, there is insufficient indication of the primacy of construct validity. If it is about asking (as the last question on Cronbach's list, not the first) how scores on a test can 'be explained psychologically', it is clearly concerned more with the interpretation of test behaviour than with what should be elicited in the first place.

Davies (1977;1 p.64) does provide, in his 'schema for education' a logical sequence in which reference to language (learning) theory and validation processes are incorporated in a possible hypothesis-testing model:

\[\text{Figure 4.2: Schema for education}\]

Construct validity finally assumes its rightful place. As Davies says:
"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 that construct that validity, of the content and predictive kinds is based" (p.63).

'Proficiency' is seen by Davies as more than a label for tests to determine whether a student's 'language ability corresponds to specific language requirements' (Valette 1977 p.6). Proficiency considerations are a design influence, involving 'an assessment ...... of just 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' (Davies 1977;1 p.62). This latter definition reveals quite a lot about Davies' own construct, a construct that will be returned to below, but his useful chronology of validating events needs to be reinforced, for the purposes of the present theoretical discussion, by my schema in Figure 4.3, which further underlines the hierarchical nature of validity categories:

Figure 4.3: A logical hierarchy of validity categories (1)

The tree is hierarchical in the sense that the relationships between levels are entailment relationships. You cannot validate the content of a test without reference to the theoretical construct. Once test content is specified, it is validated for face, that is in terms of
how it appears (and appeals) to the candidate, the interested institutions, the administrators and so on. The content of a test is also validated ecologically as a check against the intrusion of irrelevantly disruptive factors into the testing context. Neither of these categories are trivial as will be indicated in further discussion below, when the nature of the validity concepts rather than their hierarchical status comes under closer scrutiny. Concurrent and predictive validity, in circular cells to distinguish their predominantly quantitative character from the primarily qualitative essence of the other categories, involve validating the construct, content, face and ecology of a test against outside criteria. These may be a relevant, already validated existing test or set of language behaviours or, in the case of predictive validity, performance in some future test or situation. Cronbach (1961) states that 'construct validation is much more complex than other forms of validation' (p.105). But he would appear to be comparing a superordinate with its hyponyms. Construct validity subsumes 'the other forms of validation'. The aim of these other validating processes is to ensure the best possible renewal of connection between theory and practice. Cronbach gives a more accurate picture when he claims that 'construct validation is established through a long-continued interplay between observation, reasoning and imagination' (p.121). This statement captures the combined qualitative and quantitative dimensions of validity and underlines the need for both deductive and inductive validation processes.

4. Communicative Competence and Performance: scope and status

The statements of rationale and the exploration of precedents contained in the first three chapters indicate strongly that the construct shaping
my means of eliciting and evaluating language data will derive from a broad view of competence and performance. My research is, by definition, concerned with the question of communicative competence; the discussion of it here arises, therefore, from the demands of my construct. Two quotations from Chomsky offer support for my need to take the broad view (as well as being a reminder that his acceptance of a structuralist view of the scope of linguistics was considered and not, as one might assume from some of the comments of his detractors, an error of omission):

"To study actual linguistic performance, we must consider the interaction of a variety of factors, of which the underlying competence of the speaker-hearer is only one. In this respect, study of language is no different from the investigation of other complex phenomena" (1965 p.4).

"Knowledge of one's language is not reflected directly in linguistic habits and dispositions, and it is clear that speakers of the same language or dialect may differ enormously in dispositions to verbal response, depending on personality, beliefs and countless other extra-linguistic factors" (1970 p.10 and see Chapter Two, Section 2 above).

Now, the broad-scope interpretation of the competence concept required by my research paradigm means that I shall be informed by insights from most forms of linguistic analysis. The following quotations, from speech act theory, sociolinguistic, socio-semantic, psycholinguistic and discourse analysis sources indicate the relevance of such sources and their commonality of concern:

"I think that it is essential to any specimen of linguistic communication that it involves a linguistic act. It is not, as has generally been supposed, the symbol or word or sentence, or even the token of the symbol or word or sentence, which is the unit of linguistic communication, but rather it is production of the token in the performance of the speech act that constitutes the basic unit of linguistic communication" (Searle 1965 p.136/7).

"Verbal interaction is a social process in which utterances are selected in accordance with socially recognised norms and expectations. It follows that linguistic phenomena are analysable..."
both within the context of language itself and within the broader context of social behaviour" (Gumperz 1968 p.219).

"Although generative grammarians, in particular Chomsky, claim that their work is an attempt to characterize the nature of competence (that is, the nature of those human abilities that are specific to language), their main effort has in fact been directed towards a far more restricted competence, which we will call competence, from which by far the most important linguistic ability has been omitted - the ability to produce or understand utterances which are not so much grammatical but, more important, appropriate to the context in which they are made" (Campbell and Wales 1970 p.247).

"The particular form taken by the grammatical system of language is closely related to the social and personal needs that language is required to serve. But in order to bring this out it is necessary to look at both the system of language and its functions at the same time" (Halliday 1970;1 p.142).

"We break irrevocably with the model that restricts the design of language to one face towards referential meaning, one toward sound, and that defines organisation of language as solely consisting of rules for linking the two. Such a model implies naming to be the sole use of speech, as if languages were never organised to lament, rejoice, beseech, admonish, aphorise, inveigh (Burke 1966 p.13), for the many varied forms of persuasion, direction, expression and symbolic play. A model of language must design it with a face toward communicative conduct and social life" (Hymes 1970 p.15).

"I have suggested that a distinction might be made between language usage and language use. The first of these is the citation of words and sentences as manifestations of the language system and the second is the way the system is realized for normal communication purposes" (Widdowson 1978 p.18). (All emphases mine).

It is certainly the broader, sociolinguistic model of communication that will inform my language evaluation procedures. And there are clear design and validation implications in the statements cited above. My empirical study must sample abilities and processes at work as the target learners attempt to handle the formal and the functional dimensions of speech acts related to their social and personal needs in various situations.

But it is not only the question of the scope of the competence concept that is relevant to my construct. The vital and complex issue of the status of competence and performance must be re-assessed if...
evaluation framework is to be comprehensive and coherent. The 'problem' starts from Chomsky's apparently inconsistent use of the competence concept. Is it what his 'weak' version suggests, 'the speaker-hearer's knowledge of his language' (Chomsky 1965 p.4) (and he makes it clear that he means the grammatical knowledge that can assign structural descriptions to sentences)? Or is it something stronger, 'concerned with discovering a mental reality underlying actual behaviour' (also 1965 p.4). And the problem deepens, inevitably, as the status of the concept of performance is put in doubt. Commenting on Chomsky's (1965) clarification of the concept, Hymes (1970) asks:

"When one speaks of performance, then, does one mean the behavioral data of speech? or all that underlies speech beyond the grammatical? or both? If the ambiguity is intentional, it is not fruitful; it smacks more of the residual category and marginal interest" (p.18).

Halliday (1970;1) dismisses the distinction between 'an idealized knowledge of a language and its actualized use: between 'the code' and 'the use of the code, or between 'competence' and 'performance' as 'unnecessary or misleading' (p.145) but posits, in his concept of meaning potential, something 'which is what I understand by Hymes' communicative competence' (Halliday 1970;2 p.9). Halliday goes on to state his own position on both the scope and status issues very unambiguously:

".... I am interested in what the speaker (or hearer) can do, not in what he knows; and while the two are, to a certain extent, different ways of looking at the same thing, the ethnographic perspective does have somewhat different implications from the psychological one" (Halliday 1970;2 p.9).

In this research, however, I am interested in evidence both of what my speaker/hearers can do and in what they know to the extent that my research design has both an ethnographic and a psychological perspective.
In such a context, it seems appropriate to explore an analogy somewhat cryptically referred to by Hymes (1970) but not, as far as I am aware, explicitly pursued by him:

"There seems, indeed, to have been some unconscious shifting between the sense in which one would speak of the performance of a motor, and that in which one would speak of the performance of a person or actor (cf Goffman 1959 pp.17-76), 'Performances'" (Hymes 1970 p.20).

An exploration of these analogies provides insights for the tester/evaluator at the present theoretical level of discussion and subsequently, when the focus is on more practical questions. So, the analogies, summarised to explicate the shifting of senses noted by Hymes and for their implications for the evaluator:

The car, model X

**Performance 0:** the capacities and capabilities of model X as specified in its manual; its built-in, intrinsic, potential to meet the requirements it was designed for and/or will be expected to meet; the specification of all the cars belonging to generic group, Model X, eg all Mini Metros.

**Performance 1:** the capacities and capabilities of an exemplar of model X as evidenced in various actual uses, summarised and interpreted, probably with at least implicit reference to 'performance 0', eg 'My Metro has been economic, nifty and reliable. The Metro is a good car'.

**Performance 2:** capacities and capabilities of an exemplar of model X as evidenced in a particular instance, perhaps on a test drive, summarised and interpreted, again probably with implicit reference to 'performance 0' and 'performance 1', eg 'It went well. Seems economic and nifty. The Metro seems a good car'.

We need Hymes' second analogy, too, because the connotational equivalent to 'performance 0' is **competence**, because of its animacy and because it covers generalisation from a sample to a corpus as well as from a sample to a population. Goffman himself states:

"I have been using the term 'performance' to refer to all the
activity of an individual which occurs during a period marked by his continuous presence before a particular set of observers and which has some influence on the observers" (Goffman 1959 p.32).

(It is clear, incidentally, that he is using the term 'observer' to include both participating and non-participating members of a group.)

So:

The performer, Olivier

**Competence:** the capacities and capabilities of individual X as known to observers (and to himself); his innate and acquired potential to meet the requirements he has learnt and/or will be expected to meet ie Olivier the actor.

**Performance 1:** the capacities and capabilities of individual X as evidenced in various actual performances, summarised and interpreted, probably with at least implicit reference to competence, eg 'Olivier's Othello is powerful, creative and unorthodox. He's a great actor'.

**Performance 2:** capacities and capabilities of individual X as evidenced in a particular performance on a particular day; summarised and interpreted, again probably with implicit reference to 'performance 1' and competence, eg 'Olivier's Othello was superb last night. He seems to keep on surprising you. A great actor'.

There is no equivalent to 'performance 0' here. We do not normally use the term 'performance' in its underlying sense of 'potential' when we are referring to individuals, as we are able to do with mechanisms designed to particular specifications. In this research, it is not intended to operate at the level of psychological 'performance models' where 'the mechanism behind the behaviour' (Sutherland 1966 p.155) is a legitimate focus. It is the influence from such models, however, that, as Sutherland, for example, is pointing out in his attack on Fodor and Garrett's apparent failure (1966) to make a clear distinction between competence and mechanism, that can lead to the 'sense shifting'
that worried Rymes. For my own evaluation and profiling purposes the three levels described in the two analogies will be theoretically and practically relevant. The labels 'competence', 'performance 1' and 'performance 2' will be used to refer to them when the two analogies inform the theoretical and practical discussion of key aspects of evaluation below.

The important collection Individual Differences in Language Ability and Language Behaviour (Fillmore, Kempler and Wang 1979) contains ideas which, as the title of the volume suggests they should be, are focally relevant to the competence/performance issue and thus to the construct validity of my language evaluation approaches. J B Carroll, for example, has reservations about a notion of competence that 'assumes no variation among speakers' and calls for a broader, more heterogeneous concept:

"It seems necessary to extend the notion of competence to describe a whole range of competences (with emphasis on the plural), not only those having to do with implicit knowledge of language rules, but also those having to do with the characteristic abilities of speakers (or writers) to use their linguistic knowledge to produce effective communications, to retrieve particular types of linguistic knowledge when called for, or to adapt their speech or writing styles to the demands of different occasions" (p.15).

Carroll's extended notion and its emphatic pluralism lend support to research such as my own study that attempts to evaluate linguistic competence in conjunction with other 'characteristic abilities'. As might be expected (and as we know from Chapter Two, Section 3.2), Carroll goes on to suggest the desirability of 'considering the 'various kinds of language abilities' from performances involving the 'observation of speakers interacting in communication situations'' (p.23). The route to any evaluation or profile of competence does
indeed seem to be through performances 1 and 2, which makes Fillmore's 1979 suggestion controversial:

".... the distinction between competence and performance may not be as important for the understanding of language behaviour as some scholars have considered it to be .... In a situation in which language use plays an essential role in a speaker's engagement in a matrix of human actions .... the distinction seems not to be particularly helpful" (p.91).

While it is clear that, particularly in the more formulaic instances of communication the difference between 'knowing that' and 'knowing how' can become very fine, it is a difference that certainly seems to reveal itself in most human activity, very often, as foreign language teachers (or car owners or theatre-goers) would agree, for reasons that are important to find out about. For the evaluator concerned with individual differences in communicative effectiveness, the investigation of the respective influences of factors of knowledge and use is essential. Hymes (1979;2) would like to see individual differences granted 'foundational status' in linguistics because of their value 'as a vantage point from which to consider questions of method and theory in the study of language in general' (p.36). In this cause, he recommends the use of the term 'competence' in its 'normal and natural' sense:

"It seems to me desirable to reinstitute the term 'competence' in the study of individual differences. What a person is able or not able to do, after all, is at the heart of much interest in the subject ...." (Hymes 1979;2 p.41).

In the same paper he is insistent on the varied and sensitive sampling of 'the realization of competence' (p.43):

"It should be obvious that personal competence in language cannot be assessed by fit or lack of fit to a preconceived model, or by tests in a single type of situation." (p.41) (my emphasis).
There are key questions to be answered:

"Where and what are the satisfactions possible through uses of language? What factors encourage, what discourage them? What configurations of individual ability and skill are fostered, what frustrated? Where are there abilitis that lack occasion? Where occasions to which ability can not or does not arise?" (p.43).

In the importance granted to such questions (indeed, in the appearance at this time of a multi-disciplinary collection conceived as 'an argument for locating concern with individual differences at the center of the study of language' (Fillmore et al. 1979 p.xiii)) I find support for the validity of the broad, multi-dimensional construct against which the evaluative procedures of my empirical research are carried out. And the clear implication is that competence, performance 1 and performance 2 all need to be investigated.

In the complex, and sometimes contradictory discussions of the notions of competence and performance, then, a warning is sounded of the dangers of unwarranted assumptions. If the notions are problematic for descriptive linguists, there is a strong likelihood that they will also be for evaluators. My two extended analogies may be of recurrent help in this area.

5. Communicative Testing? Some mainly theoretical issues

It is natural that the 'psycholinguistic-sociolinguistic' era in language testing (to use Spolsky's 1975 label for the phase that seems to have followed the psychometric-structuralist period) should be characterised by a concern with the evaluation of communicative competence. Paradigm shifts, however, are quite frequently the cue for rushes to fresh judgments and there is some evidence of this as language testers attempt to come to terms with the implications of the
communicative construct. The title of B J Carroll's 1980 study is *Testing Communicative Performance*; and Morrow (1979) has, as the first sub-heading for his discussion of the 'promised land' of communicative language testing: 'Performance Tests' (p.151). Now certain important points of logic need to be tackled here if my own attempts to test and evaluate how well various people communicate are to be soundly based.

Can you really test performance? If performance is actual behaviour, 'the behavioural data of speech' (Hymes 1970 p.18), and a test is 'that by which the existence or genuineness of anything is or may be determined' (Shorter Oxford English Dictionary 1972), or a 'systematic procedure for comparing the behaviour of two or more persons' (Cronbach 1961 p.21), the logical answer might seem to be 'no'. There are at least three reasons for this. In the first place, if something is 'actual', by definition it already exists and thus needs no test to determine its existence. In the second, if it is 'actual', it is the whole real thing rather than a sample of the real thing and the function of sampling is inherent to the concept of testing; we test in order to be able to make inferences from our sample to our general population or corpus. And thirdly, 'systematic procedures' presuppose a certain artificiality by the very fact that they are 'set up'. In such circumstances any behaviour observed is unlikely to be 'actual', where 'actual' has its sense of full authenticity. Davies (1978), in his comprehensive survey of language testing history and constructs, concludes:

"Naturalism is a vulgar error; all education needs some measure of idealisation, and the search for authenticity in language teaching is chimerical. The linguist, the language teacher and the tester are all concerned with generalising from a language sample to the whole of the language" (p.225).
So if you cannot test performance, the key questions become: what can you test if it is not performance, and what can you do with performance if you cannot test it?

In answering both questions it is crucial for the researcher to remind himself constantly what purposes his evaluation procedures are supposed to be serving. Someone like Davies, for example, whose main concern has been with the development of economical, widely administrable, predictive tests of proficiency, is not looking for fine-grained diagnostic data to be used for detailed individual profiling and the investigation of presage/process relationships. He is thus more likely to favour the 'typical extension of structuralist language frameworks' which 'could accommodate the testing of the communicative skills through, for example, context' (Davies 1978 p.225). With this stance, the answer to the first question seems to be that you test linguistic competence but making 'sure that there are tests of context as well as grammar', which, since 'language is not divorced from communication and cannot be taught or tested separately' (p.225), means you can claim to be testing communicative competence. And what, with priorities such as Davies', do you do with performance? You are less concerned to tap it directly for evaluation purposes. But you certainly observe it with a keen eye for its potential generalisability and as part of your 'serious attempt to make the language provided in .... (test) items more realistic' (p.225). Davies' view represents an evolutionary approach to language testing. In a sense, perhaps, this is the 'exam view' rather than a 'test' view, one more likely to be espoused when the focus is on formal, public, accredited, global proficiency or achievement assessment. In spite of its label, the
English Proficiency Test Battery (or Davies Test) (Davies 1964, Davies and Alderson 1977) would, since it possesses all these characteristics, normally be called an exam rather than a test. However, as Pilliner (1968 pp.21-22), Davies himself (1977;1 pp.49-54) and B J Carroll (1980 p.83) remind us, the two terms are used inconsistently and sometimes interchangeably, though with 'test' the commoner choice.

Morrow (1977 and 1979) suggests that more revolutionary change is called for if language testing and evaluation are 'to take account of .... (communicative) developments in any systematic way' (Morrow 1979 p.143). It is in this context that Morrow discusses his 'performance tests' though with an immediate awareness of some of the problems involved:

"Asking the question, 'What can this candidate do?' clearly implies a performance-based test. The idea that performance (rather than competence) is a legitimate area of concern for tests is actually quite a novel one and poses a number of problems, chiefly in terms of extrapolation and assessment" (Morrow 1979 p.151).

Morrow is, in fact, avoiding the illogic of 'testing performance'; and his answer to the two key questions above is that you can test something that is 'performance-based' and that a model that enables you to create valid performance tasks will enable you to evaluate performance (both 1 and 2 in my terms). Rea (1978) attempts to outline such a model with her 'overview' for 'assessing language as communication' (p.62). Feeling, like Jakobovits (1972 p.6), that 'the language testing field represents the most reactionary wing in FL education', Rea says:

"In terms of our testing AIMS the major consideration would seem to be in the construction of a simulated communication task which closely resembles one which the testee would encounter in real life and one which makes realistic demands on him in terms of language performance behaviours" (Rea 1978 p.54).
Morrow and Rea are taking a stronger line than Davies, without trying to chase the chimera of full authenticity and in spite of the fact that, for Morrow, at least, with his developmental work for the Royal Society of Arts (1977), the concern is with exams as well as tests. The strong claim is that the evaluation of performance 2 is a requirement, not an option:

".... in language use the whole is bigger than its parts. No matter how sophisticated the analysis of the parts, no matter whether the parts are isolated in terms of structures, lexis or functions, it is implausible to derive hard data about actual language performance from tests of these parts alone .... The clear implication .... is that by and large it is performance tests which are of most value in a communicative context. The very real problems of extrapolation and assessment .... therefore have to be faced" (Morrow 1979 p.151/152).

Now, assessment procedures carried out primarily for hypothesis-testing research purposes will not normally be (and should not be) completely separable from those carried out in the interest of student placement through proficiency measures or for the measurement of learning progress on a teaching programme. My own language evaluation system certainly had to be designed with more than its research function in view. Thus I must re-assess the research : instruction relationship, starting from the way in which tests for different purposes are conventionally defined. Moller (1975) provides useful definitions:

"Achievement tests are based on a particular course of study or syllabus and are devised to find out how much of the language material that has been presented has in fact been mastered by the students. Diagnostic testing is 'in-depth' testing, and the test constructors will ensure that sufficient items are given on the selected features of the language to be tested so that the students' strengths and weaknesses can be ascertained .... A proficiency test is an external test, one which is imposed on a candidate from outside and does not necessarily arise from his immediate studies or daily occupation. It should relate to some future need to use the language in certain situations" (p.6).

Some of the more detailed, practical points made by Moller will be
returned to below; here the focus is on overlapping testing functions. My need to elicit 'in-depth' diagnostic data and to construct repeatable and comparable measures of proficiency have already been mentioned as prerequisite to my research. But I also have to find out about achievement, about learners' progress towards 'mastery' and, rooting my research tests even more firmly into a real-life context, is their place in an actual training programme (see Chapter 5 below and Rixon 1981) based on a communicative teaching approach. Given the inevitability of a backwash from evaluation to teaching and learning, it is obviously preferable if the wash carries things in the intended direction. If it does so, with empirical research procedures serving an immediate practical purpose as well as a developing theoretical one, there is a healthy research : training relationship which can reinforce construct validity. In his introduction to Savignon (1972), Jakobovits describes the dangers of a less healthy testing : teaching relationship, where, for example, the requirements of 'objective accountability' can:

"effectively counteract the best intentions of goodwill for many a teacher, and help tip the balance, in a decisive way, in favour of impersonality, of non-authenticity, of transactions based on role-prescribed patterns of behaviour in the classroom" (Savignon 1972 p.6).

It is interesting in the light of the discussion of the feasibility of communicative evaluation below, that Jakobovits appears to feel that such testing is more practical than the testing of linguistic competence. Teachers, Jakobovits suggests, should:

"insist on the validity and desirability of goal-oriented criterion measures based on practicl communicative performance rather than on theoretical (and, in my view, impractical) considerations of a linguistic sort as typified by the 'standard' language tests ...." (Savignon 1972 p.6).
Of course, neither Jakobovits nor Savignon had at that stage experienced the full, multi-variable force of the communicative competence concept on language learning or testing and Savignon's total separation of linguistic competence from communicative competence in both areas is dubious. Still, the message from her balanced and insightful study (discussed further below) is that the construct validating link from evaluation to language learning theory and practice should be close and strong:

"The research .... focuses on the development of tests of communicative competence suited to the beginning level of a college French programme and the use of these tests to measure the effectiveness of early training in communicative skills .... The most significant findings of this study point to the value of training in communicative skills from the very beginning of the FL programme and to the inadequacy of traditional tests of achievement in assessing communicative competence" (Savignon 1972 p.9).

If the communicative construct is accepted, the fundamental problem facing the researcher is the reconciliation of the realities of communication with the theoretical and practical requirements of testing and evaluation. In real life, communication is judged by communicative results. In Searle's terms, refining Grice (1957), it is a matter of a speaker/writer's intention:

"to produce a certain effect by means of getting the hearer (or reader) to recognise his intention to produce that effect and (by intending) this recognition to be achieved in virtue of the fact that rules for using the expressions he utters (or writes) associate the expressions with the production of that effect" (Searle 1965 p.145/6). (my parentheses).

All this is, of course, to be achieved in real time and in a real setting. It has become an accepted aim of communicative approaches to language teaching to help learners to develop their capacity to produce and recognise the illocutionary acts they intend to produce and recognise. Somewhere along the way towards this aim attempts are
made to create the interactive, unpredictable, physically and psycho-socially situated conditions of authentic, effect-producing intentions. It would, as we have seen, be illogical for the tester to think he can recreate fully authentic communicative conditions but the evaluator, looking to find out how well people are performing and why, needs to elicit data from 'situated language use' (Fillmore 1973 p.5). And the evaluator (especially as researcher) may take note of the fact that the crucial measure of effectiveness for any real communicator is his own evaluation of his performance. Holec (1980) recognises this.

Discussing 'auto-evaluation' he says:

".... the learner has to take part in real acts of communication, since this is the only way in which the match between results achieved and results aimed at can be confirmed .... Tests do not elicit performances which are valid enough (in terms of their authenticity) to be used in this ...." (Holec 1980 p.36).

There is a useful reminder here of the importance of the notion of self-evaluation of performance 2 (in the process of communicating) and of performance 1 and competence (before and after it). Both notions are explored by my own evaluation procedures in the empirical study itself.

6. Communicative Testing: from theory towards practice
As a prelude in the transition from aspects of a theoretical framework for evaluation to more practical evaluation issues, it may be appropriate to return to my two analogies. In many ways the role of the researcher can be compared with that of the motoring correspondent who is to test and evaluate a model in order to inform interested parties about it. Before he sets up his system for testing the car, he will certainly make sure he has the fullest available details of its performance specifications as well as any other information relevant to his purposes. This 'presage' information will then be considered
in relation to predict factors, the likely functions that the model will be expected to perform. If one of the model's key specifications, for instance, is its low in-town petrol consumption and an important potential consumer group is economy-minded city-dwellers, it may be that the motoring correspondent will set up a test drive in town and evaluate the car under those conditions, observing its performance (2), inferring possible processes during the programme he has decided on. After the test drive, it is fairly certain that he will measure and report the mpg achieved and compare it with the figure specified. Almost everything else he may or should do, however, is problematic. The extent to which he can generalise about town-driving petrol consumption, given that the particular performance will have been affected by his driving habits (and, perhaps, his driving on that particular day), the amount and behaviour of other traffic, the weather, the petrol quality and so on. And should he use the test drive only to evaluate petrol consumption or is it fair to evaluate the car's comfort, road-holding or even its 'general impression' at the same time? And what about long-term reliability? What kind of testing can predict for those who want to know tomorrow, the likely state of their Metro in five years time? At the least, our correspondent may be questioning the value (and the validity) of evaluating the model on the basis of one sampling of one performance (2) of one exemplar of the model. At the most he may be wondering whether anything has been gained by taking the car out into the real world, whether he may not have learnt more, more economically and conveniently, by having a good look at the blueprints, the assembly line, the boardroom and by seeing what the car looked, felt and sounded like in the showroom and/or on rollers. Such questions are familiar to language testers, too, and the retreat from the direct
evaluation of performance may be acceptable, as Davies seems to say, provided that relationships between data from competence testing and predicted behaviour have been established. As the car analogy is pushed further, it begins to be revealed as inadequate, but, as a worthwhile analogy should, helps make a point by its very inadequacy. Perhaps the main reasons why it does not seem wholly unacceptable to report on a car without a direct evaluation of its performance are, first that there is the minimum of variability between individual examples of the model and, second that there is a more predictable relationship between performance specifications and performance in use.

It would not, however, be conceivable for a drama critic to inform theatregoers about what to expect from actors in a play without seeing them in a performance of it. The human factor in this analogy immediately brings in a high degree of individual variability which, in interaction with the script and other production factors, will result in a performance that would have been very difficult to predict. Once he has done his homework on the cast, the script and the production, then seen the performance, the critic is in a fairly strong position to make an evaluation which will help its readers not only to decide whether to try the play or not but also to appreciate better the performance they see. But the critic, like the motoring correspondent, still faces dilemmas of focus, extrapolation and comparison. In their shared responsibilities of predicting and generalising, the two breeds of evaluators are aided by the crucial quality of consistency in contrasting elements of their task, the one by the relative invariability of particular examples of a car model, the other by the invariability of the script. The evaluator of people who have
to participate in real communication has neither his evaluatee nor his context of situation as constants. The best he can do is to find out as much as possible about their likely 'scripts' (the predict).

In language testing and evaluation terms the practical issues raised by the preceding theoretical discussion and exemplified through the analogies are the following: the relationship between validity and reliability; generalisability; criterion and norm referencing; the divisibility of competence and criteria for assessment. An examination of these issues will enable me to establish a set of general evaluational requirements as a framework for the construction, validation and assessment of all the elicitation devices I use to evaluate the communicative effectiveness of the members of my target group.

Davies' 1978 survey article contains a neat summary of what he calls the validity-reliability 'tension' (Davies 1978 p.223):

Now, the neatness of Davies' grid is deceptive but the discussion required to avoid its risk of over-simplification involves an instructive consideration of the key issues. Oller (1979;1) describes the two categories of test as follows:

"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 recognised skills or aspects of skills" (p.37)

So much for the pedigree of integrative tests. But how does he see their competence/performance status? An integrative or, for Oiler synonymously, a pragmatic test is:

".... any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language and which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extra-linguistic context" (p.38).

Oiler's integrative tests seem to resemble the communicative performance tasks indicated by Rea and Morrow above. If this is so, he differs from Davies since the latter decides whether a test is integrative or not by reference to the scale of its linguistic content and by the means employed to score performance on it. One's intuition is that the first criterion would take precedence so that, for example, a reading text of some paragraphs in length, but with discrete-item multiple choice questions would be considered integrative. But for Davies such a test is discrete point (DP) since, in terms of his 'patterns of relationship' between DP and integrative tests, only the stimulus (here, the reading text) is integrative; the tasks, items and scoring are DP. Davies himself acknowledges the definition problem:

".... we argue that the distinction between discrete point and integrative testing is not a real or an absolute one" (Davies 1978 p.155).

This underlines the need for caution in the interpretation of Davies' own grid. The match or lack of it between the task and the method of assessing performance on it must both be taken into account, not only in the cause of clearer definitions but, more important, in the interest of construct validity. In performance tasks such as note-taking,
scanning, information transfer, oral interaction and formal written communication (see Tests 1, 3, 4 and 5 in Chapter Six), there can be no question that the 'test' is integrative. Whether my use of analytic criteria to assess performance in such activities turns 'what starts off as an integrative test into a discrete point one' (Davies 1978 p.154) or, if it does, whether this affects the validity of my communicative construct, will be the subject for later discussion.

Davies' separation, on his grid, of linguistic and communicative competence also raises interesting practical questions. In the examination of the scope of the competence concept above, the former is seen as part of the latter (though Canale and Swain 1979 claim that it is 'equally common' to find the two terms used as if the two forms of competence were separate). Language teachers and testers have never been as able or willing to keep the two apart as Chomsky decided to. It is thus difficult to envisage (as the grid says we might) an integrative test that measures only linguistic competence; to accept that 'as the table shows, the most desirable test (having +V +R) is a DP test of communicative competence, the least desirable (-V -R) an integrative test of linguistic competence' (Davies 1978 p.223) is more than difficult. My own Test 2 (A) is in part a DP test of communicative competence (see Chapters Six and Eight) but these characteristics do not in themselves make it my most desirable evaluation instrument. Rather, it is included in the battery, as we shall see, to elicit particular kinds of data and to serve particular correlational purposes. Davies himself offers the sensible advice that issues such as linguistic/communicative competence, DP/integrative tests, reliability/validity relationships are best perceived as continua not dichotomies.
However, he and others reacting to the communicative paradigm shift in a less than radical way may be making unwarranted assumptions about the central issue, that of validity.

The idea of a 'tension' between two notions (Davies 1978 p.149) seems to imply that the two notions operate at the same level, that ground gained by one side in the tension (as tug-of-war) is ground lost by the other. Both these assumptions seem to be accepted if one is to judge by statements such as:

"All frameworks in tests are attempts to provide a method of scoring ie they (normally) increase reliability by the use of discrete point test items and thus decrease validity by not using integrative-type items" (Davies 1978 p.222).

But in my hierarchy of validity categories above, reliability (the issue of 'how consistently (a test) produces similar results on different occasions under similar circumstances' (Oiler 1979;1 p.4)) would operate at the same level as concurrent validity. Agreed, 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 such as those discussed in Chapters Six and Eight.

Given my communicative construct, I may rewrite my validation hierarchy as follows:
And given the communicative construct, the significance of both face and ecological validity needs re-examining. Lado (1961) regards face validation as 'a widely used way but a weak one' (p.321); Ingram (1977) 'as a public relations problem rather than a technical one' (p.18).

Morrow (1977), however, pinpoints the testee as the main focus for face validation:

"it must seem plausible to the person taking the test that the tasks he is asked to undertake are relevant to the objectives of the test" (p.16).

This is nearer the mark but does not yet reflect the full implications of candidate as participant in a performance task. If performance 2 is being evaluated, it becomes essential for the candidate to be able to pick up, from the sight or sound of the task, the cues that will involve him in the intended communicative activity. No such involvement is likely if the task lacks such cues. All the performance tasks in my main test batteries (see Tests 1, 3, 4 and 5 in Chapter Six)
attempt to supply the cues in the form of task authenticity, with this term used in the sense suggested by Widdowson (1978) with the learner 'required to deal with' 'genuine instances of language' in a way that corresponds to 'his normal communicative activities' (p.80). Now this interpretation of face validity is a strong one. If the term also has to cover 'public relations' matters like whether other interested parties not actually taking the test are satisfied by the look of it, perhaps we are expecting too much of one label. Hence the subdivision in the rewritten hierarchy above.

If face validity is achieved by designing into the test features that will help to involve the participant, ecological validation is a matter of excluding features that are likely to deter or handicap him irrelevantly. Although my empirical research is not primarily cross-cultural, the heterogeneity of my group makes insights from that area of ethnography pertinent. Cole et al. (1971) feel that the relationship between experiment (or test) and process must be finely judged since:

".... differences in cognition reside more in the situations to which particular processes are applied than in the existence of a process in one cultural group and its absence in another" (p.233).

In part this is a methodological argument. 'Ethnographers', says Cole, 'reject experiments as artificial while psychologists avoid natural behaviour as ambiguous' (p.474). And Neisser (1976 p.33), among others, has noted how too much experimental purity may in itself reduce the ecological validity of an elicitation device. Not that those on the psychometric side are unaware of such problems; J B Carroll (1974) describes the question of 'task sets' as 'one of the most unstudied problems in psychology' and quotes Newell (1973) on its importance in testing:
"The interaction of the instructions with the task performance programme is as much central to control as the internal part of the performance programme" (Carroll 1974 p.12).

When unresolved but crucial questions like this are set alongside ethnographic evidence such as that which led Cole (1975) to conclude that 'familiarity with the materials about which one is asked to reason is important if people are going to apply a cognitive skill they have' (p.476) it is not surprising that ecological validation is difficult, especially with a mixed-background population. Farhady (1979) reminds us of the kind of problem it poses for the language tester when he cites evidence of the way in which learners from different countries perform better or worse according to whether they are tested with integrative or discrete point techniques. Clearly not enough is yet known about these matters for ecological validity to be firmly established at the test or task design stage. Still, I have a good chance, given my fairly full profiles of the presage of the individuals in my group, to identify ecological problems and indeed to learn more about the individuals concerned from the problems as they emerge from test or task performance. Of course, problems emphasised by socio-cultural background but which have been predicted as likely to be faced as part of the foreign culture experience may not be excluded. As was indicated at the beginning of this discussion, ecological validity depends on the exclusion of features that may irrelevantly deter or handicap.

Qualitative, though systematic, validation has precedence at the test and task construction stage over the more quantitative validation procedures. Precedence, for example, over concurrent validation against an existing test, which is normally suggested at pilot-testing
stage. 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. In my own case, the development of a new, communicatively-informed test battery was informed by an analysis with the 1979 ODA group of performance on existing test instruments even though this analysis produced fairly negative results in terms of construct validity. More on this issue in Chapter Six. More significant comparisons of performance and judgments on validity are possible when feedback on the actual communicative performance of the group in their real-life contexts becomes available (as in Phase Two of my study). Then we are able to assess the predictive validity of our evaluations and to consider adjustments at all levels of the validity hierarchy including our actual construct.

But the issues of concurrent validity and reliability, with their essential concerns of consistency and repeatability, raise important questions about the potential lack of generalisability of certain kinds of communicative tests. If the construct implies the specification of performance tasks closely related to the learners' predicted communicative activities, two key points about generalisability immediately arise.

Firstly, will a task recreating conditions specific to a communicative
activity be susceptible to checks for reliability through readministra-
tion or will the very specificity of the task render it unrepeatable? Or, will not a performance task, the design of which is true to its
construct in aiming at communicative validity, lack generalisability
in terms of other communicative activities? Are such forms of
evaluation in other words 'only' evaluations of actual utterances
(performance 2) under explicit and implicit linguistic and social
constraints, thus not easily related to competence as 'characteristic
abilities' (J B Carroll 1979) or 'ability for use' (Hymes 1970 p.20)? My way of answering these questions is through the general
evaluational requirements proposed below which ensure tests and tasks
are amenable to the processes of communicative validation including
those needed to create and check generalisability. But other evaluators
have pointed to the potential generalisability of certain levels of
communicative activity that can be triggered even by the most specific
of tasks. Rea (1978) identifies 'higher order' skills such as 'inter-
pretative comprehension', 'self-expression', 'analysis and evaluation'
and 'synthesis' (p.57). B J Carroll (1978 , 1980) and Morrow (1979)
examine Munby's (1978) 'micro-skills', re-christening some of them
'enabling skills'. These, Morrow says, have an application that
'extends far beyond any one particular instance of performance and in
this creativity they reflect an aspect of what is generally understood
by competence ' (p.153). My own evaluation system certainly taps the
potential of such processes in a variety of different guises and
attempts to discover how well they can be reconciled with the demands
of reliability and scorability, areas in which according to Morrow
(1978) 'there are no published data' (p.154).
Another clear implication of Davies' supposed validity-reliability tension and of much of the foregoing theoretical discussion is the question of criterion-referenced testing. The concept is neither really new nor logically discrete from norm-referenced testing but it has come explicitly to the fore with the individualisation of instruction, increased emphasis on curriculum evaluation and disillusionment with or distaste for rank-ordering psychometrics.

Popham (1978) provides the following functional definitions:

"... 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 (norm-referenced) tests, a good criterion-referenced test yields a better picture of just what it is that the examinee can or cannot do" (p.2).

Davies (1978) makes the connection with language testing and, unsurprisingly given his wary view of performance tasks, has reservations:

"Ideally, criterion referenced tests have the following characteristics: they test externally defined objectives, they test on a syllabus or content rather than on a rank order, they are useful diagnostically, and they test all relevant behaviour, not just samples of it (Bormuth 1970). Naturally, 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" (p.158).

But I need the descriptive, diagnostic power of criterion-referenced tests in my research; my investigation of inter-relationships between linguistic and other variables requires detailed and comprehensive profiles of the kind produced at the end of Phase One, the initial six-week period with my group. The lack of a 'finite inventory of learning points' is not a problem unique to criterion-referencing and is handled by predict and presage analysis (see Chapter Two, Figure 2.6). The
'very many behavioural objectives' have to be taken on board, in fact, it is a key aim of this kind of exploratory research to identify them (from the kind of sources discussed in Chapters One to Three as well as here) then to check and relate them empirically. 'External criteria of success' abound as my dialogues with tutors, administrators and my group members reveal. The framework of general evaluational requirements proposed below suggests how the relationships between features of a performance task, the particular abilities required to handle it and the criteria to be used to assess these abilities may be established and validated. Further relevant description and analysis appear in the report of the empirical research in Chapters Five to Eight.

Much of the discussion of norm- and criterion-referenced tests centres on the quantitative implications, especially on how different things are if you take the latter approach. But Davies (1978) reminds us of the common ground, sensibly I find on the basis of my own empirical investigation: 'For every criterion referenced test,' he says (p.158) 'there must be a population for whom the test could be norm-referenced'. Yes, or, one could also say, for every norm reference, there must be a criterion against which that norm was established. And in practical test construction and validation terms it is not that criterion-referenced tests preclude the analysis of variance in results, eliminate the need for item analysis through discrimination indices or eschew rank ordering. It is rather that they will aim as their priority to reflect accurately the criterion behaviour. Their a priori validation (Popham 1978 p.104) may take precedence over a posteriori validity analysis but cannot replace it. Popham and Husek (1969, quoted in B J Carroll 1980) seem to get the balance right:
"If a criterion-referenced test has a high average inter-item correlation, this is fine. If the test has a high test-retest correlation, that is also fine .... The point is that a criterion-referenced test could be highly consistent, either internally or temporarily, and yet indices dependent on variability (in its statistical sense) might not reflect that consistency .... A carefully made judgment, based on the test's apparent relevance to the behaviours legitimately inferable from those delimited by the criterion, is the general procedure for validating criterion-referenced measures" (p.103). (My parentheses).

Obviously you do not find out whether such correlations or consistencies 'are fine' or not without checking them and interpreting the evidence, though all the time with construct validity as your guiding priority. Chapters Six and Eight show how I found both qualitative and quantitative validating procedures essential. They also reveal the importance of the use of rank order within my sample and of some reference to external norm-referenced measurements. General support for the use of criterion-referenced tests in research like mine is inferred from Davies' indication of their suitability with very small samples and to establish 'what is adequate proficiency' (Davies 1978 p.158).

Since I am looking into the implications of the validity-reliability tension for any hint that communicative competence may ultimately be too problematic to test, it is interesting to come across a speculative paper by Oller (1979;2) which, starting from the question "Communicative Competence: Can it be Tested?" actually investigates the antithetical query: "Communicative Competence: Can Anything Else be Tested?". And the reason why he dares to ask the latter question is evidence from research that 'the best tests were always strongly correlated with each other and that this inter-relationship was somewhat independent of the modality of the test or of the components or
skills the tests were aimed at' (p.7). Now Oller's 1979 paper represents the fullest extension of previous investigations into the discrete competence, the unitary competence and the partial divisibility hypotheses, DCH, UCH and PDH (see, for example, Oller 1975, Oller 1976, Oller and Perkins (eds) 1978) in which the preponderance of evidence is found to be in favour of the UCH. Other studies of these and related hypotheses produce less than consistent findings, however. Upshur (1975) considers the UCH inadequate, since L2 proficiency is a matter of integrative ability and discrete skills. Palmer and Bachman (1980) claim strong pro-DCH evidence in their study of oral proficiency and feel that Oller's findings in favour of indivisibility are biased by his use of Principal Component Analysis, a factor analysis technique that tends to produce inflated measures of common variance. Vollmer (1979) thinks that existing evidence on these hypotheses is inconclusive and suggests the usefulness of a 'hierarchical skills theory' (p.10).

Some of the work on the divisibility question concentrates on its relevance to the assessment of performance. Yorozuya and Oller (1980) find that careful overall assessment of oral performances is just as good as an assessment on a series of scales, mainly because of the halo effect on raters, that is the influence of a rating given on one criterion (eg 'grammar') on another (eg 'fluency'). Callaway (1980) considers that, since raters tend to make 'holistic, unidimensional evaluations' anyway, 'dividing oral performance into components is superfluous at best, and artifactual at worst' (p.111). Yet Mullen (1977) finds that each of the separate scales used to rate writing proficiency 'plays a role in the determination of the overall
The language evaluation reported on in Chapters Six and Eight is not aimed directly at testing divisibility hypotheses. But my approaches have been informed by the discussions and studies cited here and my empirical research will inevitably be to some extent interpretable in UCH, DCH, PDH terms. With the multi-dimensional competence construct that I justify in this chapter and my need to sample a variety of performances, I am bound to evaluate in a variety of ways and at a variety of levels. I also require analytic evaluation scales to help to systematise the connections between performance and assessment criteria. If this is a DCH-biased approach, its bias is open to falsification - and on more than the basis of statistical correlation, which is, perhaps, too exclusively the cornerstone of Oller's investigations.

The whole divisibility question clearly takes us back to Davies' DP/integrative discussion. It is interesting to note that when testers accepting the communicative construct come to practical grips with its implications, they tend to emphasise either an analytic system for specifying the task, or their analytic criteria for assessing it. And even when fairly equal attention is given to both, the way that the two are related is not always very clear. Savignon (1972), for example, concentrates on the role and attitude of the interviewers in her 'Discussion' and 'Interview' tests and suggests interesting evaluation scales such as:

<table>
<thead>
<tr>
<th>Effort to communicate</th>
<th>None</th>
<th></th>
<th>Great</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amount of communication</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Savignon does not, however, refer to any framework that might aid validation of the tasks themselves. On this she is fairly vague, the assumption perhaps being that native tester intuition will suffice.

"The first task was designed as an informal interaction between the students and a native speaker of French who also knew English. The object of the exercise was to see how much information the student and the native speaker could exchange on an assigned topic in the four minutes allowed. The administrator began the discussion by introducing the student and the native speaker and giving them the subject of the discussion. Three topics were randomly assigned: 1) the advantages of a large university versus a small college 2) the validity of FL requirements in a liberal arts education 3) the role of students in university administration ...." (Savignon 1972 p.42).

Similarly, Schulz (1977) in an interesting experiment on the effect of DP and 'simulated communication testing' on foreign language learning leaves the definition of performance tasks fairly open:

"A simulated communication test was defined as any test item which required the student to make a spontaneous (uncued) response (written, oral or kinetic) in a specified situationally realistic setting" (p.94).

Morrow (1977), who, like B J Carroll (1978 and 1980) develops a strong set of task construction and validation parameters from Munby (1978), seems less than certain of the practical influence of such parameters (discussed below) or of analytic evaluation criteria. Describing Palmer's (1972) 'Comtests', which involve candidate and tester in a dialogue to find out which of a set of marginally differing pictures
the examiner has selected, Morrow comments that 'this is clearly an authentic situation' (p.39). One suspects that if this task were checked in terms of parameters such as his 'function', 'role', 'status' etc, its 'authenticity' might be questioned. Morrow's ideas on assessment criteria are examined below; his 1979 comment reveals that they are tentative:

"It would be possible to use an analytic system whereby candidates' performance was marked in terms of each of the criteria in turn and these were then totalled to give a score. More attractive (to me at least) is a scheme whereby an overall impression mark is given with the marker instructed simply to base his impression on the specified criteria. Which of these will work better in practice remains to be seen ...." (p.154).

Studies such as Savignon (1972) and Palmer (1972) also raise again in their practical context the key questions of the directness and authenticity of performance tasks. My notions of performance 1 and performance 2 assume direct sampling of actual behaviour. But the motoring correspondent, it will be remembered, considered the alternative of evaluating without actually tapping performance 2. And in the language testing field Palmer's Comtest is communicative but does not recreate a situation likely to occur in real life; Oller's use of dictation and cloze tests as the best and quickest way of evaluating communicative competence is an advocacy of tasks that, in spite of his claims (e.g. Oller 1979;1), are improbable; and Morrow's interesting suggestion for DP communicative tests has parameters such as modality, pre-supposition, role, formality, mood used as foci for multiple-choice items on isolated remarks (pp.29-33), not requiring any productive activity from a candidate. One is reminded that in the field of psychology 'performance tests' are often designedly indirect:

"If an observation is to bring to light typical behaviour, the subject must not know what characteristic is being observed" (Cronbach 1961 p.32).
"Although all tests call for performance of some sort, the name performance test is usually applied to tests requiring a non-verbal response." (p.35).

(The Cattell and Witkin instruments for personality factors and cognitive style introduced in Chapter Two and investigated in Chapters Seven and Eight are both examples of indirect tests.)

My own evaluation procedures, while preponderantly communicative, integrative and direct because of my construct and, relatedly, the need for detailed profiling information, do not exclude DP or indirect assessment. Only my correlational and other forms of analysis will decide what kinds of evaluation procedures best serve which purposes. They may also help clarify the question of whether tests not involving individuals in active production (e.g., the grammatical, social survival and discourse function sub-tests of my Test 2) or tests calling for an apparently improbable activity (e.g., my cloze tests) can necessarily claim to be assessing competence.

7. General Evaluational Requirements: a framework for performance test construction and assessment

At this point, then, to my proposed sets of general evaluational requirements, my attempt to specify, taking insights from all the sources discussed above, what needs to be taken into account in the design and assessment of communicative tasks following the construct and serving the purposes of my empirical research.

Communicative testers who see the need for a set of parameters as a framework for the generalised description of language use normally derive them from speech act theory and sociolinguistic models, sometimes
via their adaptation by needs analysts. Hymes (1964), himself informed by Jakobson (1953 and 1960) suggests the following as 'general criteria for .... communicative status':

".... the concept of message implies the sharing (real or imputed) of (1) a code or codes in terms of which the message is intelligible to (2) participants, minimally an addressor and addressee .... in (3) an event constituted by its transmission and characterised by (4) a channel or channels, (5) a setting or context, (6) a definite form or shape to the message and (7) a topic and comment, ie that it says something about someone" (Hymes 1964 p.26).

TL needs analysts are quick to adopt contextual parameters such as setting, channel and addressor/addressee relationships; their sets of parameters, however, tend naturally enough to respecify 'code' and 'topic' in terms of 'functions', 'notions', 'skills' (Van Ek 1975, Munby 1978) and 'linguistic exponents' (Van Ek) or 'language realisations' (Munby). Hymes is concerned with criteria for ethnographic description, the needs analysts with syllabus definition. The communicative tester needs such parameters as a checklist for the design of appropriate performance tasks and to inform less direct competence tests.

Reservations about needs analysis models such as Munby's as course design instruments (eg Widdowson 1981) may not apply to the use of such models in testing. Like B J Carroll (1978 and 1980) I shall use Munby as my most immediate informing source. For my general evaluational requirement 1 I use the following set of parameters:

1) Participant (as subject and communicator in the test event)
2) Purpose (of the participant(s) in the event and of the event itself)
3) Activities (sub-tasks involved in achieving the purpose)
4) Setting (physical and psycho-social context of the event)
5) **Interaction** (role set and social relationships of participants)
6) **Instrumentality** (medium, mode, channel of communication for the event)
7) **Dialect** (dialect and accent of participants in the event)
8) **Code** (forms and functions conveying the messages of the event)
9) **Communicative Operations** (skills and strategies required to achieve the purposes of the event)

The test task constructor and validator will also need a systematic means of taking account of the more dynamic characteristics of communication, which I shall use as the label for my general requirement 2. The tester/evaluator with a commitment to the recreation of as many of the conditions of real communication as is feasible needs to be able to predict what happens when the parameters of communicative events trigger each other off. Now, there is no definitive description of such characteristics around but it is possible, from the speech act theory and sociolinguistic sources exemplified above and in Chapter Two Section 3.7, and from language acquisition research such as the kind also discussed in Chapter Two in connection with age and language learning, to derive a useful checklist.

The result is an inventory of characteristics that enables the evaluator to check a task described through his parameters (general requirement 1) for opportunities to get nearer to authentic performance 2. Thus my set of dynamic characteristics of communication, still making no claim that it is comprehensive or that there is no overlap between categories, is:
1) the existence of relevant information gaps (i.e., a universe of discourse from which the participant has to process not yet known information of a kind relevant to his real-life situation and fill a similarly relevant gap for his interlocutor)

2) inter-subjectivity (i.e., that the speaker/writer should want to know that the hearer/reader knows that the former knows what the latter knows ...)

3) authenticity of setting (i.e., that scene and setting are specified or inferable as appropriate to subject matter and participants)

4) unpredictability (i.e., that the task allows for the occurrence of activities that could not normally be foreseen from the initial given or shared knowledge)

5) creativity (i.e., that the task offers scope for participants to be 'novel', to assert their communicative independence as they might in a naturalistic event)

6) self-monitoring (i.e., that the task allows participants to use their discourse processing strategies to evaluate communicative effectiveness and make adjustments during the course of the event)

7) natural chunking (i.e., that the size and scope of task activities are such that participants are processing the kind of input segments that they would normally expect to)

8) real time (i.e., that the task has to be accomplished under normal time constraints)

The specification of these eight characteristics assumes that no significant allowance is being made in the design of tasks for the fact that those being evaluated are not native speakers. If such allowance were being made, for example by 'simplifying' text or other resource materials, then the list would have to be altered radically, if not discarded. As they stand, they are suitable for the qualitative validation of both mainly speaking and mainly writing tasks, though certain categories, inter-subjectivity, unpredictability and self-monitoring, for instance, would seem to play a more frequent and immediate role in oral interaction.
At this stage the evaluator is in a better position to select and describe a particular test event than to check it for communicative authenticity. My general requirement 3 is a framework for describing or predicting the dimensions of a particular event as they might be with native speaker participation. At one level, this provides a description in more objective linguistic/stylistic terms, of the target task. At another, it allows the tester, thinking ahead now to the evaluation of participants' performance in the event, to plot his performance evaluation criteria (fluency, accuracy etc) against the dimensions (size, complexity etc) inherent in the task itself. Munby (1978) has as one of the parameters of his needs analysis model a 'target level guide', a category used to specify communicative dimensions. In a testing context these are more relevant as objectives than as determinants, as means of describing an event as it might, or even should, be. However, I do not intend to use them in the way B J Carroll (1980) and Morrow (1980) do, as criteria for the specification of general levels of proficiency:

- **Size**: Can understand and take part effectively in an extended dialogue or group discussion" (Carroll 1980 p.137).
- **Basic Level Complexity**: Texts may be simple, showing little cohesion. Simple sentences with little attempt at cohesion are acceptable" (Morrow 1980 p.1).

For my test construction and validation purposes it is relevant to check the dimensions of an event in terms of:

1) **Size** (the amount of communication, receptive and/or productive, that is involved in the event)

2) **Complexity** (the degree of grammatical complexity, the range of cohesion devices likely to be required)
3) **Functional Range**
   (the degree of variety of illocutionary acts involved in the event)

4) **Referential Range**
   (the breadth and depth of lexical knowledge required to handle activities in the event)

This set of dimensions considerably modifies Munby's categories and definitions. I narrow down his 'complexity' to the syntactic/propositional level. When Munby (p.165) includes 'discourse coherence' under his complexity label, he is already into the area of the 'relationship' between 'illocutionary acts' (Widdowson 1978 p.28) and thus overlapping his own 'micro-functions' under 'range' and 'delicacy'.

Cohesion, in the delimited interpretation given to it by Widdowson (p.28) as 'the overt relationship between propositions expressed through sentences' is included under my complexity label. Coherence features belong here under 'functional range'. I drop Munby's 'delicacy' category for similar reasons. Whether it is concerned with 'specificity and detail' (p.165) or 'Delicacy of forms/micro-functions/micro-skills' (p.165), the latter, apparently referring to something akin to 'subtlety' (cf 'an endpoint to grammar where lexis takes over' Halliday 1961 p.72), it can be handled in my inventory under 'referential range' and 'functional range'. The dimension 'speed of communication' is also dispensed with. For the tester's validation and scoring purposes, it is more relevant as an aspect of the time-constrained characteristic of communicative dynamics and as a criterion for evaluation. Munby's dimension 'flexibility of communication' which 'covers the capacity to handle novelty or communication unrelated to the participant's own purposive domain, and adaptability to switching of subject, style or interlocutor' (p.165) is subsumed under my functional range category.
'Novel' or 'unrelated' communicative acts are still part of a participant's involvement in that event, an aspect of the 'variety of illocutionary acts' I mention. The useful notion of flexibility is also one of my criteria for assessment below.

In his needs analysis model Munby goes on to suggest a means of modifying the predicted dimensions of communicative activities for tolerance, glossed by Bowers (1978) as:

"a measurement of that latitude which is conventionally allowed to the participant by those with whom he communicates in the role-sets specified by the model, or by those who in such settings (as eg in academic contexts) have the authority to evaluate his communicative ability" (p.A4).

Now the applicability of the notion of tolerance of linguistic error, stylistic failure, reference (to sources of help), repetition and hesitation to the evaluation of competence and performance is noted by both J B Carroll and Morrow. However, the way they choose to combine dimensions and tolerance features as 'performance criteria' (Carroll 1980 p.31) or 'operational specifications' (Morrow 1978 p.156) to move directly into their descriptive bands is problematic. I choose to keep the dimensions used to describe a particular event separate from (though not unrelated to) criteria for the assessment of test task performance. The rationale for this is that the tester needs to be balancing the achieved against the required result (see Holec (1980) above). 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, for example, the complexity of usage a native speaker might feel was
demanded by it, and the criteria for evaluation, for example, the degree of fluency that a testee demonstrated when participating in it. This fluency may well vary according to the inherent dimensions of a communicative event; a learner will probably perform more fluently in a task where the inherent complexity, functional and referential range required are low, than he will if they are high. It will be useful, especially in a case such as my own where group members are evaluated in the performance of such a large number and variety of tasks, to be able to relate, in a reasonably systematic way, the evaluation of the performance and the dimensions of the task.

For this latter purpose, it would be enough to rate the dimensions of a particular task in fairly approximate terms eg:

<table>
<thead>
<tr>
<th>Dimensions of note-taking task Test 1</th>
<th>(see Chapter Six)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Size:</td>
<td></td>
</tr>
<tr>
<td>Complexity:</td>
<td></td>
</tr>
<tr>
<td>Functional Range:</td>
<td></td>
</tr>
<tr>
<td>Referential Range:</td>
<td></td>
</tr>
</tbody>
</table>

The practical usefulness of the notion of parameters, dynamic characteristics and dimensions was demonstrated when I 'piloted' my initial version of them to describe, and qualitatively validate a 'Text Interaction Proficiency Test' (TIPT), part of the post-test battery used with the 1979 ODA group. The result was not only an extremely comprehensive description of the parameters, characteristics and dimensions of the task but also the exposure of some possible
shortcomings in the TIPT as a test. Under the interaction parameter of role-set, for example, it was found that an ambiguity in the task instruction caused uncertainty about the addressee-addressee relationship. This in turn affected the inter-subjectivity characteristic (the negotiation of meaning in the light of shared knowledge about shared knowledge) of the event and thus reduced its communicative validity. If we had used the parameters, characteristics and dimensions at the a priori validation stage, this kind of problem might have been avoided. My general evaluational requirements were themselves modified in the course of the 1979 group trials. I dropped Munby's parameter of 'communicative key' (referring to the tenor or tone of an event) as a separate heading, since it can be variously subsumed by aspects of social relationships, psycho-social setting and communicative function, though I do not underrate the importance of key, as will be seen from the criteria for evaluation described below. Then, there is the change from my original parameter 'enabling skills' to the two parameters 'code' and 'communicative operations'. The analysis of the expected linguistic output of the TIPT revealed a need to distinguish between lexical and syntactic components and higher order cognitive operations such as 'scanning to locate specifically required information' (cf Munby 1978 p.130). The absence of any real hierarchical structure in Munby's 'Taxonomy of language skills' has always been a problem for those using his model for syllabus definition purposes. It will be one for testers, too, unless some attempt is made to make at least the basic distinctions. This is not, however, to deny the descriptive utility and potential generalisability of many of the 'micro-skills' identified and collated by Munby.
My general requirements 1 to 3 are directly concerned with test events and only indirectly (though crucially) concerned with the assessment of people's performance in them. In the case of the TIPT used with the pilot group, the connection between performance 2 and assessment in terms of descriptive levels on a banding chart is by the direct route. The nine levels, labels and band rubrics were taken from B J Carroll (1978), modified slightly by the test constructor to fit the TIPT and the oral interaction test that went with it. It will be noticed from the extract from the Banding Chart shown below that some of the descriptions of level refer to criteria such as 'accuracy' and 'appropriacy'. However, categories that in my approach are only used as 'dimensions' (eg range, complexity), here also form part of the band descriptions. The test constructor follows B J Carroll and Morrow in combining Munby's target level and tolerance factors as this extract from his performance test banding chart shows:

<table>
<thead>
<tr>
<th>BAND</th>
<th>Expert Communicator</th>
<th>Usually only achieved in L₁, or bilingual plus education/extended practice</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Very good Communicator</td>
<td>Gives/gets all information wanted with almost complete control of range complexity and appropriacy. Interacts successfully in informal complex situations and only needs procedural training for formal complex events/activities.</td>
</tr>
<tr>
<td>7</td>
<td>Good Communicator</td>
<td>A responsive and effective communicator in all interactions, maintaining several points of information, initiating or backtracking well. Performance is appropriate/accurate in usual situations but needs to adjust still in more complex events/activities.</td>
</tr>
</tbody>
</table>
6 Competent Communicator

Can operate independently in a good range of interactions. Can pursue details in information giving/getting and cope with the known. Can communicate/receive at length though with concentration. Gaps in performance can be covered by rewording quite effectively. Increasing ability in handling different text types, though information giving/getting through implication or nuance is still to be developed.

etc.

The most fundamental issue raised by the use of this kind of band description must be the question of generalisation. A learner is assessed as a 'good communicator' on the basis of a single, fairly restricted information retrieval, selection and representation task. Such an assessment generalises not only to other similar events but also to different contexts and different levels of difficulty. An excerpt from the chart is cited again here with my emphases indicating where assumptions seem to have been made about the generalisability of performance on this test:

"A responsive and effective communicator in all interactions, maintaining several points of information, initiating and backtracking well. Performance is appropriate/accurate in usual situations but needs to adjust still in more complex events/activities".

Morrow's investigations in connection with new tests for the RSA show a somewhat more explicit and refined route from performance on a task to description of performance level, although the levels themselves are broader ('basic', 'intermediate', 'advanced') than the nine levels specified by B J Carroll. Morrow identifies four evaluation criteria for the assessment of writing performance, provides a description for each criterion and puts the descriptions together to summarise a general level of performance. The excerpt here is from a 1980 working paper:
"Intermediate Level

Accuracy: Grammatical, lexical and orthographical accuracy is generally high, though some errors which do not destroy communication are acceptable. Handwriting is legible without effort.

Appropriacy: Use of language is appropriate to function. Some adaptation of style to the particular context is demonstrated. The overall intention of the writer is always clear. Layout appropriate.

Range: A fair range of language is available to the candidate. He is able to express himself clearly without distortion.

Complexity: Texts will display simple organisation with themes and points linked and related." (Morrow 1980 p.7)

There is less generalisation here and the separate criteria allow a more analytic evaluation. Neither this approach nor that used by B J Carroll, however, seem to cater for learners whose performance levels vary in terms of different criteria. The individual with a fluent, appropriate use of language but a tendency towards grammatical inaccuracy in discourse that is not particularly coherent is a far from unlikely candidate. He would be hard to place in either of these hierarchies.

The most logically refined connection between performance and assessment is probably exemplified by the United States Foreign Service Institute Rating Procedure (eg Adams and Frith (eds) (1979)). If all the available steps are taken, the FSI assessor would carry out the following operations:

- the evaluation of performance 2 according to five criteria (accent, grammar, vocabulary, fluency, comprehension) for each of which any of six descriptive levels may be selected eg:
Fluency
1. Speech is so halting and fragmentary that conversation is virtually impossible.
2. Speech is very slow and uneven except for short or routine sentences.
3. Speech is frequently hesitant and jery; sentences may be left uncompleted.
4. Speech is tonally hesitant, with some unevenness caused by rephrasing and getting for words.
5. Speech is effortless and smooth, but perceptibly non-native in speed and evenness.
6. Speech on all professional and general topics is effortless and smooth as a native speaker's.

(from Valette 1977)

- the processing of each assigned criterion through the FSI weighting table is:

- the summing of the weighted scores to give a total which is then convertible into a descriptive proficiency rating combining functional and criterial elements eg:

Level 4: Able to use the language fluently and accurately on all levels normally pertinent to professional needs. Can understand and participate in any conversation within the range of his or her experience with a high degree of fluency and precision of vocabulary; would rarely be taken for a native speaker, but can respond appropriately even in unfamiliar situations; errors of pronunciation and grammar quite rare; can handle informal interpreting from and into the language.

The actual FSI criteria and their respective weightings are not the main point of interest here. What is most relevant is the comprehensiveness and logical completeness of the FSI route from performance to assessment. The three different routes exemplified by the TIPT, the RSA trial test and the FSI approaches may be represented as follows:
The actual selection of criteria for evaluation will derive from my testing construct. Typically, as can be seen from the examples already cited, the selection and weighting reflect the tester's priorities. Thus Jakobovits and Gordon (1974) emphasise the communicative rather than the linguistic aspect of competence, fluency rather than accuracy. They choose to rate:

- accuracy of information; amount of information related; fluency;
- naturalness of discourse organisation; style of expression; clarity of expression; naturalness; complexity of transactional performance.

But one detects a fair amount of overlap between criteria here which might make actual rating difficult. Schulz's 1977 criteria are actually described as a 'communicative competence scale'. She uses:

- fluency (taken as is from the FSI); comprehensibility; amount of communication (from Bartz 1974); quality of communication.

In order to satisfy my general requirement 3 which covers my evaluation criteria and how they are used to score and profile performance 1, 2 and competence, I am informed by the construct
developed in this chapter and by the practices of the other testers quoted. My criteria, which are described in detail in Chapter Six and presented in full in the Appendices, include some that are common to all or most of my language tests and tasks and others that are used only where they are required by particular tasks or for particular hypothesis-testing reasons. The common criteria with a summary of the features that are developed into functional descriptions, normally for six different levels of performance for each criterion, are:

<table>
<thead>
<tr>
<th></th>
<th><strong>Formal Accuracy</strong></th>
<th><strong>Referential Adequacy</strong></th>
<th><strong>Socio-cultural Appropriacy</strong></th>
<th><strong>Fluency</strong></th>
<th><strong>Flexibility</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>(covering range and control of major and minor patterns; effect on communication)</td>
<td>(covering range and accuracy of lexical usage; repair and avoidance strategies; effect on communication)</td>
<td>(covering adaptation of utterances to the social context; effect on communication)</td>
<td>(covering speed, evenness, ease and intersentential connection)</td>
<td>(covering ability to adapt to and initiate topic switches)</td>
</tr>
</tbody>
</table>

Other criteria used when required include: quality, relevance and accuracy of information processed; text reduction skills; organisation ability; listening comprehension; 'accent'; 'survivability' factors and creativity. Whenever feasible Route 3 is taken, because it allows for greater flexibility in the selection of criterion level descriptions, leads to a fuller, more detailed profile but at the same time yields scores that are amenable to statistical analysis.

The summary of my four general evaluational requirements in Figure 4.6 below may serve as a reminder of the need for the communicative evaluator to have a framework that will help him to understand, construct and validate an event in terms of its parameters, dynamics
and dimensions. When an event can be understood in these terms, appropriate criteria for the evaluation of participants' performance in it are more easily identified. The connections between the sets of categories in this working model are made by the 'cybernetic and emergent properties' of the event itself (Hymes 1970 p.20) as well as by the evaluator.

Figure 4.6: General Evaluational Requirements
As predicted at the beginning of the chapter, the decision to take the fundamental question of validity as the starting point, has ensured coverage of the key theoretical and practical issues in testing and evaluation. Apart from the examination of validity itself with all its dimensions and their hierarchical relationships, I have re-assessed the implications of the notions of competence and performance in testing, tests for different purposes, the teaching:testing relationship, reliability, synthetic and analytic views of language and assessment, direct and indirect evaluation approaches, generalisibility and the nature and use of criteria for assessment. I now have a rationale and a stance. How the evaluation instruments I have designed reflect and affect these will be seen in the report of my empirical study. Only this can reveal whether my tests actually achieve their required comprehensiveness, flexibility, comparability, in a practicable, acceptable, reliable and reasonably economical way. In short, only this can reveal whether my tests are valid.

8. Research Hypotheses 3 Formulated

The re-assessment of testing theory and practice here does not suggest that there are insurmountable barriers to the continued search for support for my main thesis (TI) that sensitive profiling of EAP learner/users can help provide them with better preparation for the training events of their lives in C₂. There are certainly going to be problems eliciting and measuring data in line with my communicative construct, but the fact that so many of these problems are still unsolved increases rather than reduces the value of the search for at least some of the solutions. RH₃, the most basic, instrumental of my research hypotheses must be formulated:
This time we know the direction of the hypothetical relationship.

Chapters Six and Eight will present the means, observations and analyses involved in the testing of the numerous specific hypotheses essential to the search for the validation of RH₃.
180.
PP 181 - 188
Missing
From Thesis
CHAPTER FIVE

THE DESIGN FOR THE EMPIRICAL STUDY
1. Introduction

This chapter has the following aims:

1. to characterise and present the design for my empirical study
2. to discuss theoretical and practical aspects of the selection of participants in such a study
3. to describe my participant group in relation to relevant reference populations

This chapter's pivotal role is shown in Section 5 in Chapter One. It represents the transition from hypothesis formulation to hypothesis testing. The three general Research Hypotheses ($RH_1$, $RH_2$ and $RH_3$) derived from my initial problem via its examination in the light of previous discussion and inquiry, are now to be put to empirical test. The plan for the empirical study is described here; its implementation is reported in the three succeeding chapters.
2. **The Design of the Empirical Study**

The following key features of the design reflect its hypothesis-testing functions:

1. The design allows for the required *multi-dimensional focus*. Factors of presage, programme, process, community and product (see Figure 1.1 Chapter One) are at issue. Their status and interrelationships as variables must now be examined. Hence the elicitation of data on TL, personality, cognitive style, attitude/motivation, academic and other C2 experience.

2. The design must take account of both *individual and group characteristics*, the ideographic as well as the nomothetic dimensions. It therefore needs a combination of standardised (or standardisable) data and individual-sensitive information.

3. The design must be *longitudinal*, measuring more than one set of elements on more than one occasion, tracing and evaluating changes in people in changing situations.

4. The design is *methodologically eclectic*. The study combines aspects of survey, case study, causal comparative and developmental inquiry methods because the objectives in 1 to 3 require them.

Figure 5.1 summarises the focus, methods, chronology and locations of the data collection.
Certain points in explanation of the diagram should be made here. They will touch on key theoretical and practical aspects of the functioning of the design.

1. The investigation is a 'pure' longitudinal study in that it attempts to follow all participants rather than a sub-set through the entire period or the crucial first year of their stay in Britain. My interest in describing and explaining changing individuals in a changing environment, in postulating causal relationships between early and subsequent events, requires contact 'over a time span long enough to encompass a detectable change in developmental status' (Moore 1968 p.152). The fact that the contact is not uninterrupted may suggest Zazzo's 1966 category of 'cross-sectional developmental' as an additional appropriate label for the study. This certainly fits the division of the inquiry into two phases, the first involving the creation of potentially predictive learner profiles (Profile 1) from data elicited during the initial intensive pre-sessional programme at ELTI, the second extending and refining this profile into Profile 2, which becomes the criterion for validating the first.

2. Figure 5.1 also indicates the way in which the study repeats various elicitation devices in identical or parallel forms. Participants' TL communicative competence is measured four times, for example. Pre-Test and Post-Test 2 are identical whereas Post-Test 2A, administered at the first reunion, is an alternate form, itself re-administered identically at re-union 2. A free-writing measure (T5) is again taken four times, but never in identical-task form. 'Objective' data on motivation and attitude are collected three times with part of the instrument remaining unaltered but parts modified
according to changes in time and environment. The aim is to balance the quantitative advantages of repeatability with the qualitative need for flexibility.

3. The design as a whole permits the inter-twining of quantitative and qualitative threads. Personality factors are objectively measured three times in Phase One. But the continuous assessment procedures during that phase along with the in-depth follow-up interviews in Phase Two, add more qualitative information to the objective data. In fact, all parts of the profiles are based on both kinds of evidence.

4. The design has a priori and a posteriori features. The theory and practice of the TL test battery, the use of Cattell’s 16PF inventory and Witkin’s Group Embedded Figures Test, the instrumentation for the collection of programme, process and community data on the visits to receiving institutions were, for example, designed into the study from the beginning as a result of insights gained from the literature and from experience in my pilot work. Other elements, such as the topics for group discussions at the reunions or some of the variables to be included in Profile 2, however, emerged as the empirical study proceeded:

".... a well-designed, longitudinal study anticipates the possibility of spotting the unpredictable influence, the critical period - and any worthwhile definition of the process of research must include the occurrence of insight along the way, unforeseen and arising from living with the data obtained ...." (Wall and Williams 1970 p.24).

Of course, there are times when the study has to handle oversight as well as fore- and hindsight. When this happens, adjustments will be made and seen to be made.
3. **Sampling: theory and practice**

Since empirical research by definition seeks some measure of **generalisability**, it must be informed by the theory and practice of **sampling**. Even where the aim is to explore 'the complexities and uniqueness of the individual' (Wall and Williams 1970 p.1) in complex and unique environments (individual variabilities in both cases that it is often the purpose of sampling to control out or to submerge), the potential generalisability of findings must be established.

The most rigorous form of sampling, the **simple random sample**, where every subject in the reference population is equally likely to be selected for the survey or experiment, is most obviously required when the objective is 'the determination of population parameters' (Eysenck 1975 p.196). If your need, for production planning purposes, is to find out how many customers are going to prefer red Mini Metros, estimates based on a non-representative sample of the general population may be misleading. The **simple random sample is ideal for the straight** survey focusing on the prevalence of a single variable. When the focus is on a range of inter-related variables, this brand of sampling becomes less suitable.

**Stratified sampling techniques**, where the researcher divides his target population on the basis of variables 'known to relate to the characteristics under study' (Bennett 1973 p.42) might seem more appropriate. It could mean, for example, using 'strata' such as CI and TL background, specialist subject area, levels of main course, then selecting learners (randomly) from each stratum. In fact certain **stratified sampling techniques** are used in my study (see Experiment 3.
in Chapter Seven), but for the purpose of controlling variables rather than for the selection of a representative sample.

The students on the ELTI/ODA pre-sessional course, who became the participants in my empirical study were not a selection based on any systematic form of probability sampling. In standard research terms they would be classified as an opportunity sample, that is a group available, willing and permitted to take part in the study as well as, if evidence from my pilot work with the previous year's parallel group could be believed, suitable for it. Now, discussion of the validity of this kind of sample almost inevitably reflects people's research-methodological stance. And within this framework for debate the key issues seem to be:

1. should there be probability sampling?

and 2. can there be probability sampling?

The strong case for systematic random sampling comes from the psychometric camp informed by the methods of normal science:

"Random selection assures that the observations will be representative of the performance of the appropriate reference group and free from systematic bias" (Calfee 1975 p.47).

"Two of the most important (assumptions of inferential statistics) are that the variable in question should be normally distributed in the underlying population and that error variance within each treatment group must be homogeneous" (Cochrane and Duffy 1974 p.118).

According to this methodological paradigm your sample has to be properly random or generalisation from it by means of parametric statistics is not justifiable. Yet it is not even necessary to leave the psychometric camp to find objections to this hard line. In his provocatively entitled paper 'Who needs a random sample?' Eysenck (1975) makes points that are relevant to my study. One is that there is no
inherent logic in an insistence on a random sample of the general population:

"The selection of a proper sample should derive from one's theoretical position, and not be imposed by irrelevant Baconian principles. In certain circumstances a random sample may be appropriate, in others not." (p.198)

One circumstance where a random sample would be inappropriate is where an investigation or experiment has as its focus 'the specific nature of a sample or even of individuals' (p.197). When I investigate, for example, the effect of a particular study task on analytic learners, (see Experiment 3 in Chapter Seven) the use of a random selection of subjects would make more difficult the maximisation of systematic variance (in this case field-dependency).

Eysenck makes another relevant point:

"Differences between samples would be of considerable interest, and would require new hypotheses (or possibly they could be accounted for in terms of (the original) hypothesis, which could be used to generate predictions regarding different samples)" (p.197). My emphasis.

My study is, as has already been established, essentially exploratory, attempting to discover things about selected variables in learners and learning in a range of contexts. Provided that the participants concerned can be related to identifiable underlying populations, the exploration should create useful opportunities for further research with related but not identical samples.

From the alternative, non-psychometric methodological paradigms comes a message emphasising the non-use rather than the misuse of systematic sampling. The often convincing case for case studies has strongly influenced my empirical work. Wall and Williams (1970) put it thus:
"The main advantage of the case-study approach lies in the greater detail and greater precision of information, particularly of a qualitative kind, which it permits compared, for example, with the retrospective and contemporary quantitative data usually obtained in large-scale surveys ...." (p.4).

Hamilton and Delamont (1974) add potential generalisability to the advantages of the approach:

"to an ethnographer (as opposed to an interaction analyst) the development of generally or universally applicable statements is quite a different task: one that is never achieved by merely surveying the field. Despite their diversity, individual classrooms share many characteristics. Through the detailed study of one particular context it is still possible to clarify relationships, pinpoint critical processes and identify common phenomena. From these, abstracted summaries and general concepts can be formulated which may be germane to a wider variety of settings. Case studies, therefore, are not necessarily restricted in scope" (P.327).

Like Wall and Williams I seek the advantages of 'detail and greater precision of information' that derive from a case-study approach. Like Hamilton and Delamont I shall want to claim a certain degree of generalisability and feel entitled to do so given the typicality (if not the strictly statistical representativeness) of my group in terms of the strata identified in the sample population description below. As a collection of case studies carried out almost entirely by a single researcher the group is in fact above average in number. And I interpret as additional support for my small-sample, case-study approach the respected tradition in child language research of the use of 'samples' as small as one, three and five respectively. (See Leopold (1954), Imedaze and Uznadze (1967); Schumann (1977), Brown et al. (eg 1963); Wong Fillmore (1979)). In such cases, the use of small non-random samples is not considered an insurmountable obstacle to generalisability even when, like Wong Fillmore's, the focus is on individual variation. Provided that constraints on generalisability are made
explicit, there is no reason why it should be.

The following answers can thus be abstracted to the question of whether samples should meet the requirements of systematic randomness. Yes, if the purpose is to predict the prevalence in the reference population of a quantifiable variable on which planning decisions are to be made. No, if the purpose is to explore the scope and interrelationships of variables (including randomly-occurring biases) to provide insights and leads for further research. When this is the primary purpose of a study, only an improbably atypical or irrelevant sample is a bar to fruitful discovery or discussion. Such a sample is actually very unlikely to occur when an opportunity group of learners such as my own is taken, as constituted by the normal events of pre-sessional selection.

It is a further paradox of competing research methodologies that the ideal of random sampling never really seems to be achieved. Cochrane and Duffy's four-year survey of studies in the Bulletin of the British Psychological Society (1974) found that only 1 to 24 'based their findings on a true sample of the general population' (p.120). Key 'errors' were sampling method, size, bias in favour of captive audiences of university students or volunteers and inadequate population descriptions. Eysenck would feel that the researchers concerned were not so much unable to randomise seriously as unconvinced of the need to do so. But the problem in educational inquiry is just as often the practical difficulties involved in assembling and keeping a random sample. It is interesting to note that most of the empirical studies of language learning variables discussed in Chapter Two use the label
'subjects' rather than 'sample' in the description of their methodology. The amount and type of population description that follows varies significantly. Wesche (1979), one of the few to use the term 'sample', describes her students according to key variables, relating them explicitly to her sampling frame:

"The students were English-speaking Canadian public servants learning French in a six-hour daily, audio-visual, beginning level training programme. Three-week cycles of full-time training were alternated with mid-week periods at their regular jobs. The classes observed were in the highest of four initial ability groupings, meaning that student aptitude scores on the MLAT short form were not lower than the average score for the population on training. This high ability group, consisting of 37 students was compared on certain characteristics with a sample of 163 students selected randomly from the total population receiving training at the same time. The high ability group closely resembled the random sample in age (with respective averages of 33.8 and 34.8 years) and level of formal education (three years of university training on the average for both, with a range between high school diploma and the Ph.D). The high ability group included a higher percentage of males (81 percent versus 73 percent) and its MLAT Short Form average scores were significantly higher than those of the random sample (64.6 versus 53.1, with standard deviations of 13.2 and 16.2 respectively)" (p.416).

Although Wesche's sample is not random in the strictest sense (it is in fact an opportunity sample) it is adequate for her purposes, namely 'to identify characteristics of those students who were particularly successful in acquiring listening and speaking fluency in French during nine weeks (three cycles) of training' (p.416). Above all, it is comprehensively yet relevantly described.

Oller, Perkins and Murakami (1980), however, give a description of subjects that illustrates the difficulty of getting the sample you want and the confusion that may then arise:

"In all, 182 foreign students at the Center for English as a Second Language .... were tested as part of the spring testing project in 1977. Owing to absenteeism and the voluntary nature of participation in the attitude part of the study, between 45 and 101 students completed relevant portions of the questionnaires, the oral interview
and the language tests. There was some selectivity favouring the better students because the weaker ones tended to complete fewer language tests and fewer attitude questionnaires, but all levels of CESL were represented. Practically all the subjects were males between the ages of 19 and 30 and the largest language backgrounds represented were Arabic, Persian and Spanish" (p.235).

Such deficiencies in sampling theory and practice would certainly disturb strict psychometrists, especially in a study seeking predictor variables 'important to learning English as a second language for adult foreign students in the United States' (p.233). However, it is probably the lack of precision and the question-begging in the description of subjects that are its main weaknesses rather than the fact that the sample is non-random. By what criteria, for example, were some students 'better'? The impression is that either the sample should have been modified in the light of the patchiness of the data or that the reasons for the gaps were potential issues for fruitful investigation, therefore not belonging in the description of subjects. It is perhaps one of the inherent problems of 'difficult' samples that their description becomes complicated, tending to overlap with the reporting of the procedures of the study itself. In the Good Language Learner (1978) there is a detailed account of how the selection of students for the main classroom study was affected by the mismatch between test and teacher ratings of French proficiency:

"It was thought to be of interest to include such mismatch cases whenever possible. It was hoped that classroom observation and the subsequent interviews with both the teachers and students concerned would reveal some of the underlying reasons for the discrepancies" (p.43).

The point is fair but it confuses the description of subjects. Perhaps this should be confined to their description according to the strata selected or accepted as necessary categories of base line data plus
their status in relation to the reference population. In my case the
description of participants will be in terms of certain presage factors
identifiable at the outset of the longitudinal study ie sex; age; Cl;
TL background and need; specialist subject; level, location and duration
of training. And although some of these factors are re-examined as
part of the actual empirical study, which, after all, has as one of its
key functions the re-description of participants in a finer-grained
way, the new data will be kept clearly separate from the base line
description. Discussion of the sample status of the group, however,
will accompany the initial description since no type of sample is
meaningful unless it is related explicitly to its frame. (see Section
4 below).

So, can simple random of populations be achieved in practice? The
answer seems to be no. The great majority of the groups in the
studies surveyed in this thesis are opportunity samples used partly
because 'there is no alternative to the researcher' (Bennett 1973),
partly because such 'natural groups' as school classes or pre-sessional
learners are actually just the kinds of groups applied linguists with
interests or problems deriving from language learning and teaching
experience feel are the most relevant.

But even if the general message is that random samples are not often
appropriate for in-depth, multi-variable, developmental research, and
anyway not often possible to achieve, there are lessons to be learnt
from this brief look at the sampling question. The most important are:

1. that the group under investigation should be clearly and
   relevantly described so that
2. the generalisability of findings can be established.
If these lessons are learnt, Bennett's assessment of the potential of this kind of study could be justified:

"Even an opportunity sample may yield information of considerable value - particularly in an exploratory study of an area previously under-researched or in 'case' or 'clinical' studies using a small number of individuals. The value of any conclusions from these types of study will depend on an intelligent assessment of how far the conditions have been satisfied for valid scientific generalisation" (1973 p.50).

4. Describing the Participant Group

First, then, certain technical points of explanation of the categories in the descriptive matrix, Figure 5.2, then the matrix itself and finally in this section some summarising statistics and discussion of the data characterising the group and its relation to various possible reference populations.

In Column 1 the letter arbitrarily assigned to each participant is for identification purposes when individual cases are mentioned in the body of the report of the empirical study. Column 2 states the sex of each participant, Column 3 his or her age on July 1 1980, the month of the beginning of Phase One. The country named in Column 4 is the country of the participant's normal residence.

The categories covered in Columns 5 and 6 are rather less straightforward. Column 5 attempts a differentiation between English as a foreign and second language with 'FL' used unless a significant proportion (e.g. the secondary or tertiary period) of academic education has been in the medium of English, that is 'SL'. Even this distinction is not always clearcut especially for countries with recent changes in language policy; the issue has to be one for closer scrutiny in the course of the study itself.
The question of establishing a summarising initial category for proficiency in English is an even more complicated one. After close examination of official British Council documentation on each participant, which always includes details of English language learning experience and at least one in-country assessment of competence for the UK training programme to be undertaken, it was decided to enter in Column 6 the number of weeks of intensive pre-sessional language training recommended by those who had evaluated participants' English most immediately prior to their departure for the UK. Other possible criteria such as length and type of English instruction, contact with native speakers, previous performance on standardised tests are not used at this stage. It is notoriously difficult to equalise such data given the heterogeneity of backgrounds, variety of tests used and the differences in timing of the assessments. Besides, the whole question of TL learning and evaluation is a key focus of the study itself, especially in the testing of Hypothesis 3. Meanwhile the figures in Column 6 should be seen as a relevant official assessment of participants' immediate TL needs.
<table>
<thead>
<tr>
<th>Identity</th>
<th>Sex</th>
<th>Age</th>
<th>Country</th>
<th>Status of English</th>
<th>Recommended Pre-specialist (in Weeks)</th>
<th>Area of Specialist Training</th>
<th>Level of Specialist Training</th>
<th>Receiving Institution</th>
<th>Proposed Time (months)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>m</td>
<td>29</td>
<td>Thailand</td>
<td>FL</td>
<td>12+</td>
<td>Devt.Econ/s</td>
<td>MA</td>
<td>U. of East Anglia</td>
<td>12</td>
</tr>
<tr>
<td>C</td>
<td>f</td>
<td>25</td>
<td>Thailand</td>
<td>FL</td>
<td>12</td>
<td>Devt.Econ/s</td>
<td>PGD</td>
<td>Birmingham U.</td>
<td>10</td>
</tr>
<tr>
<td>D</td>
<td>m</td>
<td>36</td>
<td>Brazil</td>
<td>FL</td>
<td>12</td>
<td>Bio-Aeronautics</td>
<td>MSc</td>
<td>Cranfield Inst.Tech/y</td>
<td>12</td>
</tr>
<tr>
<td>E</td>
<td>m</td>
<td>29</td>
<td>Bolivia</td>
<td>FL</td>
<td>8</td>
<td>Paediatrics</td>
<td>Post-doct.</td>
<td>Southampton Gen. Hosp.</td>
<td>24</td>
</tr>
<tr>
<td>F</td>
<td>f</td>
<td>27</td>
<td>Thailand</td>
<td>FL</td>
<td>10</td>
<td>Devt.Econ/s</td>
<td>PGD</td>
<td>U. of East Anglia</td>
<td>10</td>
</tr>
<tr>
<td>G</td>
<td>m</td>
<td>33</td>
<td>Korea</td>
<td>FL</td>
<td>10</td>
<td>Digital Systems</td>
<td>PhD</td>
<td>Brunel U.</td>
<td>36-48</td>
</tr>
<tr>
<td>H</td>
<td>m</td>
<td>26</td>
<td>Thailand</td>
<td>FL</td>
<td>8</td>
<td>Mining &amp; Mineral Scs.</td>
<td>MPhil</td>
<td>Leeds U.</td>
<td>24</td>
</tr>
<tr>
<td>I</td>
<td>m</td>
<td>29</td>
<td>El Salvador</td>
<td>FL</td>
<td>11</td>
<td>Fibre Sc. &amp; Tech/y</td>
<td>MSc</td>
<td>Leeds U.</td>
<td>24</td>
</tr>
<tr>
<td>J</td>
<td>f</td>
<td>39</td>
<td>Egypt</td>
<td>SL</td>
<td>8</td>
<td>Teaching tech.subs.</td>
<td>PG</td>
<td>Bolton Coll.</td>
<td>10</td>
</tr>
<tr>
<td>K</td>
<td>m</td>
<td>30</td>
<td>Mali</td>
<td>FL</td>
<td>12</td>
<td>Tropical Prod.storage</td>
<td>Acad.</td>
<td>Inst.Slough</td>
<td>6</td>
</tr>
<tr>
<td>L</td>
<td>m</td>
<td>40</td>
<td>Sudan</td>
<td>SL</td>
<td>12</td>
<td>Devt.Admin.</td>
<td>PGD</td>
<td>Birmingham U.</td>
<td>10</td>
</tr>
<tr>
<td>M</td>
<td>f</td>
<td>48</td>
<td>Egypt</td>
<td>SL</td>
<td>6</td>
<td>Food Analysis</td>
<td>Non-Gr. Public Analyst</td>
<td><strong>MSc</strong></td>
<td>6</td>
</tr>
<tr>
<td>N</td>
<td>m</td>
<td>30</td>
<td>Ecuador</td>
<td>FL</td>
<td>16</td>
<td>Min.Explor. &amp; Min./Geo.</td>
<td>MSc</td>
<td>Leicester U.</td>
<td>12</td>
</tr>
<tr>
<td>O</td>
<td>f</td>
<td>25</td>
<td>Mexico</td>
<td>FL</td>
<td>12</td>
<td>Urban Design</td>
<td>MA</td>
<td>Oxford Poly.</td>
<td>18</td>
</tr>
<tr>
<td>P</td>
<td>m</td>
<td>33</td>
<td>El Salvador</td>
<td>FL</td>
<td>12</td>
<td>Rural Devt.</td>
<td>MA</td>
<td>U. of East Anglia</td>
<td>12</td>
</tr>
<tr>
<td>Q</td>
<td>m</td>
<td>30</td>
<td>Nepal</td>
<td>SL</td>
<td>10</td>
<td>Law</td>
<td>LLM</td>
<td>King's Coll.</td>
<td>12</td>
</tr>
<tr>
<td>R</td>
<td>m</td>
<td>25</td>
<td>Sudan</td>
<td>SL</td>
<td>12</td>
<td>Vet.Lab. Sciences</td>
<td>Non-Gr. Central Vet.</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>S</td>
<td>m</td>
<td>35</td>
<td>Korea</td>
<td>FL</td>
<td>16+</td>
<td>Mineral Processing</td>
<td>PGD</td>
<td>Leeds U.</td>
<td>10</td>
</tr>
<tr>
<td>T</td>
<td>f</td>
<td>34</td>
<td>Bangladesh</td>
<td>SL</td>
<td>10</td>
<td>Food Resources</td>
<td>PGD</td>
<td>Q. Elizabeth Coll.London</td>
<td>10</td>
</tr>
<tr>
<td>U</td>
<td>f</td>
<td>30</td>
<td>Turkey</td>
<td>FL</td>
<td>4</td>
<td>Sheep</td>
<td>Post-doct.</td>
<td>Edinburgh U.</td>
<td>6</td>
</tr>
<tr>
<td>V</td>
<td>f</td>
<td>44</td>
<td>Indonesia</td>
<td>FL</td>
<td>16</td>
<td>Tropical Prod.storage</td>
<td>Acad.</td>
<td>Inst.Slough</td>
<td>6</td>
</tr>
<tr>
<td>W</td>
<td>m</td>
<td>39</td>
<td>Ecuador</td>
<td>FL</td>
<td>12+</td>
<td>Tropical Prod.storage</td>
<td>Acad.</td>
<td>Inst.Slough</td>
<td>6</td>
</tr>
<tr>
<td>X</td>
<td>m</td>
<td>28</td>
<td>Indonesia</td>
<td>FL</td>
<td>12</td>
<td>Applied Entomology</td>
<td>MSc</td>
<td>Imperial Coll.London</td>
<td>24</td>
</tr>
<tr>
<td>Identity</td>
<td>Sex</td>
<td>Age</td>
<td>Country</td>
<td>Status of English</td>
<td>Recommended Pre-Training (weeks)</td>
<td>Area of Specialist Training</td>
<td>Level of Specialist Training</td>
<td>Receiving Institution</td>
<td>Proposed Time (months)</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>-----</td>
<td>------------------</td>
<td>-------------------</td>
<td>-------------------------------</td>
<td>-----------------------------</td>
<td>-----------------------------</td>
<td>------------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Y m 32</td>
<td>4</td>
<td>FL 10 Irrigation Engineering</td>
<td>MSc Southampton U.</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z m 41</td>
<td>6</td>
<td>FL 6 Text.Dyeing &amp; Printing</td>
<td>MSc Salford U.</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>z m 31</td>
<td>12</td>
<td>FL 12 Insect Contr. res.methods</td>
<td>MSc Southampton U.</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| 8 f 3 A | 5 NME 7 SL 2 Edn | 7 Non-Ac 18 Univ | 5 |
| 19 m 10 B | 6 C 3 S.As 9 SEA/FE 2 LA | 7 Phys.Sc 11 Med | 11 M | 4 Poly 1 Hosp | 14 |
| 6 C 3 S.As | 9 SEA/FE | 7 Bio.Sc | 1 PhD | 5 Govt | 2 |
| 5 D 9 SEA/FE | 2 LA | 9 Agr. | 2 Post D | 9_Soc.Sc | 1 A.H. |
| 2 E 7 LA | | | | |

Mean 32.81 |
SD 6.08 - - 10.78 - 2.89 - - 14.4 |

Figure 5.2: Summary description of the participant group
Where the notes preceding Figure 5.2 briefly explained the column labels, the following discussion points focus on the information in the columns. It is the purpose of the colour-coding, aggregating and the averaging, along with the comment to which they give rise, to clarify the status of my group as a sample with reference to various populations which readers of the research may be seeking to relate it to. Some of the theoretical aspects of sampling explored in Section 3 above are seen here in more practical application.

1. Column 2 specifies the sex of the participants as an essential feature of any case study portrait, not necessarily because it is seen as a key variable. It is noticeable that neither the studies reviewed in Chapter Two nor the 'surveys' surveyed in Chapter Three make much of the male/female issue. However, the overall ratio of overseas student men to women in universities in 1976/77 was 78:22 (postgraduates 75:25) (Commonwealth Universities Yearbook 1979 Appendix II, Table C) and in the Blaug and Woodhall sample (N=1484) in A Survey of Overseas Students in British Higher Education 1980 (Williams ed. 1981) 75% of all respondents were again male. Of the total number of Technical Cooperation and Training Department (TCTD) study fellows arriving in Britain between June 1 1980 and May 31 1981 (N=3873), 85% were male; the proportions of students in this population who received remedial English (N=627) were 80:20. My own 19:8 ratio (ie 70%:30%) seems representative of the three possible underlying populations, the overall overseas student group, TCTD study fellows and the remedial English sub-group.

2. The mean age of the group at 32.81 (SD 6.08) and the mode, which, as the aggregates indicate is the 26-30 interval, reflect the fact
that the participants are post-academic training and post-experience. The overall range (25 to 48 years old) and distribution (none of the six intervals is unoccupied) are helpful if aspects of age as a factor in language learning emerge at all. Most statistics on overseas students do not mention the age factor at all, though Blaug and Woodhall (1981) describe 'the typical overseas student' in their sample as 'aged 0 or below (82%)' (p.241), a fact that can certainly be explained by the more than 50% first degree and HND students present in their sample.

The 15 different countries of origin as given in Column 4 are grouped in terms of geographical region according to normal atlas conventions. These are close to the British Council's general categories (eg The British Council Annual Report 1979/80 p.24) rather than the geo-political categories used by TCTD (eg 'Commonwealth Africa', 'Colombo Plan' etc.) or the classification by per capita income used in the World Bank Atlas (IBRD 1978). The range of countries and regions represented in my group is influenced by two main factors. The first is that all participants come from countries receiving British aid through the ODA. The second is that members of a pre-sessional remedial English group are less likely to come from areas with a strong English medium/ESL tradition. If the British Council's statistics of Overseas Students in Britain 1978/79 are analysed and translated into my regional categories, approximate percentages (based on a grand total of over 87,000 students from the regions concerned) are as follows:
Near and Middle East: 34%
Anglophone Africa: 20%
Non-anglophone Africa: 3%
'ESL' South Asia: 6%
Non-ESL South Asia: 0.2%
'ESL' S E Asia & Far East: 27%
Non-ESL S E Asia & Far East: 5%
Latin America: 5%

If statistics for TCTD Study Fellow arrivals for the financial year 1980/81 are similarly analysed (N=3749), percentages are distributed in a significantly different way:

Near and Middle East: 74%
Anglophone Africa: 50%
Non-anglophone Africa: 4%
'ESL' South Asia: 22%
Non-ESL South Asia: 34%
'ESL' S E Asia & Far East: 3%
Non-ESL S E Asia & Far East: 64%
Latin America: 54%

Clearly the aid factor is the main reason for the higher proportions for Africa and South Asia in the TCTD population compared to the general overseas student figures. It also explains the lower percentages for the Near and Middle East and for S E Asia, the latter chiefly because Malaysia, the biggest single sender of students to the UK in 1978/79, is no longer a major recipient of TC fellowships (only 1% of the 3749 analysed here). But the third, and perhaps most important underlying population given that the need for pre-sessional English is my key
variable, gives a different picture again. The TCTD statistics for the study fellows (from the same population) who were given remedial English instruction show a regional distribution, as follows (N=612):

- Near and Middle East: 19%
- Anglophone Africa: 11%
- Non-anglophone Africa: 9%
- ESL South Asia: 14%
- Non-ESL South Asia: 8%
- ESL S E Asia & Far East: 1%
- Non-ESL Asia & Far East: 20%
- Latin America: 18%

Here the language background factor clearly dominates. The analysis of countries of origin according to my categories shows that students from countries of the Near and Middle East, non-ESL S E Asia and the Far East, and from Latin America constitute much more significant proportions in the remedial English group. Although there are still students from ESL backgrounds, their percentages are down.

My own case study group of 27 participants is too small to describe validly in percentage terms. However, it is worth noting that the three largest regional sub-groups are non-ESL S E Asia and the Far East, Latin America, and the Near and Middle East. All the other regions in the TCTD remedial English population are also represented in my group with the exception of ESL S E Asia and the Far east (in any case a mere 1% of that population). The participant group for the empirical study is then a reasonable, class-size reflection of the underlying remedial English population in terms of regional background.
Now, accepting the fact that a true sample population must satisfy the criterion of size as well as that of representativeness, something a group of case study proportions is unlikely to be able to do, this discussion is not to be taken as a belated claim to systematic randomness; as we have already seen, the priorities and methodology of my study do not require such an approach. What would be disturbing, however, would be a group of participants that really did seem atypical as regards national background. This, fortunately, my group does not.

4. The 20:7 FL:SL ratio in Column 5 is again based on pre-arrival data. A closer investigation of the realities of the second versus foreign language distinction will be part of the study. It is relevant at this stage to note that in the TCTD remedial English population there is an ESL representation. 26% of the students are from African, South Asian, and SE Asian/Far East countries that are conventionally considered as ESL. The proportion in my group is thus apparently typical.

5. The mean recommended number of weeks of pre-sessional English instruction for the group (see Column 6) is 10.78, SD 2.89. The validity of this and other measures of TL proficiency will of course be a crucial focus of this study. For the moment just two inferences. The first is that participants in my group were judged to have an above-average need for remedial English (TCTD's English Teaching Co-ordination Unit estimates the norm for their remedial population at about 8 weeks). The second is that the six-week pre-sessional course at ELTI was officially predicted as insufficient in length for all but two of the participants. In short, mine is the kind of learner group that is normally considered as fairly borderline in TL proficiency terms.
6. The use of colour-coding and aggregating to characterise the group reveals certain points of immediate interest in Column 7, which deals with areas of specialist training and suggests that the DES categories of Science; the Social Sciences; Engineering and Technology, and Agriculture are the commonest subject areas in my group. The first three of these are the traditional leaders:

"These three categories (Engineering and Technology; Science; Social, Administrative and Business Studies) account for almost three-quarters of all overseas students enrolled at institutions in the tertiary education sector" (Weir forthcoming).

A fairly typical set of figures is the breakdown of the fields of study of overseas students in UK universities for 1976/77 (Commonwealth Universities Yearbook 1979 Appendix II, Table C):

<table>
<thead>
<tr>
<th>Subject Group</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education</td>
<td>1493</td>
</tr>
<tr>
<td>Medicine, dentistry &amp; health</td>
<td>2588</td>
</tr>
<tr>
<td>Engineering and Technology</td>
<td>10742</td>
</tr>
<tr>
<td>Agriculture, forestry &amp; veterinary science</td>
<td>741</td>
</tr>
<tr>
<td>Science</td>
<td>6799</td>
</tr>
<tr>
<td>Social, administrative &amp; business studies</td>
<td>6913</td>
</tr>
<tr>
<td>Architecture &amp; other professional &amp; vocational studies</td>
<td>793</td>
</tr>
<tr>
<td>Language, literature &amp; area studies</td>
<td>2274</td>
</tr>
<tr>
<td>Arts other than languages</td>
<td>2104</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>34447</strong></td>
</tr>
</tbody>
</table>

Interestingly enough, in the TCTD overall statistics for 4/80 to 3/81 as re-analysed according to my geographical categories a new candidate
moves into the top three specialist fields, namely 'Education, ELT', at 17%, ousting Science at 8%, though with the Social Sciences (38%) and Engineering and Technology (18%) still in the top two places. The Agricultural Sciences (10%) also occupy a higher place in this TCTD population than in the overseas student population at large (see, for example, the meagre 2% of university overseas students in the table on page 213 above). The table that follows as Figure 5.3 allows some significant inferences to be drawn about the interrelationships between Cl, TL and specialist subjects.
Social Science is clearly the leading subject field for the TCTD group as a whole, a fact, it may be assumed, that reflects the agreed perception of both aid receiving and giving countries that the field is crucial developmentally. Since the very high figure of 38% is made up largely of students from anglophone Africa and ESL South Asia, however, we may expect a less predominant, though still leading place, for the social sciences in the remedial English sub-group (see Figure 5.4 below and the proportion in my own group).

Engineering and technology occupies second place in the overall TCTD population, again dominated by ESL Africa and South Asia. But this area of specialisation, at 18%, is rather less strong than it is in the general UK overseas student population as exemplified in the 1976/77 figures quoted on page 214 above. Presumably the third-world bias of the TCTD group explains the lower, though still significant, proportion of engineering and technology students and the relatively higher figures for education and agriculture.

It is interesting to note that education, the third best represented subject at 17% in the TCTD population, is actually the leading field for three of the non-ESL regions, for the Near/Middle East at 34% of its TCTD fellowships, Africa (FL) at 53% and South Asia at 35%. The reasons for this are not immediately obvious unless education is an area where Britain has already exerted an influence that is now being reviewed and, in some ESL countries, actually reversed after a period during which educational systems had followed the British model quite closely. Such reasoning is speculative, however, especially as the FL numbers are generally smaller than the SL in the TCTD population (23% to 76%) which means that proportions can more easily be distorted by a
single factor, for example the fact that as many as 67% of the Near/Middle East education fellowship holders are from a single country, Egypt.

Agricultural studies are in fourth place in the TCTD population. In this subject area there does not seem to be any SL:FL-related pattern, the nature of the climate and economy of individual countries taking precedence. The third-world factor again probably explains the relatively lower popularity of the physical and biological sciences here compared to its normal second or third place in the overall UK overseas student population. The perceived need in developing societies may well be for a focus on the applied rather than the pure sciences. Still, the latter remain a significant choice, at 8% of the TCTD group, only 2% fewer than for agriculture.

In the two regions that, in my classification Figure 5.3 above, are not subdivided according to the status of English, the Near/Middle East and Latin America, the social sciences and engineering/technology share prominence and almost identical percentages (just over and just under 20% respectively). But whereas the Near/Middle East has education in first place, Latin America has agriculture.

Now, the table in Figure 5.4 has the same axes as the one in 5.3 above to allow direct comparisons between the general TCTD population for 1980/81 and the sub-group for whom pre-sessional English was considered necessary. Such comparisons will also provide further pointers to the representativeness of my own group.
The following are key points with reference to the TL:subject area relationships within the TCTD remedial English group:

1. **Social studies** remains the leading area of specialisation though, as was suggested above, the absence of most of the anglophone African and the Indian study fellows from the remedial English group does reduce the dominance of this subject quite significantly, from 38% to 23%.

2. This change in regional balance is also responsible for the fact that **engineering and technology** now occupies third rather than
second place, though with an almost unchanged percentage.

3. Agriculture is the second most common subject in the remedial English group. Interestingly enough in view of the increasingly complex SL/FL question in post-colonial societies, this is mainly because of the fairly high number of Bangladeshi and Sudanese students in the remedial English group.

4. Education is in fourth place as opposed to the third place it occupies in the overall TCTD group. This may be to do with the fact that about 25% of the specialists in this area are in 'ELT and Linguistics' and are thus less likely to be included in the remedial pre-sessional English group. The tendency noted above for a bias towards FL countries among the education students remains.

5. The positions of science, medicine and arts remain unchanged compared with the general TCTD population though the increased percentage of medical students (up from 5% to 11%) is an interesting result of the relatively high proportion of Latin American students who study in this field in Britain.

6. The most significant point that emerges from my attempts to characterise the areas of specialist training for my own group (see Column 7 in Figure 5.2) is that even at the preliminary stage of archival data collection it was clear that discrete classifications for the courses or programmes concerned were difficult to make. The aggregate of subject fields in Column 7 is greater than the number of participants, an indication of the inter-disciplinary nature of their training. This is obviously an issue for further investigation in Phase Two of the study where the reasonably small number of participants makes
it possible to clarify what the actual subject matter of particular programmes is.

The significant proportions of the students in my group specialising in social studies, agriculture and engineering/technology make it rather more like the TCTD remedial English group than the overall TCTD population in terms of subject areas. The pure sciences appear somewhat more prominent than in either of the TCTD populations, but, as the colour-coding in Column 7 shows, there is not a single programme that is exclusively 'physical' or 'biological' science. A similar situation would almost certainly be found if the subject classifications for larger populations were more closely examined. The medicine component in my group seems fairly typical of other UKOS populations as are 'arts and humanities', which hardly figure at all in either TCTD reference group, and figure in mine only because the British Council classification includes 'architecture and town planning' under this heading.

In summary, the combination of subject areas represented in my group is by no means untypical of TCTD study fellows as a group except that the field of education seems somewhat under-represented.

7. The coding of the levels of training in Column 8 underlines the fact that the group is essentially post-graduate, like 52% of the overseas students at British Universities in 1977/78 (British Council 1980), and like over 70% of the TCTD populations. My most common level of post-graduate training is the Master's degree but the post-graduate diploma course is also well-represented in the group. Most statistical sources do not make the Master's:diploma distinction
but in Blaug and Woodhall's sample (Overseas Students' Trust 1981) there were three times as many of the former as the latter. One reason for the higher proportion of Diploma students in a TCTD sample may be the existence, for example at Birmingham University, the University of East Anglia and Bolton College of Education, of courses established in consultation with the ODA for the particular benefit of overseas aid students. The ODA connection also explains the relatively high proportion of students on special courses or programmes (36% of the total 1980/81 TCTD population, 20% of the remedial English group and 8 out of my 27 participants). In Column 8 the label 'non-academic' is given to all training not leading to an academic qualification. Thus, K and W's six-month course at the Tropical Products Institute at Slough, M's specially arranged attachment at the Public Analyst's Laboratory in Reading and U's familiarisation and training period at the Edinburgh University School of Agriculture. It is clear from Figure 5.2 that the term 'non-academic' does not imply a sub-academic level of training. Both E and U, for instance, do 'non-academic' training at post-doctoral level. Nor should overseas students not attending taught courses be considered untypical. British Council statistics for 1977/78 indicate a not insignificant tradition of overseas students 'outside the public sector' and include categories labelled by OST (1979) as 'industry/professional/government etc.' which has for years accounted for between 20 and 30% of the 'non-public sector' overseas student population (a population of around 37,000 annually since 1967).

8. Column 9 shows that universities are the majority type of receiving institution for my group (18 out of 27). Overall TCTD figures suggest that this is reasonably typical although they specify
level of study rather than type of institution. British Council statistics for 1978/79 show 62% of all higher education overseas students at universities with the remaining 38% at Colleges of Education, Polytechnics and Further Education Colleges. Apart from such conventional non-university institutions, of course, my group has an interesting sprinkling of less typical training contexts because of the non-academic attachments mentioned under point 7 above. It is also useful given the importance of community factors in this study, that the geographical spread of the receiving institutions is so varied, with participants in the North, East, South and Midlands as well as in or around London.

9. The final column, Column 10, codes the proposed duration of UK training, excluding pre-sessional ELT. 19 of the participants, like the majority of TCDT study fellows, came into the 3 to 6 and 7 to 12 month brackets and were thus scheduled to complete their C2 stays before the end of the empirical study. The differences (in both directions) between proposed and actual durations will of course be revealing.

The following summarising statements can now be made about the heterogeneity and homogeneity of my group:

.1 The group may be characterised as heterogeneous in regard to sex; age; country of origin; background status, experience and level of English; area of specialisation; level, location and duration of UK training.

.2 It may be characterised as homogeneous in the following senses:
.1 in that all participants had been awarded ODA study fellowships in fields considered developmentally important in countries receiving British educational aid; all 15 countries represented had annual per capita incomes below $2000 according to World Bank figures for 1978.

.2 in that all participants are post-experience, that is in Britain to receive additional training in a field they were already qualified for, working in and likely to return to; all could be described at the beginning of Phase One as 'EAP students'.

.3 in that all had been assessed before their departure for Britain as needing intensive pre-sessional remedial English, but in the expectation that between 4 and 16 weeks of such instruction would bring their TL proficiency to the level required by their training programmes.

.4 in that they became the ELTI/ODA 'class of 1980' and were aware that they were helping in this particular longitudinal study.

We now have an empirical research framework, a description of the participant group in terms of base-line presage data and an idea of the status of the group in relation to various reference populations. From this starting point it is feasible to embark on the testing of specific hypotheses derived from the three general Research Hypotheses (RH₁, RH₂, and RH₃) formulated in Chapters One to Four. The design and the nature of the group participating in the study should allow useful conclusions to be drawn from a combination of quantitative and qualitative findings.
CHAPTER SIX

THE EMPIRICAL STUDY: PHASE ONE – TARGET LANGUAGE EVALUATION
1. **Introduction**

This chapter has the following main purposes:

1. to summarise implications for the validity of my test battery of the theoretical and empirical discussions in earlier chapters
2. to describe individual tests and sub-tests in relation to their construct
3. to analyse and interpret data from the pre- and post-test batteries in terms of test reliability
4. to examine intra- and inter-test relationships for further evidence on test validity
5. to assess, through selected case studies, the use of Phase One evaluational data in the profiling of individual participants' communicative competence.

This chapter is the first of three reporting on the empirical investigation of the participant group described in Chapter Five. Its main concern is the testing of Research Hypothesis 3, which posits fruitful relationships between language testing data and learner profiles.
2. **Test Construction and Construct Validity**

The construction of the TL performance tasks, and decisions on their place in my test battery take account of:

1. **real-life study events and problems** as identified by the authorities cited in Chapter Two, Section 4 and in Chapter Three
2. **the general evaluational requirements** for communicative performance tests developed in Chapter Four, Section 7
3. **the specific functions** my tests had to fulfil on the British Council/ODA course of 1980
4. **the normal practical demands** on any test battery forming part of any course design.

To begin with, then, a brief look at the influence of 1. and 2. on my test design. It is immediately clear that the decision to build most of the tests around performance tasks (performance 2 as a sampling of performance 1) of the kind identified as typical of the EAP situation in C2 makes the checking of the tasks against my Chapter Four communicative parameters and dynamic characteristics (see Figure 4.6 above) fairly straightforward. In all four performance tests (Tests 1, 3, 4 and 5), **participants** are communicating on their own behalf, as students in interaction respectively with their course director/lecturer; writers and compilers of information on overseas students (including such students themselves); their teacher/interviewers and fellow interviewees. In no case are they required to play roles that are 'pretended' in any more artificial sense than is inherent in all testing or many real-life situations. The intended purposes of participants in the performance events vary in their explicitness, contextual authenticity and degree of external imposition.
In at least two of the tasks (eg Tests 1 and 4 below) the purposes of the exercise are genuine in that the information exchanged is actually required at the time by all parties in the interaction. My communicative parameter of 'activities', namely what has to be done to achieve the purposes of the tasks, represents an inter-level between the task as a whole (eg in Test 3 the processing of written information and the reporting or evaluation of key points from it) and the linguistic code, skills or strategies involved in achieving communicative success. Here we are concerned with complex, varyingly overt and overlapping features such as those exemplified in Chapter Two, Figure 2.6, where EAP needs from the phonological to the study habit levels are cited. In general, however, the activities are a 'natural' part of the performance task, for instance participants expressing their own reactions to the overseas student opinions quoted in Test 3, Part 3.

Under the setting parameter, which, it will be remembered, covers temporal, physical and psycho-social aspects of the activities, the realities of the pre-sessional context are allowed to rule. If descriptors from Munby's 1978 inventory of psycho-social environments (pp.158/159) were to be used, the settings for the test tasks might be characterised as:

- intellectual; public to fairly private; quiet; demanding; hurried; formal to fairly informal.

Such an ambiance does not seem untypical of the study experience at real receiving institutions. But where there are question marks over the appropriacy of some aspects of task settings, they will be commented on in the descriptions of individual tests. The parameters selected in Chapter Four also include Munby's category, 'instrumentality'
covering medium, mode and channel. The tests in my battery involve all media (ie (spoken + written) x (receptive + productive)) and the most normal of his communicative modes (monologue: spoken to be heard + spoken to be written + written to be read; and dialogue spoken to be heard). Print and face-to-face 'channels' are also used. Since the real-life 'role others' (Merton 1957) in the pre-sessional programme were the people also participating in the administration of the oral interactive test tasks, the factor of dialect remains uncontrolled. As it happened, the students were in contact with standard English, Northern, and South-Eastern English dialects from their teachers, as well as with the various versions of the various national standards represented by their colleagues. Again such a mixture does not seem unrepresentative of their later main course settings.

It is only in Test 2, the one that attempts to tap competence rather than to recreate performance conditions, that features of the TL linguistic code are sampled and assessed per se. In the four performance tests, the forms and functions required to convey the required messages are a dimension of the task itself and the participants' chosen ways of handling it. In all these tests, formal accuracy and referential adequacy are nevertheless used as criteria for the evaluation of performance. Similarly with my parameter 'communicative operations' that is, the skills or strategies people activate in order to perform various sub-tasks. Test 5, for example, requires participants to use 'indicators for introducing, developing and concluding an idea' (Munby 1978 p.181) but it is not a task constructed in order to bring into play this or any other pre-selected set of 'enabling skills', strategies, processes etc so that they can
be assessed. Rather, the activation and inter-play of such skills is judged through my analytic criteria on the basis of how effectively the purpose of the task is achieved.

The second set of general evaluational requirements identified in Chapter Four, Section 7 attempts to take account of the dynamics of communication. Given that the performance tests can be described using my parameters for who is to do what, why, where, with whom, how and using which forms, functions or skills, it is helpful also to be able to check out their communicative dynamics, both predicted and in-the-event. An initial fact-finding interview such as Pre-Test 4, entails for instance, the inter-subjectively-motivated filling of relevant information gaps in a setting which is authentic in that it involves the creative handling of both predictable and unpredictable events, naturally chunked and allowing normal self-monitoring under normal time constraints. The test thus comes out well in terms of dynamic characteristics. This is important for the communicative construct, though in practice every effort still had to be made to detect other dynamic effects (for example, here, the influence of the pairing of interviewees) that might still disturb ecological or other aspects of validity. Some of these influences will be mentioned in the detailed test descriptions below.

The dimensions of the test tasks will also be considered then. Clearly questions of the size, complexity and range of each task (see Chapter Four, Section 7) have a bearing on the weight it should be given in the whole battery. But these dimensions are again used mainly as post facto checks. It is the intended communicative relevance of the overall tasks that takes priority; the tasks are not
in general manipulated in order to fit preconceived ideas of size, complexity or range.

In Chapter Four, Section 7 it was decided that the assessment of the performance tests would be based on communicatively appropriate, functionally described analytic criteria for levels of performance that could be weighted and aggregated for an overall score (cf 'Route 3' in Figure 4.5 above). The criteria of formal accuracy, referential adequacy, socio-cultural appropriacy, fluency and flexibility were seen as applicable to most testing events; other criteria would be added as required by the nature of particular tasks. The actual evaluation criteria used for each test are summarised in the descriptions below but some discussion is needed here of how they are used to assess performance in a way that reflects my communicative construct. The following are key points in this connection with quotations taken from the actual criterion scales reproduced in full in Appendices 1.1, 1.3, 1.4 and 1.5.

.1 The criteria need to be comprehensive. Thus the criteria of flexibility and socio-cultural appropriacy are added to FSI-derived criteria for the oral interview test. After all, a participant may understand, respond in a reasonable accent with a reasonable degree of grammatical and lexical accuracy and at a reasonable speed (compare the FSI criteria of accent, grammar, vocabulary, fluency and comprehension) but with a lack of communicative effectiveness not apparently caused by any of these features. Perhaps it is a matter of the socio-cultural suitability of his utterances. There can be few users of any non-native language who have not experienced breakdowns in communication caused by infringements of rules of use. So my
criteria for the evaluation of performance on Test 4 (oral interaction) and Test 5 (writing about C2 hopes, plans or experiences) include a scale of socio-cultural appropriacy. Level 3 from this scale may be quoted as an example of the attempt to take account of the effect of problems caused by 'the required sense of relation to contextual features' (Hymes 1970 p.23):

Socio-cultural Appropriacy

"3. Frequent errors in the rules of social language use result in communication that is often inappropriate to the setting, role-set or in tone. Errors significant enough to cause occasional social misunderstanding."

The FSI criteria also reflect their pre-communicative pedigree by failing to allow explicitly for the unpredictable, topic-sensitive nature of most communication. In addition to, 'fluency', which in FSI terms is all about 'speed and evenness', in mine about these plus intersentential connection, we need a category covering the receptive and productive capacity to handle changes of subject. Level 3 of my flexibility scale, used in Tests 3, 4 and 5 may be quoted to give the essence of this additional criterion:

"Quite frequently thrown by changes of topic. Occasionally attempts to initiate new topics, sometimes unsuccessfully."

.2 The criteria need to be adaptable to the differing requirements of different performance tasks. Not only must criteria be excludable, for example listening comprehension as obviously irrelevant in the writing task, or supplementable, for example by the criterion of compositional organisation, but they must also be easily modifiable according to the demands of medium. So, we have two versions of the fluency criterion:

For the oral interview (Test 4):
"3. Utterances fairly slow, hesitant and uneven. Some utterances incomplete but some are suitably inter-connected."

For the 'free' writing (Test 5):

"3. Written discourse still shows low productive speed and unevenness. A broader range of suitable inter-sentential connectors but some ideas left incoherent."

.3 The criteria must be amenable to differential weighting in line with the different priorities of different tasks. Thus, for example, socio-cultural appropriacy is weighted at x2 in the oral interaction test because of its perceived importance in successful oral communication. In the 'free' writing task, however, where we, the interlocutors, were inviting a personal account of individual expectations or experience, we were implicitly suggesting that the accounts would be read with a high degree of tolerance of unpredictabilities or 'errors' in the rules of use. So the socio-cultural appropriacy criterion is weighted only at x1.

.4 Empirical (as well as programme) requirements meant that the criteria must be quantifiable. In Tests 4 and 5, where 'Route 3' is followed, the descriptive criterion scales are numerically coded (by level), weighted and aggregated. Since each description is numbered on an ordinal scale, this is possible. Thus, the agreed assessment of Participant F according to the two interviewer/assessors at her initial interview:
<table>
<thead>
<tr>
<th>CRITERION</th>
<th>Rating</th>
<th>Weighting</th>
<th>=</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening Comprehension</td>
<td>4</td>
<td>x 3</td>
<td>=</td>
<td>12</td>
</tr>
<tr>
<td>'Accent'</td>
<td>3</td>
<td>x 2</td>
<td>=</td>
<td>6</td>
</tr>
<tr>
<td>Formal Accuracy</td>
<td>4</td>
<td>x 2</td>
<td>=</td>
<td>8</td>
</tr>
<tr>
<td>Referential Adequacy</td>
<td>4</td>
<td>x 2</td>
<td>=</td>
<td>8</td>
</tr>
<tr>
<td>Socio-cultural Appropriacy</td>
<td>3</td>
<td>x 2</td>
<td>=</td>
<td>6</td>
</tr>
<tr>
<td>Fluency</td>
<td>3</td>
<td>x 2</td>
<td>=</td>
<td>6</td>
</tr>
<tr>
<td>Flexibility</td>
<td>3</td>
<td>x 2</td>
<td>=</td>
<td>6</td>
</tr>
<tr>
<td>Extra-linguistic Factors</td>
<td>6</td>
<td></td>
<td>=</td>
<td>6</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td></td>
<td>=</td>
<td><strong>58</strong></td>
</tr>
</tbody>
</table>

So F's performance on Pre-Test 4 can be expressed in qualitative descriptive terms, in the form of a communicative portrait composed of the criterion level descriptions selected by the assessors (and exemplified in detail in Section 6 below) or as a set of scores and a total score amenable to quantitative investigation in comparison with her peers, her own concurrent or subsequent performances or competence etc. As will be seen in the descriptions and analyses of the tests below, the Likert-like scales that are used (for example in Tests 1, 3 and 5) are also both qualitatively and quantitatively describable.

Quantifiability and comparability demand that the levels on a criterion scale should be consistently constituted and clearly differentiated. Raters should as far as possible be spared the problem of feeling that a performance on a particular criterion warrants a description made up partly of features of one level on the scale and partly
of those from others. The following excerpt from the referential adequacy scale for Test 5 indicates how attempts are made to make different levels distinct yet linearly related:

"4. Vocabulary adequate for most of the topics in the intended communication given occasional inaccuracies and/or circumlocutions.

5. Vocabulary adequate all round for the intended communication. Only very occasional inaccuracies and/or circumlocutions."

And finally in this brief renewal of connection between what has been established as desirable and what has actually happened, a reminder of the specific requirements of my tests as part of the ELTI/ODA course design. We needed evaluation procedures that would help us:

1. to group students according to their communicative competence
2. to diagnose communicative strengths and weaknesses
3. to measure progress during and by the end of the programme
4. to follow participants' progress after the course.

Detailed information on these proficiency, diagnostic, progress and achievement assessment functions emerges from the description and analysis below. If the tests perform these roles in an acceptable, reliable, practical and reasonably economic way they will also have satisfied their conventional requirements as assessment instruments.

3. The Tests Described
3.1 Test 1: Lecture and Note-Taking; Dictation

There is general agreement among the EAP-involved authorities cited in Chapter Two, Section 4 that lectures and note-taking are important activities. That the ELTI pre-sessional course should begin with an orientation lecture was to be expected by the participants and was the genuine intent of the Course Director. Her actual talk (on the subject
of programme schedule, groupings and aims, delivered from detailed lecture notes) was therefore designed into the test battery as Test 1 (see Appendix 1.1). It had immediate communicative validity in terms of participation, purpose, activities, setting, instrumentality, dialect, code and operations. Since the event itself was authentic there was no reason why its communicative dynamics should not also be; certainly the information gaps were genuine and inter-subjective, the information processing demands properly chunked and time-constrained.

The key evaluation criterion used to assess performance on the task of interpreting the message of the talk is the quality and quantity of information processed and noted by the participants. The scoring scheme thus selects points from the lecture that it is considered should be noted 'in a way likely to facilitate accurate recall/re-use of the information given' (see Notes for Assessors, Appendix 1.1). The actual validity and occurrence of these points of information was checked during the live lecture by three native-speaker teachers, who made necessary adjustments on the pre-typed assessment sheet. In addition to this major criterion, three others relating to more specific note-taking skills are designed into the evaluation scheme. These are: division into logical sections, text reduction skills and presentation skills, all scored on Likert-type scales and described in detail in Appendix 1.1.

With the major dimensions of a priori validity apparently established in this task, a key concession to reliability was made. The Course Director was invited to prepare a dictation to be given at the end of her talk as a summary of the rationale of the course. The perceived value of dictation as a test of competence (see Chapter Four) made its
inclusion in the battery worthwhile, especially as its relevant content and convenient administration at this point in the testing day meant that it should not interfere too much with the communicative situation.

All scripts were photocopied so that the originals could be returned to participants for immediate orientational use. The initial scoring was completed before midday by three raters (those who had attended the lecture, to validate and amend the information criterion scoring scheme to an agreed final form). Test 1 had thus satisfied the need for economy, administrability and scorability. How it actually fared as regards reliability and predictive validity will be investigated in Sections 4 to 6 below.

Like most communicative tasks designed for maximum authenticity, Test 1 is, in terms of the strictest logic, unrepeatable. However, the purpose, format and means of assessment involved are eminently replicable. And in fact, Post-Test 1, administered at the end of the ELTI course, is an alternate form in almost every respect, based on the Course Director's round-up lecture (and dictation). The requirement of repeatability, so important in a longitudinal study, is thus satisfied.

In Appendix 1.1 the following relevant data will be found:

1. The original lecture notes for Test 1, Post-Test 1 and follow-up Test 1 (the latter analysed in Chapter Eight).
2. Note-taking forms
3. Notes for assessors and scoring schemes.
3.2 Test 2: Sentence Structure, Use in Spoken Context, Use in Written Context, Cloze

In the terms of some of the testing continua discussed in Chapter Four, Test 2 belongs mainly on the competence/discrete-item/reliability rather than the performance/integrative/validity side of things. The debate on such issues suggests, it will be remembered, that the inclusion of a test instrument of this kind is advisable, however strong the inclination towards 'performance' testing. Clearly there is much to be said for a combination of both types of tests in any battery.

Test 2 has four sub-tests. The three sub-tests in Part A are all objective, 4-choice items, 15 on sentence structure, 10 on language use at remark and response level and 10 on language use in written discourse from sentence to paragraph level. The informing sources on the type, level and balance of items selected are:

1. tests used by current EAP authorities at UK universities (some of whose insights are cited in Chapter Two, Figure 2.6)
2. my qualitative and quantitative analyses of an existing objective test as used with the previous year's ELTI/ODA group
3. recent theoretical insights into information structure (again as sampled under 'Cohesion' and 'Coherence' in Figure 2.6).

The analysis of the test used during my pilot work proved relevantly informative, even taking into account the designer's caveat that the test is "a broad, coarse screening instrument for group testing; it is not designed for making fine diagnostic decisions about individual students, a job best done by diagnostic tests coupled with teacher assessments".

A qualitative validation of this test produced the following useful insights for my Test 2:
1. That item selection needs particular care especially when the number in a sub-test is relatively low (35 in my Part A) and the test items basically heterogeneous (see Section 4 below). There is not much room for communicatively fairly redundant items, particularly more than one of them testing the same grammatical rule such as the 1979 test's two items on question tags. And there is enough evidence from the field to suggest that both modals and logical intra-sentential connectors deserve reasonably prominent coverage (see Figure 2.6).

2. That careful attention has to be paid to the elimination of acceptable distractors. Checks with 1979 ELTI staff suggested that even conservative native-speaker intuition would probably rate 9 or 10 of the 20 sentence items in the screening test as ambiguous in this sense. One reason for the tester's apparent difficulty in finding unambiguous distractors may well have been the low degree of contextualisation provided by the short sentences used. Longer, more clearly situated sentences, sentence pairs, or groups, or remark and response excerpts can make items more meaningful, less neutral and thus more contextually constrained with regard to acceptability.

3. That weaknesses in item construction technique can seriously affect the power of items to discriminate between levels of competence. This is the case for example, when distractors that seem intuitively unlikely to distract are included.

My quantitative validation of the previous test based on its performance with the 1979 group (N = 42) also provided useful lessons:

1. An analysis of the test's item discrimination (ID) index (that is the power of an item to discriminate between the upper and lower 27½% of a sample as based on their overall performance on the
test) showed that only 50% of the 20 items in the sentence test had an ID of 25% or above. In only one of these 10 items were all the distractors active. In 5 of the rest, the discrimination index could be explained by contaminating factors such as the presence of a second acceptable (but not accepted) distractor.

2. A detailed analysis of distractors reveals that only a rather low proportion of them did indeed distract; more than a third were not selected by anyone at all on the post-test; another fifth only by a single candidate.

3. The mean score of the top group (n = 14) was lower on the post- than the pre-test. Though some individual regression is not unusual, a general downward trend is, especially when there is no evidence from concurrent validating sources that levels of proficiency in the group actually fell.

Although the three sub-tests in Part A of my Test 2 were informed by mainly negative experiences from use of the existing test, I did take note of items that had performed well. I also benefited from in-confidence access to well-tried university pre-sessional tests, which were particularly helpful in my selection of structural/semantic topic areas for sub-test 1. As a guard against misleading distractors I included a fair amount of contextual information and checked each emerging item with five interested non-naive native-speakers, including ELTI course teachers. The resulting items can be seen in full in Appendix 1.2. In both forms of Test 2 the structural foci are:

---

I am grateful to Brian Heaton at Leeds, Tim Johns at Birmingham and George Blue at Southampton in this connection.
The subject matter of all 15 items is either the test itself or matters connected with students in a position similar to that of the participants. This is more than simply a nod in the direction of shared interest; it often helps to situate the utterance concerned so that there are extra clues of a semantic and pragmatic kind constraining the selection of the appropriate alternative.

The second sub-test (Part A, items 16 to 25) taps features of underlying knowledge of spoken use such as those suggested by Morrow (1977) (see Chapter Four, Section 6). All the items are at the remark + response level, some asking students to provide one or the other to complete the dialogue, others posing metalinguistic questions about the situation envisaged. In this way the following interactive features are covered:

- formal greeting; accepting sympathy; requesting room to pass; suggesting; informal query/request; formulaic telephone/library communication; strong informal agreement; introducing an opinion; showing polite disagreement.

Again the subject matter is appropriate to the participants' current or future situations, social and academic.

Sub-test 3 (Part A, Items 26 to 35) focuses on rhetorical function, using almost exclusively authentic written discourse ranging from single sentence to paragraph length. Cohesion and coherence features examined are:
The question format is metalinguistic, the contexts again overseas-student related.

The second part of Test 2 in both its versions consists of a cloze passage, included mainly on the basis of its reputation for predictive validity and reliability rather than with any preconceptions as to what particular skills, strategies or underlying competence it taps. It is in tune with my communicative construct and with the emphasis in Test 2 on both accuracy and appropriacy, that the two cloze texts used are completely original, that they are on the immediately relevant topic of ODA Study Fellows and that any semantically and syntactically suitable entry is accepted in the text gaps. A fixed-ratio, seventh-word deletion rate is used with opening and final sentences of the text left complete. The passage in the parallel version of Test 2 is the continuation of the first cloze text. Native-speaker informants (n = 5) were asked to complete both passages looking for as many alternative completions as they felt were appropriate. They were also consulted on additional items discovered by the students themselves. Analysed in a similar way to the items in Test 2 Part A, the most probable cloze completion items are:

Text 1 (used in Pre- and Post-Test 2):

- noun (NP premodification); well (adverb of degree); being (passive construction); a/the; defining relative; noun (lexical meaning); in (verb + preposition); be (present/past aspect); verb (past participle passive lexical meaning); in (complex preposition); a/the; noun (lexical meaning); the; verb (infinitive, lexical meaning); a/the; verb (infinitive, lexical meaning); noun (lexical meaning); noun (lexical meaning); noun (lexical meaning); noun (lexical meaning); preposition (subject matter); adverb (location); this/that (substitute); noun (noun phrase); adverb (probability); pronoun (substitute, indirect object).
Text 2 (used in the Phase Two parallel form of Test 2):

and/with (co-ordination); noun (noun phrase); to (infinitive); noun (NP premodification); adjective (lexical meaning); verb (infinitive, lexical meaning); who (defining relative); noun (lexical meaning); to (infinitive); a; and; noun (plural, lexical meaning); of (preposition, division); adjective (lexical meaning); be (passive construction); be (infinitive); adjective (lexical meaning); that (reported statement); noun (plural, lexical meaning); that (result); noun (plural, lexical meaning); noun (NP premodification, lexical meaning); by (passive, agent); should (modal, probability); out (prepositional adverb).

The item descriptions underlined are those that may be considered to be tapping participants' **lexical competence** (or, to use the criterion label used in the performance tests, 'referential adequacy'). The other objective tests I studied tend to include essentially lexically-oriented items in their multiple-choice sections. Part A of my Test 2 does not do this though I accept, of course, that it is not always possible to say whether an item is testing morphological or syntactic or lexical or semantic meaning. Nevertheless, the existence in the cloze tests of items sampling a reasonable range of lexical competence in context left me free to concentrate Part A where I wanted it and increased the economy of Test 2 as a whole.

These, then, were the procedures for the construction and qualitative pre-validation of Test 2. It took 45 minutes to administer and was scored initially on the day of the tests by a team of three raters using lay-over sheets for Part A and the cloze passage key. Scoring was on the basis of a straight one point or zero for each item.

Like most objective tests, Test 2 is eminently repeatable. It was used in identical form at the end of the 1980 course and twice in its alternate form during Phase Two of my study (see Chapter Eight).
Qualitative and quantitative validation analyses and discussion appear in Sections 4, 5 and 6 below.

Appendix 1.2 includes the following relevant data:

1. Copies of Test 2 and Test 2A.
2. Answer sheets.
3. Notes for assessors and scoring scheme.

3.3 Test 3: Reading and Report Writing

Like the study modes sampled in Test 1, the real-life academic relevance of the activities of reading, reporting and evaluating information is in no doubt, as Figure 2.6 in Chapter Two shows. Whether their authentic conditions can be recreated in a testing situation, however, is more problematic as a check against my parameters and dynamic characteristics reveals. In this test, the participants are not processing information needed for their immediate pre-sessional purposes (as they were in Test 1); nor are they working on materials directly connected with their areas of specialisation, since the wide variety of subjects represented in such a group makes the use of all or one of the relevant domains of discourse impracticable. My compromise is to present data related to their situation as overseas students. Their purpose, therefore, is to assimilate and report on what should be of current interest to them. Sub-tasks involve the search for various kinds of information and reacting to it according to varyingly specific instructions. The test compromises, too, with regard to authenticity of setting. In most genuine academic reading and reporting contexts, tasks are carried out over an extended, often non-continuous or non-time-pressured period in a range of settings (in libraries, alone in one's room etc) and with help from a variety of sources (books, journals,
lecture notes etc). And the interactive dimensions of academic reading and writing may include not only a reader's developing relationship with writers of texts but also the active participation of others (colleagues, tutors etc) with a shared interest.

Test 3 attempts to distil the essence of such factors, not the whole reality. In all its versions, participants were given a data booklet divided into three parts, each containing material for a separate task. In Part A the data are a combination of descriptive text and information presented in tabular or other diagrammatic forms. In the answer booklet participants are asked to report on the facts and figures with certain topics identified for particular attention. No guidance at all is given on report format. Evaluation criteria are:

1. The quality, relevance and accuracy of information processed, with marks assigned to particular pre-selected points from the texts.
2. Formal accuracy, scored according to the criterion scale levels as used for the writing task in Test 5.
3. Organisation, covering 'logical sequencing, inter-topic connections, focusing, headings, sub-headings, underlining and other features of lay-out'. (see Appendix 1.3).

The texts for Part B are purely verbal, taken from government reports on higher education. This time the instruction for the task is general but specific headings and lay-out are provided on the report sheet. Criteria 1 and 2 are as for the first task but the third criterion (in the first version of this test) is flexibility using a modified version of the scale developed for Test 5. Flexibility is an appropriate criterion here since participants are expected to infer,
evaluate and switch to their own view of the issues to a greater extent than in the first reporting task.

Part C consists of authentic quotations from overseas students which participants are asked to summarise and react to, the interaction here thus being with the original speakers as well as the potential reader. For this task all the previous criteria are seen as appropriate, ie quality, relevance and accuracy of information; formal accuracy; flexibility and organisation.

Given the special difficulties of achieving close authenticity of topic and task, Test 3 can nevertheless claim to replicate some of the dynamic characteristics of natural academic reading and reporting. It certainly requires participants to perform a variety of activities and operations with a variety of relevant text types with a varying degree of creative, evaluative or subjective involvement. It can certainly be said to exercise a wide range of higher order linguistic, study or 'enabling' skills of the kind so often identified as important. For example, in Munby's 1978 terms:

- distinguishing the main idea from supporting details
- extracting salient points to summarise topics or the underlying point of a text
- interpreting a text by going outside it
- skimming, scanning, transcoding etc.

(pp.129/131)

Although participants 'are processing the kind of input segments that they would normally expect to' (my definition of natural chunking from Chapter Four, Section 7), the size of the texts and tasks is probably not representative, in that they are relatively short. But then the time allowed for their processing is probably unrealistically limited (at 1½ hours for the three tasks). The post facto validation of
Test 3 will be particularly important. Will the compromises balance each other out or invalidate the test?

Test 3 scripts were initially scored by one rater the day after the test. They proved the slowest of all the tests to score (at around 15 minutes per script). Post-test 3 was an alternate form with similar tasks, topics, format and evaluation criteria. The follow-up version for May 1981 was a combination of tasks from the pre- and post-test versions.

In Appendix 3.1, the following data on Test 3 will be found:

1. Data booklets for Test 3, and Post-Test 3.
2. Answer booklets.
3. Notes for assessors and scoring schemes.

3.4 Test 4: Oral Interview

The expert sources cited in Chapters Two and Three almost unanimously pinpoint spoken TL proficiency as important in the academic and social survival of overseas students. Test 4 makes no attempt to recreate the conditions of academic seminars or tutorials as I felt that the heterogeneity of the group and the difficulty of assessing largish sub-groups of potential interactors all at once made this impracticable, especially at the very beginning of our course. Instead, our genuine need to get to know about the participants (and, probably, theirs about us) was used as the basis for an oral assessment built around an interview and taking advantage of the types of authenticity of mutual purpose mentioned in Section Two above.

Pairs of participants were interviewed by two teacher/assessors for up to 25 minutes. On the very first day of my acquaintance with the
group it was not possible to arrange pairings according to any sophisticated group-dynamic criteria but it was decided as far as possible to mix the participants' cultural backgrounds (because some of the questions we wanted them to ask each other would not make much sense otherwise) but not to mix the sexes (because this could be an irrelevantly constraining factor for members of some of the cultures represented). In the event, administrative factors resulted in one female and one same-nationality pairing becoming inevitable, though without any serious repercussions according to the initial interviewers or subsequent video assessors.

The topics for discussion at the interviews were: base line identification information, main courses, CI job, interests, hobbies, hopes etc. These were covered with each participant in turn though with the sequencing of the groups of questions arranged so that everyone had to respond first on two of the topics. Interviewers were at pains to create a friendly, reasonably informal atmosphere, with one of them pre-designated to lead on the first three rather more formal topics, the other on the more personal fourth. Aspects of role-set interaction were further explored by designing the final part of the interview as a participant-to-participant conversation, with the interviewers fading out unless they genuinely needed to join in. The topics for this part of the interview were: TL learning and use in a participant's own country and views on needs or wants on the ELTI course. To start things moving on these fronts, each interviewee was given a card with suggested discussion questions. (This rather artificial aid had been found more of a help than a hindrance in the piloting phase of this test with other similar overseas students.)
All interviewers had been oriented to Test 4 and trained in the use of its assessment criteria using a video-ed interview from the pilot phase. During the course of the interviews, each interviewer completed the specially designed information sheet for one of the interviewees. At the same time both interviewers assessed both interviewees on the 7 criteria shown in the scoring box in Section 3 above, scores on these scales to be weighted, balanced and aggregated individually before joint negotiation on an agreed score during the short break between interviews. An additional 10 points were allowed to each interviewer for an evaluation of the 'extra-linguistic' or 'social survival' impression made by each participant, this assessment also to be negotiated into an agreed joint score. This last, exploratory and more obviously subjective criterion was to tie in with my interest in individual cognitive/affective and social factors, though none of the interviewers considered it by any means irrelevant to the communicative competence they were assessing.

Within a few minutes of the end of each interview, then, the assessing was complete in a conveniently quantified form representing the agreed opinion of the rating teams who had, in the process of actively participating in an interaction with their new students, learnt some useful information about them. Post-test 4 followed a similar pattern but with the topics changed to fit its end-of-course timing. All pre- and post-test interviews were video-recorded to allow later checks on the progress of the students and the reliability of the raters (see Section 4 below). At the two reunions during Phase Two of my study, interviews were less formalised but were assessed through exactly the same criteria.
The following supporting data on Test 4 appear in Appendix 1.4:

1. Test 4 information sheets and participant cue cards.
2. Notes for assessors.
3. Criterion scales.

3.5 Test 5: Free Writing

A number of the key TL activities and problems for overseas students were, it will be remembered, identified in Chapters Two and Three as originating from writing modes. Participants have already been called upon to write in various ways in Tests 1, 2 and 3, but what they have not yet been asked to do is to write directly and extensively about their own situation. Test 5 fills this gap and in a way meeting genuinely inter-subjective demands in that we, as course designers and teachers, and they, as course participants, found meaningful. The pre- and post-test tasks asked the students to write about their plans for their UK stay based respectively on initial impressions and the experience of a couple of months here. At the reunions in Phase Two, the task was appropriately modified to allow for later events, the May 1981 account being suitably retrospective. Repeatability was thus designed into the task. In all versions the task was left very open, offering the maximum opportunity for individual decisions on interpretation of purpose, self-monitoring, creativity etc. Initial and later assessments used the full set of criterion scales (formal accuracy, referential adequacy, fluency, flexibility, socio-cultural appropriacy, compositional organisation and creativity) weighted and aggregated for an overall score.

Participants were allowed up to an hour to write what they wanted to, before and/or after their scheduled Test 4 interviews. Initial scoring
by a single rater was completed for the first Monday of the course after the Friday day of testing.

In Appendix 1.5, supporting data consist of:

1. Test 5 topics for all versions.
2. Criterion scales and weighting instructions for raters.

3.6 Continuous Assessment

The test days as a whole, particularly the initial one, can claim a kind of coherence which may well have added to their communicative validity. If you think of the test events in sequence, they make a global communicative sense that is worth more than the sum of their parts: in Test 1, participants are briefed on the content of their course; in Tests 2 and 3 they receive and work on information adding to their knowledge of their situation as overseas students in Britain; it is then fairly natural that they should talk and write about their own particular roles in this context. All this is particularly important when, as with the 1980 ELTI programme, the testing system is an essential part of the course design. It may well have been a contributing factor in the positive answer to the question of non-participant face validity, (identified in Figure 4.4 above). The pragmatic question of how 'outsiders' view a test battery was pragmatically answered in the case of the tests under discussion here; the battery was generally considered to have served its purpose in 1980 and was requested by ELTI for the equivalent course in 1981.

Those responsible for the design of both courses, however, were keen that the detailed 'before' and 'after' profiles that the test battery could provide should be bridged by a complementary system of
continuous assessment covering development during the programme.

And as my empirical study design in Chapter Five shows, the monitoring of progress (or even process) was to play a key role in the profiling of participants and in the validation of test data.

On the 1980 programme 13 index cards were provided, open to all teachers at all times for checks and comments on the progress and achievements of all participants. The first card in the set gave base line data on each participant (age, CI, future receiving institution, specialist area, level of training) along with results on the pre-test battery with an interpretation of them in terms of TL competence and study skill abilities. This card was updated at the end of the course with a parallel post-test profile. The cards on which teachers actually recorded their own comments were classified according to:

- the core evaluation criteria as used in assessing test performances (cards 1 to 5)
- study skills x study modes (cards 6 to 9)
- self-access, individual and project work (card 10)
- 'extra-linguistic' factors (card 11)
- results on set work (card 12)

Teachers had the freedom they were entitled to regarding when and what they wrote on the continuous assessment cards. At its best, the system produced revealing dialogues between teachers, as this excerpt of entries for Participant 0 shows:

"FLUENCY (Spoken)

Communicates with fluent confidence in spite of frequent formal errors (see above). Teacher A 13/8
Yes, because ambitious in what she wants to say and can't keep it all together yet. Teacher B 13/8
Will significantly improve. At the moment, frustrated. Teacher C 29/8."

The importance of this kind of teacher evaluation becomes especially
clear in the case study profiles discussed in Section 6 below.

4. **Statistical Validation (1): Reliability**

Now that we have some idea about what the tests are supposed to be testing we need to establish how much credence can be put on the measures of performance they produce. Without this, their contribution to descriptions of participants' proficiency and progress cannot be evaluated. In this section, therefore, the following issues will be explored:

1. Do measures on the various performance tests by different people at different times seem stable and sensible?

2. Are measures on the discrete-item competence test reliable?

As will have become clear from some of the discussion in Chapter Four, much of the classical testing theory of reliability, based as it is on objective, discrete-item, norm-referenced assessment, is irrelevant to criterion-referenced communicative performance tests. It would not be possible, for example, to check tasks such as those in Tests 1, 3, 4 or 5 for split-half reliability. A significant proportion of the reliability checking in this section, therefore, is devoted to the conventionally most difficult problem of global performance testing, that is stability in scoring by different raters of performance on the same or parallel tasks.

4.1 **Test 1**

The reliability of the assessment of Test 1 (note-taking from a lecture) was checked in the following ways:

1. Through the validation by teacher/assessors concerned, as participants at the actual lectures, of the scoring scheme for the test.
Modifications to the criterion of 'quality and quantity of information processed' were made individually then agreed jointly before scoring began.

2. Through a comparison of mean scores and standard deviations of three separate ratings of Test 1 from:

- the three raters working together (R1 in both tables in Figure 6.1)
- two individual raters working independently at leisure after the event (R2 and R3 in both tables)

3. Through Spearman rho rank-order correlations and Kendall W coefficients of concordance for all three ratings (see Table 2).

**Figure 6.1 (a): Table 1: Means, standard deviations and % scores on three ratings of the note-taking section of Pre- and Post-Test 1 with t-tests for significance of pre-/post-test differences.**
We are looking for evidence on the stability of Test 1 as a measure of note-taking performance in its alternate forms and of agreement, the extent to which different raters assign similar scores and ranks to the same individuals on the same test. Key points on these aspects of reliability emerging from Figure 6.1 and related analyses are now summarised.

1. Raters, whether working together or individually (R1 vs. R2 and R3) have tended to assign scores that are close in terms of central tendency (see $\bar{x}$'s in Table 1) and with similar dispersions (see standard deviations (SD)). This applies both to criterion 1, scored on points per acceptable item of information and criterion 2, measured on the three scales mentioned in Section 3 above.

2. Table 2 gives evidence that as well as scoring the group as a whole at similar levels, raters have tended to rank individuals within the group similarly. The range of rank-order correlations (.74 to .94) is significant at the $p<.001$ level. As Guilford and Fruchter (1978)
point out, the Spearman rho rank-order correlation formula is an appropriate substitute for the Pearson product-moment correlation (based on scores rather than ranks) when samples are, like mine, small and heterogeneous. The Spearman formula is conventionally agreed to give marginally lower coefficients than the Pearson (an assumption that I check out periodically in the analyses below). So the inter-rater reliability figures in Table 2 can be taken as encouraging.

Particularly interesting from a practical point of view is that R1 (the quick initial team rating to get scores for immediate programme placement and diagnostic purposes) appears no less reliable than R2 or R3 carried out by single raters at their leisure. Table 2 also gives the indices for Kendall’s coefficient of concordance, worked out from comparisons of the rankings on the three ratings on the test as a whole. The figures, at over .8 on pre- and post-test are corroboration of the high level of agreement across raters (see McIntosh 1976 p.87).

3. This high level of agreement does not seem to be achieved by the spurious effect of 'halo error', that is the tendency for raters to 'decide', on the basis of general impression or on the measure of one criterion, to assign equivalent scores on all criteria. Intra-rater correlations across criteria average .54 when the information criterion (criterion 1) is compared with the combined criteria of logic, text reduction skills and presentation (criteria 2, 3 and 4). When the three latter scales are analysed separately, 159 out of a total of 624 scores (pre- + post-test x 3 ratings x 4 scales x 26 participants) show raters giving the same scores to an individual on two or more criteria. This is certainly not a high enough proportion to confirm a halo error effect especially when you allow for the existence of valid overlap between criteria: 'logical division', for example,
sharing variance with 'presentation'.

4. The difference between pre- and post-test scores is insignificant on the note-taking test both statistically and, given the intervening training programme, educationally. It is most important to investigate in Sections 5 and 6 below, where the emphasis is on quantitative and qualitative validation and interpretation, whether this lack of gain is evidence of Test 1's invalidity (with validity sacrificed, perhaps, on the altar of reliability) or of the fact that the programme itself did not succeed in improving the set of skills measured.

The second part of Test 1 was the dictation. Scripts were checked three times and a consensus reached on the few cases where mechanical factors such as hand-writing or lay-out raised doubts about correctness. This was the only inter-rater reliability issue, given the totally objective scoring system (1 point from 50 for each formal error). Means and standard deviations on the dictations are given in Figure 6.2, the post-test statistics based on scores adjusted for the fact that the end-of-programme dictation was longer than the first one, as the Course Director had more to say.

<table>
<thead>
<tr>
<th>TEST 1(B)</th>
<th>PRE-TEST</th>
<th>POST-TEST</th>
<th>t-Test of Pre-Post</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\bar{x}$</td>
<td>5</td>
<td>%</td>
</tr>
<tr>
<td>DICTATION</td>
<td>(max.50)</td>
<td>SD</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>30.8</td>
<td>11.1</td>
<td>61.6</td>
</tr>
</tbody>
</table>

Figure 6.2: Means, standard deviations and % scores on the dictations in Pre- and Post-Test 1 with t-test for significance of difference.

Again the indices of central tendency and deviation for the two dictations suggest high test re-test stability. This time, however, there is evidence of significant gain over the period of the ELTI
course. When the focus switches to the question of validity in the next two sections of this chapter, it will be important to establish what it is that performance on a dictation test predicts, if it is considered a measure of competence rather than as a skill in its own right.

4.2 Test 2

As has already been explained, Test 2 is different in kind from the other tests in the battery in that Part A is an objective discrete-item test not attempting to re-create conditions of communicative performance and Part B is a cloze test, which may also be regarded as tapping competence rather than performance. Observations from both parts can be expected to contribute information to my profiles without some of the risks inherent in the scoring of global performance tasks. The selection of areas of focus and of individual items for Part A was discussed in 3.2 above. Here, various checks on reliability are made.

Guilford and Fruchter (1978) suggest that an initial check on the comparability of test-retest means, standard deviations and skewness is a reasonable guide to the overall reliability of a test (p.419). The table in Figure 6.3 and the frequency polygon in Figure 6.4 follow this advice. In their interpretation it should be remembered, of course, that Pre- and Post-Test 2A are not used as a check on stability in the classical statistical sense since a 'treatment' (ie the ELTI course itself) intervened. Thus gains in mean scores would be expected, though they are not inevitable even on an identical-form objective test as we have seen from pilot work with the 1979 group.
Figure 6.3: Table of means, standard deviations and average % for Pre- and Post-Test 2 Part A with t-tests for pre-:post-test significance of differences.

<table>
<thead>
<tr>
<th>SUB-TESTS</th>
<th>PRE-TEST 2</th>
<th>POST-TEST 2</th>
<th>t-test of Pre- &amp; Post-Test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>x</td>
<td>SD</td>
<td>% average</td>
</tr>
<tr>
<td>1 A1 Structure (max 15)</td>
<td>9.96 3.1</td>
<td>66</td>
<td>10.7</td>
</tr>
<tr>
<td>2 A2 Spoken use (10)</td>
<td>6.3 1.8</td>
<td>63</td>
<td>7.6</td>
</tr>
<tr>
<td>3 A3 Written use (10)</td>
<td>5.1 1.9</td>
<td>51</td>
<td>5.6</td>
</tr>
<tr>
<td>Total Test A (35)</td>
<td>21.33 5.4</td>
<td>61</td>
<td>24.0</td>
</tr>
</tbody>
</table>

Figure 6.4: Frequency polygon for the distribution of scores on Pre- and Post-Test 2 Part A.

The following points from Figures 6.3 and 6.4 are crucial on the issue of the test's reliability:

1. that the central tendency and dispersion of scores on all sub-tests as well as Part A as a whole are stable for pre- and post-tests.
that a statistically significant gain occurs between pre- and post-tests ($p < .01$)

that the most significant gain ($p < .005$) occurs on the spoken use sub-test; gains on the structure and written use sub-tests are not statistically significant.

that the frequency polygon shows the kind of overall stability and gain that we would hope to see on Part A of Test 2.

It should be remembered that we are talking about statistical facts here, though, not about educational values. For the latter kind of interpretation we must again wait until Sections 5 and 6 below.

The next series of quantitative checks on reliability are carried out through various item analysis operations. Oller's (1979;1) warning that item statistics alone cannot be used to select or reject items (p.199) has been heeded. I have already considered certain 'higher levels of validity' (op cit.) (see Chapter Four and the discussion of item selection in Section 3 above). And such considerations will be of continuing concern throughout the study. It is nevertheless necessary to examine how Test 2 Part A actually performed in practice at the level of individual items. However worthy the appeals to higher levels of validity, neither ELTI course nor my own research purposes could be served by a test where items performed unreliably or inexplicably.

Indices of item difficulty are conventionally expected to fall between .15 and .85 (percentage error) if items are not to be considered too easy or too difficult. Although this is a norm-referenced testing notion and even my Test 2, with its data base on recognised EAP
problems, is essentially criterion-referenced, we have seen enough of the actual relationship between the two approaches to test construction to appreciate that the latter should be informed by experience with the former (see Chapter Four, Section 6). It is important not to waste the diagnostic/predictive opportunity of a test with items that are too easy or too hard in terms of appropriate norms. In the pre-test, 30 of the 35 items in Part A had difficulty levels between .15 and .85. In the post-test 26 did so. Only in the post-test did any item come up 100% correct and even then it was only a single item. Figure 6.5 summarises the levels of difficulty for all Part A sub-tests and the test overall. It also, of course, provides evidence on which sub-tests became how much easier by the end of the course.

<table>
<thead>
<tr>
<th>TEST 2 (PART A)</th>
<th>Item Difficulty Indices (% error)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Structure</td>
</tr>
<tr>
<td>Pre-Test</td>
<td>x = 33.8</td>
</tr>
<tr>
<td></td>
<td>SD 17.5</td>
</tr>
<tr>
<td>Post-Test</td>
<td>x 28.9</td>
</tr>
<tr>
<td></td>
<td>SD 17.8</td>
</tr>
</tbody>
</table>

Figure 6.5: Means and standard deviations of item difficulty for Test 2 Part A, sub-tests and whole test.

But a general impression of a reasonable level of difficulty could be misleading unless items discriminate between more and less proficient participants as judged against the whole test or sub-test as criteria of what we are supposed to be measuring. As Oiler (1979;1) points out (p.248) information on item discrimination can be a check on both reliability and validity provided that the tests are in fact sampling the criterion. Flanagan's 1939 method for ascertaining item discrimination (ID) values prescribes a comparison of performance on each item between the top and lowest 27½% of testees. Given my small sample,
the advantage of using data from the middle group as well as the
groups above and below, and the chance fact that dividing the group
into thirds gave my three equal sub-groups of n=9, I based my ID
statistics on 33 1/3% sub-divisions rather than Flanagan's recommended
27½%. (I accept the implication of Guilford and Fruchter (1978 p.467)
that this minor adjustment makes no difference to the validity of the
ID values obtained.)

The figures summarised in Figure 6.6 are taken from an analysis of all
responses on all items in both Pre- and Post-Test 2, Part A. The
indices in the (a) cells in the table represent ID's worked out against
performance on Part A seen as one test; the (b) cells give the figures
with performances on the structure, spoken use and written use sub-tests
treated separately. It will be noticed that the (b) statistics are
rather higher than those in the (a) cells, which lends support to later
suggestions that each sub-test did in fact tap different areas of
competence where different participants showed different levels of
competence. Only the spoken use sub-test in its post-test
administration has rather marginal ID's. This is probably because
its level of difficulty was rather low (see Figure 6.5 above) by the
end of the pre-sessional course.

<table>
<thead>
<tr>
<th>TEST 2 (PART A) + SUB-TESTS</th>
<th>MEAN I.D. INDICES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-Test</td>
</tr>
<tr>
<td>1  Test A1 (Structure)</td>
<td>(a) .38</td>
</tr>
<tr>
<td></td>
<td>(b) .41</td>
</tr>
<tr>
<td>2  Test A2 (Spoken use)</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td>.37</td>
</tr>
<tr>
<td>3  Test A3 (Written use)</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>.41</td>
</tr>
<tr>
<td>4  Whole Test (Part A)</td>
<td>.35</td>
</tr>
<tr>
<td></td>
<td>.40</td>
</tr>
</tbody>
</table>

Figure 6.6: Table of mean item discrimination indices for Test 2, Part
A, sub-tests and whole test.
A level of .25 is sometimes regarded by testers as the lower limit of acceptability for an ID index. A figure of .35 (the mean of all the indices in Figure 6.6) signifies in my case, for example, an item where five of the best nine performers had the correct answer but only two of the nine lowest scorers did. In a total of 140 items analysed (35 questions x two uses of the test (pre- and post) x two analyses (see cells (a) and (b)), only 10 discriminate in a marginally deviant direction. These will be given special attention in the alternate form of Test 2 administered in Phase Two (see Chapter Eight) and in the analysis of the results of the 1981 ELTI course group.

The inclusion of the middle group in the analysis reveals that 86 out of 140 items show the 'ideal' discrimination pattern, that is with more of the top 1/3 of the group getting an item right than the middle 1/3 and more of the middle third succeeding than the low 1/3.

The quantitative analysis so far is encouraging. Means, distributions, difficulty and discrimination behave well under statistical scrutiny. A final useful check on individual items to see that they involved participants in a satisfactorily broad range of selection decisions is a distractor analysis. Were the distractors in Part A active and unambiguous? Did the lessons from the pilot group and the face validation by the native speaker informants bear fruit? In general terms, yes. In the pre-test, only 11 of the possible 105 distractors were not active and only 5 of these were also ignored in the post-test. On the post-test, 13 previously active distractors lost their power to distract more probably, we may assume, because of intervening TL acquisition than because they were inherently impotent. We could summarise by claiming a 90% positive 'distraction rate' on Pre-Test 2.
Part A, 77% on the Post-Test and 98% if the distractors are analysed with both administrations of the test taken together. Given the small sample, these rates are high. A comparable figure with the post-test used with the 1979 group and taking the first 35 multiple choice items (the same number of items as in my Test 2 Part A) was only 67% with a sample half as large again as mine.

Test 2 Part A as a reliable, quantitatively valid instrument based on considerations of construct and content validity already agreed, has now been checked in terms of test-retest and item discriminating validity. Findings indicate stability, with changes in means suggestive of mainly systematic rather than mainly non-systematic fluctuation. The reliability of the alternate form (Test 2A) will be discussed when I report on its use in Phase Two (see Chapter Eight, Section 3). The chief remaining classical statistical check on reliability, therefore, is the question of internal consistency. Here, though, it is first of all necessary to decide whether my test is amenable to this kind of check at all.

In some senses Test 2 Part A is a heterogeneous rather than a homogeneous test. Viewing the issue in Guilford and Fruchter's terms (1978), a heterogeneous test measures different variables in its different parts or items and is thus not necessarily looking for consistency across the scores on these parts or items. In such tests, the test-retest method of comparison would be the logical check on reliability. Stability rather than internal consistency would be the need. Since I tried to tap as broad a sample of the features indicated as problem areas or likely needs in as convenient and economic a way as possible, I deliberately excluded clusters of items testing
individual features. In this sense, then, Test 2 is heterogeneous and a procedure such as the split-half check on internal consistency would be very likely to produce a low reliability coefficient.

Yet, each of my sub-tests claims a homogeneity of focus and the test as a whole is to be used to help profile participants' communicative competence according to criteria more embracing than a series of discrete features. This being so, I cannot avoid all internal consistency responsibilities in the name of higher-order construct validity or of criterion over norm-referencing principles. Guilford and Fruchter make the telling point that the notion of internal consistency is closest to the intuitive interpretation of reliability. Whatever the logical heterogeneity of constructs, we expect discrete-item tests to say something about how close a participant's score at the time comes to his score on 'the perfect measurement instrument' - even if such reliability 'tells us nothing about the functional stability of persons or of tests' (1978 p.420).

If we were to hypothesise the kind of 'on-the-spot' reliability expected of a short test attempting to combine economic item heterogeneity with criteria such as sentence structure, spoken interactive and discourse cohesion/coherence competence (ie Parts A1, A2 and A3 in my Test 2), the expected internal consistencies might be:

quite high on A1, given the theoretical homogeneity of the grammatical competence concept (and the pragmatic advantage of its familiarity to practising test constructors)

can be low on Part A2, given the deliberate attempt to test a range of spoken interaction features and parameters

can be low on Part A3, given the variety of rhetorical functions and
information-processing skills involved in interacting with texts

quite high on Part A taken as a whole since it is seen, and to
be used, as a homogeneous test (of 'competence').

These somewhat informal hypotheses are best tested through Kuder-
Richardson (1939) formulae for estimates of internal consistency.
The table in Figure 6.7 shows Kuder-Richardson reliability co-
efficients (\(r_{tt}\)) for Pre- and Post-Test 2 Part A with the (a) cells
giving results from formula 20, which uses variances and covariances
of all items and the (b) cells giving results from formula 21 using
average right and wrong response proportions on items, the latter
formula producing conventionally rather lower coefficients.

| TABLE REDACTED DUE TO THIRD PARTY RIGHTS OR OTHER LEGAL ISSUES |

Figure 6.7: Kuder-Richardson coefficients of internal consistency for
Test 2, Part A, sub-tests and whole test.

In terms of internal consistency, the results are in general as hypoth-
esised, with the structure sub-test and the whole of Part A producing
respectable figures for internal consistency on both uses of the test.
On the pre-test, the spoken and written use sub-tests also perform as
predicted given the mixed homogeneous-construct/heterogeneous-item
nature of the tests. So does the written use test (A3) on the second
administration. In fact in all these cases the indices not only show
the predicted levels of consistency, they also show a consistent pattern
across pre- and post-testing uses. The one non-hypothesised result
is A2 (spoken use) in the post-test. Here the table actually shows negative coefficients from both formulae. The main explanation of these figures is, though, already implicit in previous analyses. The test-retest mean and difficulty figures in 6.3 and 6.5 above show that the gain on A2 is greater than on the other two sub-tests or on Part A as a whole. Figure 6.6 shows a related significant fall in item discrimination in A2. Since by the post-test most participants are performing well in this part of the test, the range of performance variation has become narrow, the variability over items low. This inevitably results in apparently poor internal consistency indices as each of the relatively few errors carries disproportionate statistical weight. It would be premature, however, to consider the sub-test unreliable on this evidence.

If we examine individual responses on pre- and post-tests in a way that is only possible with a small sample, and relevant because of the individual focus that this study is taking, it emerges, for example, that in spite of the many and varied TL events that intervened between the two test administrations, 27 out of 44 of the erroneous distractor selections by participants getting items wrong twice were the same in both cases. Participant S, for instance, performs with model consistency. In his pre-test he made 6 errors in A2; in the post-test 3 of these had been corrected but in his three remaining problem items, his erroneous distractor selection was the same as before. Apart from its interesting implications for learning and teaching, this not untypical behaviour suggests greater actual consistency in sub-test A2 than the statistics in 6.7 might lead one to infer.

Given the intentional heterogeneity of individual items in Test 2
Part A, its internal consistency is reassuringly high. On the test-retest and ID checks, where it should achieve acceptable statistics irrespective of the homogeneity/heterogeneity question, it has already been seen to perform adequately. We have reasonable reliability. The inter-test correlation analysis and interpretation in the succeeding sections will determine whether it has the complementary validity in terms of the criteria it is supposed to reflect.

My summary of reliability findings on the cloze test, Test 2 Part B will follow the pattern used for Part A with the difference, of course, that since the cloze passage responses were participant-selected not limited choice there can be no distractor analysis. Figures 6.8 and 6.9 below show the means, deviations and distributions of scores on the identical form pre- and post-tests:

<table>
<thead>
<tr>
<th>TEST 2 (PART B)</th>
<th>PRE-TEST</th>
<th>POST-TEST</th>
<th>t-Test of Pre-: Post-Test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Max 25</td>
<td>Z</td>
<td>Max 25</td>
</tr>
<tr>
<td>Cloze Passage</td>
<td>x</td>
<td>SD</td>
<td>Average</td>
</tr>
<tr>
<td></td>
<td>9.48</td>
<td>5.99</td>
<td>38</td>
</tr>
</tbody>
</table>

Figure 6.8: Means, deviations and average % on cloze tests (Test 2 Part B) with t-test for significance of difference between pre- and post-test means.

Figure 6.9: Frequency polygon for the distribution of scores on Pre- and Post-Test 2 Part B (Cloze).
Again the indications are that the cloze test has reasonable test-retest stability. Mean scores are significantly up ($p < .001$) at the end of the course, deviations and distributions stable. The scoring of the pre-test, however, revealed a key practical point that affected statistics on the cloze performances. Coming as the second part of Test 2, the cloze question caused timing problems for some students. This is not necessarily a matter of ecological invalidity (irrelevantly invalidating the task) since time, as we have seen in Chapter Two, has been found to be a real problem for overseas EAP students. Besides, there is no evidence in the post-test, after participants had had some training in working more quickly, that getting finished before the deadline was a problem. Statistically it is a different matter, however. Clearly if up to 30% of testees do not complete the last five items because time is up, figures on item difficulty, item discrimination and internal consistency are going to be affected. This fact will be taken into consideration in the interpretation of such statistics below and also in the correlational analysis and case study profiling in Sections 5 and 6. As regards the figures in 6.8 and 6.9 above, it is probably sufficient to say that some of the rather large gain in the cloze scores can be put down to the speed factor, a factor one feels is more relevant and legitimate in performance than in competence tests.

Figure 6.10 combines data on item difficulty and item discrimination for the cloze tests:
may be tapping some stable feature of language learning or use. Whether these quantitative indications have any qualitative validity is a question that will be pursued below.

Since Test 2 will be compared with other performance criteria as a single test as well as through its sub-tests, we need summarising statistics for Parts A and B combined to supplement the data already presented on the component parts. Figures 6.12 and 6.13 summarise means, deviations, % scores, gain and frequency distribution on the whole test, pre- and post-:

<table>
<thead>
<tr>
<th>TEST 2 PARTS</th>
<th>Overall Scores</th>
<th>Post-Test</th>
<th>t-Test of Pre-: Post-Test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pre-Test</td>
<td>Post-Test</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(max 60) SD Ave-rate</td>
<td>(max 60) SD Ave-rate</td>
<td>df t-val p =</td>
</tr>
<tr>
<td>A + B</td>
<td>30.8 10.4 51.3</td>
<td>38.3 9.8 63.8</td>
<td>26 3.88 .001</td>
</tr>
</tbody>
</table>

**Figure 6.12:** Means, deviations and average % for Pre- and Post-Test 2 (Parts A and B) with t-Tests for pre- and post-test significance of difference.

**Figure 6.13:** Frequency polygon for the distribution of scores on Pre- and Post-Test 2 (Whole Test).
Figure 6.10: Test 2 Part B (Cloze) Item Difficulty and Item Discrimination Statistics.

All four statistics here are acceptable (see the equivalent indices for Test 2 Part A above). Given the speed problem in the pre-test, the post-test figures are more valid statistically but the general message is encouraging anyway. This authentic cloze text scored on semantically and formally acceptable responses seems to have proved appropriate in terms of level, with items consistently (in fact in 10 of the cases over the two uses of the test) discriminating positively between participants ranking in the top and low thirds on the whole passage and in 60% of cases between top, middle and low thirds.

Figure 6.11 estimates the internal consistency of the cloze test again using the two Kuder-Richardson formulae in the (a) and (b) cells.

Again the pre-test coefficients are affected (in this case, inflated) by the fact that speed - as opposed to power-test conditions impinged. Nevertheless, the quite high figures for the post-test confirm interrier consistency and provide quantitative indications that the cloze test
The tables in 6.14 summarise item analysis data for statistics on item difficulty and item discrimination on Test 2 complete. The ID indices use the high and low thirds of the group ranked (for the first time) according to scores on Test 2 as a whole.

And finally, in Figure 6.15, the internal consistency of the whole Test 2 is estimated using the Kuder-Richardson formulae again, 20 in the (a) cells, 21 in the (b).

The statistics on Test 2 have been encouraging throughout but perhaps the most interesting figure of all is this high estimate of internal consistency for the test as a whole. The implication could be that we have a reliable test of something consistent, relatively homogenous. Whether this something validly represents the kind of underlying competence which is usefully predictive with overseas EAP students may be clarified when we compare Test 2 with other criteria below. The evidence of reliability is crucial, of course. The test could not be a valid test of anything without it.
4.3 Test 3

Test 3 (the reading and report-writing performance test) will be investigated in a way similar to that used for Test 1 to see if it shows evidence of the reliability that is essential if its validity as a criterion-referenced measure is to be meaningfully interpretable.

The table in Figure 6.16(a) (see overleaf) summarises performance on Test 3 by giving means, standard deviations and average percentage scores on all three sub-tasks in terms of the selected evaluation criteria, as well as for the test as a whole. The statistics are based on pre- and post-test scores (with the differences tested for significance) on two independent ratings. The first rating (R1) is the original course rating carried out so that results were available by the first teaching day. This rating was checked later to make sure there were no obvious marking errors. Although very few alterations were in fact made, all the component scores for R1 are means of the initial scoring and the marginally adjusted re-check. R1 is thus a 'double' rating. R2 is a later independent rating by a different assessor.

The table in Figure 6.16(b) estimates the statistical reliability of Test 3, in A using rank order correlation coefficients (Spearman rho) tested for significance, in B using Kendall's coefficient of concordance with a chi square significance test. The high Spearman rho correlation indices are given their promised computer check against the Pearson product-moment formula using scores rather than ranks. The coefficient thus obtained is .91 based on total Test 3 scores assigned by the two raters.
<table>
<thead>
<tr>
<th>TEST 3: CRITERIA</th>
<th>PRE-TEST</th>
<th>POST-TEST</th>
<th>t-Test of Pre-Post-Test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
<td>R1</td>
</tr>
<tr>
<td>Report A</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Information</td>
<td>Max</td>
<td>x</td>
<td>SD</td>
</tr>
<tr>
<td>2. Formal Accuracy</td>
<td>(6)</td>
<td>3.4</td>
<td>.9</td>
</tr>
<tr>
<td>3. Organisation</td>
<td>(6)</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td>4. Total</td>
<td>(42)</td>
<td>16.6</td>
<td>5.1</td>
</tr>
<tr>
<td>Report B</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Information</td>
<td>(18)</td>
<td>7.1</td>
<td>3.6</td>
</tr>
<tr>
<td>2. Formal Accuracy</td>
<td>(6)</td>
<td>2.8</td>
<td>1.0</td>
</tr>
<tr>
<td>3. Flexibility</td>
<td>(6)</td>
<td>2.7</td>
<td>1.2</td>
</tr>
<tr>
<td>4. Total</td>
<td>(30)</td>
<td>12.6</td>
<td>5.4</td>
</tr>
<tr>
<td>Report C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Information</td>
<td>(15)</td>
<td>7.7</td>
<td>3.6</td>
</tr>
<tr>
<td>2. Formal Accuracy</td>
<td>(6)</td>
<td>2.8</td>
<td>1.0</td>
</tr>
<tr>
<td>3. Flexibility</td>
<td>(6)</td>
<td>2.7</td>
<td>1.3</td>
</tr>
<tr>
<td>4. Organisation</td>
<td>(6)</td>
<td>1.4</td>
<td>.8</td>
</tr>
<tr>
<td>5. Total</td>
<td>(33)</td>
<td>14.7</td>
<td>5.9</td>
</tr>
<tr>
<td>TOTAL TEST 3</td>
<td>(105)</td>
<td>43.8</td>
<td>12.2</td>
</tr>
</tbody>
</table>

Figure 6.16(a): Comparative means, standard deviations and average % scores for two ratings on sub-tasks and whole test for Pre- and Post-Test 3 with t-test for significance of difference between pre- and post-test means (R1 and R2 combined).
### Estimates of Inter-Rater Reliability

<table>
<thead>
<tr>
<th>TEST 3: Reading and Report Writing</th>
<th>(A) Rank Order Correlation p&lt;sub&gt;RL:R2&lt;/sub&gt;</th>
<th>(B) Coefficient of Concordance p&lt;sub&gt;R1:R2&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>R Report A: 0.82</td>
<td>W = 0.94</td>
</tr>
<tr>
<td></td>
<td>E Report B: 0.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>T Whole Test 3: 0.88</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>S Report A: 0.78</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T Whole Test 3: 0.82</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>T Whole Test 3: 0.82</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.16(b): Table of rank-order correlations and concordance coefficients for two ratings of Pre- and Post-Test 3, tested for levels of significance.

The scatter diagram (Figure 6.17) graphically confirms that R1 and R2 assign scores very close to the regression line that would represent perfect agreement; it also confirms R2's tendency towards slightly lower scores.
Figures 6.16 to 6.17 summarise a great deal of information about Test 3 and will be referred to again later. Key points on the present issue of reliability are the following.

1. There is, in spite of the performance construct of the test, a level of agreement between raters that makes later interpretation of participants’ performances possible. But Rater 2 is consistently less generous in her evaluation of the information content of reports, sometimes significantly so, for example in Report A of the post-test.

2. The high level of inter-rater reliability does not appear
to come from halo error. In fact a close comparison of inter-
criterion scores by the same rater and inter-rater scores shows that
the former are always more highly differentiated than the latter (the
mean difference when all possible pairs of criterion scores, on formal
accuracy, flexibility and organisation, are compared on all sub-tests
of both uses of Test 3 is .93 or approximately one scale level). As
with Test 1, the indication is that overlap in scores across criteria
is a dimension of systematic shared variance rather than the spurious
assigning of 'the same score on everything'. There is clearly a
relationship between the 'quality and quantity of information'
processed from the data booklet and its organisation in report form.
This may well explain why Rater 2 consistently scores participants
higher on organisation than Rater 1, who is more generous with his
scores on information. R2 is more optimistic about how the students
handle the medium than the message. Deriving from this kind of
interplay between criteria, is one advantage of the analytic-criteria
approach to scoring performance tasks, namely that it allows a rater
to negotiate between criteria. When you are not quite sure you have
selected the appropriate level on one scale, you may be able to balance
out your rating by a compensatory or complementary selection on another,
related criterion scale.

3. The apparent lack of gain by the group on post- as compared
with pre-test raises reliability and other questions. Figure 6.16(a)
indicates that there is no statistically significant difference in
performance on any of the three reports except Report C, where the
difference is in the wrong direction. Overall the test behaves with
admirable alternate-form equivalence reliability. The only problem
is that six weeks of training have intervened. It will be a crucial
part of the concurrent and predictive validation analyses and interpretation in Sections 5 and 6 below to find out whether participants really did not get any better at this kind of reading and reporting or whether the test is invalid for construct, ecological, administrative or even statistical reasons.

4.4 Test 4

Once again the reliability of the test must be established before we can draw any valid inferences about participants' performances. In the case of Test 4 this is especially crucial given the history of reliability problems associated with attempts to assess aural/oral proficiency. The table in Figure 6.18 summarises pre- and post-test performance over different ratings in a form similar to that already used for Tests 1 to 3.

<table>
<thead>
<tr>
<th>TEST 4: ORAL INTERACTION: OVERALL SCORES</th>
<th>PRE-TEST</th>
<th>POST-TEST</th>
<th>t-test of Pre-Test/Post-Test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1  R2   R3  Aver.</td>
<td>R1  R2  R3  R4  Aver.</td>
<td>df  t-val  P</td>
</tr>
<tr>
<td>Max=100 x</td>
<td>53.6 56.7 57.5 56</td>
<td>63.5 62.6 61.9 62.1 63</td>
<td>24 5.83  .001</td>
</tr>
<tr>
<td>SD 16.3 14.6 12.7 13.8</td>
<td>13.3 13.2 13.3 14.0 12.5</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.18: Table showing means, standard deviations and average scores on individual and combined ratings for Pre- and Post-Test 4 with t-test for significance of pre-/post-test difference.

The three pre-test ratings cover various rating conditions: R1 is the 'live' rating arrived at by the two pairs of assessor/participants at the time of the actual interviews. R2 and R3 are independent ratings by two individual trained raters from the video-recordings of the interviews, the rater for R2 having already taken part in the paired assessment of half the live interviews, the rater in R3 not having done so. The point of these variations in rating conditions is not to test hypotheses about condition x versus condition y but rather to
investigate the general evaluational robustness of the Test 4 task and its criteria.

For the Post-Test, four ratings were obtained, R1 again being the original 'live' assessment. R2, R3 and R4 are this time the negotiated and the individual ratings of two raters not involved in the actual interviews, Rater 3 being the same trained rater/teacher as in the Pre-Test, Rater 4 an independent outsider, ELT-trained, and oriented on Test 4 with the pilot training video.

The frequency polygon, Figure 6.19, is again used to give the general picture of pre-post-test comparison; and is based on participants' scores averaged from all ratings.

```
F 10
R 9
E 8 / = Pre-Test
Q 7 / = Post-Test
U 6
E 5
N 4
C 3
I 2
E 1
S
```

**Figure 6.19:** Frequency polygon for the distribution of scores on Pre- and Post-Test 4.

Figure 6.20 also follows the precedent we have established by focusing on the inter-rater agreement aspect of reliability using Spearman rho
rank-order correlations from all pairs of ratings of the test and the Kendall formula for concordance which combines sets of rankings for each participant into a single index of agreement.

<table>
<thead>
<tr>
<th>TEST 4</th>
<th>Estimates of Inter-Rater Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(A)</td>
</tr>
<tr>
<td></td>
<td>R/O Correlation</td>
</tr>
<tr>
<td>Ratings</td>
<td>Spearman rho</td>
</tr>
<tr>
<td>P</td>
<td>R1:R2</td>
</tr>
<tr>
<td>E</td>
<td>R1:R3</td>
</tr>
<tr>
<td>E</td>
<td>R2:R3</td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
<tr>
<td>P</td>
<td>R1:R2</td>
</tr>
<tr>
<td>O</td>
<td>R1:R3</td>
</tr>
<tr>
<td>S</td>
<td>R1:R4</td>
</tr>
<tr>
<td>T</td>
<td>R2:R3</td>
</tr>
<tr>
<td>E</td>
<td>R2:R4</td>
</tr>
<tr>
<td>S</td>
<td>R3:R4</td>
</tr>
<tr>
<td>T</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6.20: Estimates of inter-rater reliability for Pre- and Post-Test 4.

A further check on inter-rater reliability was carried out on a subset (n=16) of the participants on the Pre-Test. Although I had reasonable reliability figures from the four assessors involved in the actual interviews and from two additional ratings, I was interested to know whether evaluation of the task would be significantly affected by exceptional expertise and experience in the field of testing (compared with the more normal level that characterised the other raters). So, two well-known ELT testers were invited to assess 8 of the 13 interviews; individually then reaching an agreed joint rating for further comparison. The two testers were encouraged to use my criterion scales as they felt appropriate, rather less strictly, in other words, than the course raters. In this way, I felt I would get completely
independent assessments combining the essence of my scoring scheme and
the special expertise of the two testers and, of course, providing
invaluable extra feedback on the test.

The correlation matrix in Figure 6.21 summarises the results of this
experiment. Rank-order correlations between 'insiders' and 'outsiders'
are encouragingly high (see (A), the top half of the matrix) though
the outside raters tend to award higher overall scores (see (B), the
bottom half of the matrix).

### TEST 4 (PRE-TEST) CORRELATION MATRIX

<table>
<thead>
<tr>
<th></th>
<th>(A) Spearman rho</th>
<th>(B) Mean Z score differences</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R1</td>
<td>R2</td>
</tr>
<tr>
<td>R1 (Mixed, Original)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2 (Individual)</td>
<td></td>
<td>3.7</td>
</tr>
<tr>
<td>R3 (Individual)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4 (External Individual)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R5 (External Individual)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R6 (External Combined)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R1-6 (Mean Rating)</td>
<td></td>
<td>8.2</td>
</tr>
</tbody>
</table>

Figure 6.21: Matrix showing inter-rater rank-order correlations and
mean % score differences.

So far there is evidence that different raters rank participants on
Test 4 in similar orders, assigning similar scores and distributions.
What has not yet been checked is the extent to which a global score
assigned by one rating to a particular participant is likely to be made
up of scores on the constituent criteria scales that correlate closely
with those assigned to the same participant on the same test by a
different rater, yet without the close agreement being caused by halo
error. For illumination on this question it was necessary to analyse
all criteria ratings by all raters on all participants. With this
set of data in front of us, it is possible, for example, to see how
typical participant H's ratings on the Post-Test were. On R1, R2 and
R3 he is given an overall score of 55%. The three ratings also agree
entirely on his listening comprehension, 'accent', formal accuracy,
referential adequacy, fluency and 'extra-linguistic' (or 'social
survival') performance. Reading off the levels on the criterion scales
(see Appendix 1.4) on which all three ratings agree, H's communicative
competence would be described as follows:

"Understands intended communication at normal speed given the need
for occasional repetitions and/or rephrasings. Pronunciation, stress
and intonation errors require concentrated listening but only
occasional misunderstanding is caused or repetition required.
Quite frequent grammatical inaccuracies show some major patterns
not under control. Occasional breakdowns in communication caused.
Vocabulary inadequacies hamper a significant part of the intended
communication. Fairly frequent lexical inaccuracies.
Utterances fairly slow, hesitant and uneven. Some utterances
incomplete but some suitably inter-connected."

'Genial, self-possessed' (R1); 'relaxed and outgoing' (R2); 'confident'
(R3).

On the socio-cultural appropriacy scale, however, R1 notes only
'occasional', insignificant errors in the rules of use whereas R2 and
R3 select the level suggesting such errors are 'frequent' and
'significant enough to cause occasional misunderstanding'. Perhaps the
R1 raters felt H's problem stemmed from a lower level of flexibility;
they rate H at level 3 here, i.e. 'quite frequently thrown by changes of
topic' and 'sometimes unsuccessfully initiating new topics', whereas
at R2 and R3, H's relatively low socio-cultural performance is
compensated for by the fact that he 'usually adapts successfully to changes of topic' (level 4 on my flexibility scale).

Now, the extremely close agreements not only on overall scores but also on individual criteria found with participant H are not in fact typical of all the scores on Test 4. If they were, it would almost certainly mean that my system of analytic criterion scales had turned what was supposed to be an integrative test into a discrete-point one (see Chapter Four, Section 6). However, the general pattern of the assessments of H, as interpreted from a matrix showing all criterion scale levels as selected by all raters for all participants, does help answer my questions about inter-criterion correlations. If different raters were giving similar global scores but with widely differing ratings on the criterion scales, something would probably be wrong. Yet of the 208 sets of ratings (8 criteria x 26 participants) on the Pre-Test and 216 (8 criteria x 27 participants) on the Post-Test, only 18 and 23 respectively show a difference of more than one scale level across raters, compared to 48 and 46 where, as with most of the levels selected for H, all raters agree on a criterion level for a participant and 142 and 147 where the difference of level assigned is no more than one. Different raters do, therefore, tend to agree quite closely on individual criterion ratings for participants.

But what about the second question? Is this apparent agreement caused by the halo effect? To investigate this, deviations on all ratings on all criteria were computed. Out of 78 deviation statistics for Pre-Test 4 (26 participants x 3 ratings) and 108 for Post-Test 4 (27 participants x 4 ratings), there were only three cases where the same criterion level was assigned to a participant on all criteria. The
average deviations across criteria for all ratings ranged from .53 to .6, which indicates that there is a tendency for participants' performance as assessed on different criteria to converge but not to an extent that suggests insignificant variability or disinclination on the part of raters to differentiate. Nor is the variability consistent among participants. In fact, an interesting side issue emerging from this analysis is that certain participants are assessed as varying much more significantly across criteria than others, by all raters. This is important for the interpretation of TL profiles. The whole question of criterion variability is, of course, germane to the unitary vs. discrete vs. partial discrete competence hypotheses discussed in Chapter Four, Section 6.

A final summary of key points on the reliability of Test 4, then.

1. In spite of the global performance nature of the task and variations in rating conditions, measures on both pre- and post-tests are close in terms of percentage scores, criterion scale scores and distribution.

2. Course ratings correlate healthily with 'expert' ratings in rank-order terms, though the experts in my experiment tended to award higher overall scores.

3. The variations and similarities in criterion scores assigned by different rates intra- and inter-subjectively are more in line with the actual relatedness of criteria than with halo error. Some raters do assign rather more similar ratings across criteria than others, but significantly often, individual participants are assigned quite widely differing scores on different criteria. Just as the objective competence Test 2 seems to combine heterogeneity and homogeneity, so
apparently does the communicative performance tapped in Test 4.

4. Overall pre-:post-test comparison shows a statistically significant gain by the group (p=.001). Whether this gain is so significant educationally will be investigated when the focus switches from reliability to validity below.

Test 4 seems reliable. This is an essential first step in the direction of valid profiling usefulness.

4.5 Test 5

The reliable assessment of 'free' composition has traditionally raised almost as many problems as reliability in the assessment of oral interaction. The check on results on Test 5, therefore, should be thorough and searching. It will initially follow the pattern of investigation used with the other three performance tests: then, as with Test 4, extra questions will be asked.

The table in Figure 6.22 summarises scores assigned on three different ratings. R1, for both pre- and post-tests, is the initial rating made for immediate programme purposes. R2 a later at leisure rating by the same assessor and R3 an independent outsider rating using my criteria. As usual, the table also shows the statistical significance of the pre-test:post-test difference in overall mean scores, ie those represented by averaging participants' scores over the three ratings.
<table>
<thead>
<tr>
<th>CRITERIA</th>
<th>Pre-Test Max</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>Aver</th>
<th>%</th>
<th>Post-Test R1</th>
<th>R2</th>
<th>R3</th>
<th>Aver</th>
<th>%</th>
<th>t-test of Pre-/Post-Test Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formal Accuracy</td>
<td>12</td>
<td>5.7</td>
<td>6.4</td>
<td>6.0</td>
<td>6.0</td>
<td></td>
<td>7.0</td>
<td>7.3</td>
<td>6.3</td>
<td>6.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Referential Adequacy</td>
<td>12</td>
<td>6.3</td>
<td>6.6</td>
<td>6.7</td>
<td>6.5</td>
<td></td>
<td>7.3</td>
<td>7.4</td>
<td>8.0</td>
<td>7.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Fluency</td>
<td>12</td>
<td>6.4</td>
<td>6.6</td>
<td>6.0</td>
<td>6.3</td>
<td></td>
<td>7.4</td>
<td>7.3</td>
<td>7.0</td>
<td>7.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Flexibility</td>
<td>12</td>
<td>6.3</td>
<td>6.1</td>
<td>5.6</td>
<td>6.0</td>
<td></td>
<td>7.5</td>
<td>7.2</td>
<td>7.4</td>
<td>7.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Socio-Cultural</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Appropriacy</td>
<td>6</td>
<td>3.2</td>
<td>3.7</td>
<td>2.8</td>
<td>3.2</td>
<td></td>
<td>3.8</td>
<td>3.7</td>
<td>3.8</td>
<td>3.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Compositional</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organisation</td>
<td>12</td>
<td>5.0</td>
<td>5.3</td>
<td>5.8</td>
<td>5.4</td>
<td></td>
<td>5.9</td>
<td>6.5</td>
<td>5.9</td>
<td>6.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Creativity</td>
<td>6</td>
<td>2.1</td>
<td>1.9</td>
<td>2.6</td>
<td>2.2</td>
<td></td>
<td>2.2</td>
<td>2.0</td>
<td>1.4</td>
<td>1.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTALS</td>
<td>72</td>
<td>35.0</td>
<td>36.6</td>
<td>35.5</td>
<td>35.6</td>
<td>49.4%</td>
<td>41.1</td>
<td>41.4</td>
<td>39.8</td>
<td>40.9</td>
<td>56.8%</td>
<td>26</td>
</tr>
</tbody>
</table>

**Figure 6.22:** Table of mean scores on all criteria on three separate ratings of Pre- and Post-Test 5 with average total scores, percentages and pre-:post-test gain.
The frequency polygon in 6.23 adds, graphically, summary information on pre- and post-test distribution and gain.

![Frequency polygon](image)

Figure 6.23: Frequency polygon for the distribution of scores on Pre- and Post-Test 5.

Figure 6.24 gives the results of the usual check on inter-rater reliability in terms of Spearman rho and Kendall's coefficient of concordance, both tested for significance level.

<table>
<thead>
<tr>
<th>TEST 5: FREE WRITING</th>
<th>INTER-RATER RELIABILITY STATISTICS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ratings</td>
</tr>
<tr>
<td>1. Pre-Test</td>
<td>R1:R2</td>
</tr>
<tr>
<td></td>
<td>R1:R3</td>
</tr>
<tr>
<td></td>
<td>R2:R3</td>
</tr>
<tr>
<td></td>
<td>R1:R2:R3</td>
</tr>
<tr>
<td>2. Post-Test</td>
<td>R1:R2</td>
</tr>
<tr>
<td></td>
<td>R1:R3</td>
</tr>
<tr>
<td></td>
<td>R1:R3</td>
</tr>
<tr>
<td></td>
<td>R1:R2:R3</td>
</tr>
</tbody>
</table>

Figure 6.24: Table of rank-order correlations and concordance coefficients for three ratings of Test 5, tested for levels of significance.
In Figure 6.25 the results of extra checks on the reliability of the essay test are shown. It was possible that the reasonable agreement among ratings was merely a function of the analytic criteria themselves. For this reason the essays of both pre- and post-tests were re-rated by other assessors under three different conditions. The first rater (a different qualified and experienced outside assessor for pre- and post-tests) divided participants into three groups, 'good', 'average' and 'poor', glossed by the pre-test assessor as follows:

"Good = sorted out from average re fluency, range of expression. Poor = sorted out from average re lack of ability to communicate basic notions, paucity of content, gross grammatical inaccuracies."

The second extra assessment (by the same qualified teacher each time) was through the 'three-pile' technique, eg separating the essays into three equal groups (high, mid, low) by an impressionistic, global evaluation.

In the third supplementary check, the rater (not the same person on the two occasions) was asked to assign essays to five sets, A to E or, broadly speaking, very good, good, quite good, poor, very poor, but with no constraints on how many should be assigned to each category. As Figure 6.25 shows, the pre-test A to E rater decided that only the latter three labels applied and the post-test rater, under the same conditions, assigned only one participant to the 'very good' category. Now, the contingency table, with statistical explanations to follow it.
### Figure 6.25: Table comparing rank-orders assigned on the original analytic criteria for Test 5 with the same essays grouped independently by global assessments.

* indicates no mean rank-order possible as only one essay assigned to category.

The figures in each cell are the mean rank-orders on the Test ratings by my criteria of the participants assigned to the categories used by the extra assessors. Thus the rising mean figures through 'good', 'average', 'poor'; 'high', 'mid', 'low' or A to E are in the right direction. The steeper the rise, in fact, the greater the agreement between R1, R2 and R3 and the global assessors.

The following key points emerge from all these statistics and some further analysis.

1. Measures on both pre- and post-test are close across raters with regard to overall percentages, performance on each criterion, rankings and distribution.

2. Inter-rater reliability is reasonable, though not as high as on performance Tests 1 or 3. This is probably because the most
objectively scorable information criterion is missing, as it should be when students are invited to write freely.

3. Comparisons between analytic scoring on my criteria and more global assessments are positive. This suggests that the required diagnostic and profiling potential of the analytic-criterion approach is not achieved at the expense of more impressionistic (but diagnostically less useful) validity.

4. But Test 5 criterion raters do not appear to have arbitrarily converted an impressionistic evaluation into a quasi analytic one. A detailed analysis of all criterion scores for all participants by all raters again reveals the kind of pattern of systematic relationships between criteria that emerged from the assessments of Tests 1, 3 and 4 above. Of a possible 567 pairs of criterion scores (27 participants x 21 criterion combinations) the range of identical ratings assigned to individual participants is between 20 and 30% by all raters on both tests. In terms of Campbell and Fiske's (1959) notion of convergent/discriminant validation, consistent relationships or non-relationships between criteria as revealed by the analysis are interesting. All raters, for example, perceive a relationship between referential adequacy and fluency and between referential adequacy and socio-cultural appropriacy. All but Rater 3 on the Post-Test connect formal accuracy and referential adequacy in their assessments. But then Rater 3 is clearer in her differentiation between syntax and lexis (see R3 in 6.22 above); there is a certain logic, therefore, in the fact that she finds a relationship between referential adequacy and flexibility, between formal accuracy and compositional organisation. Such suggested relationships are probably the kind only noticeable when you are dealing with a small sample and looking for individual
variability. They are certainly sources of convergent validation, that is, evidence on the existence of the same measurable construct from different methods or (as in this case) different assessors.

The analysis also provides evidence of discriminant validity, in the cases where there is little or no evidence of relationships between criteria. No rater, for example, suggests connections between referential adequacy and creativity or, perhaps more surprisingly, that flexibility and creativity are linked. For Campbell and Fiske this would be at least a tentative indication that the three criteria exist as reasonably independent traits.

The ease with which reliability investigations can lead into the realm of validity foreshadows Section 5 now, where the question of what the tests mean is taken up in earnest.

5. **Statistical Validation (2): Validity**

An analysis of the correlations between performances on the various tests in the battery is a logical transition from an emphasis on reliability to an emphasis on validity. Its focus on relationships between performances should clarify what TL communicative abilities the tests actually require and measure.

The correlation matrix in Figure 6.26 gives correlation coefficients between all pairs of tests and the designedly most autonomous sub-tests in the pre- and post-batteries. Spearman rho rank-order correlations are used throughout though, as in Section 4 above, periodic cross-checks with Pearson product-moment figures will be made to make sure findings would not differ if this computationally less convenient formula were used.
### POST-TESTS

<table>
<thead>
<tr>
<th>Pre-Tests</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>P Test 1</td>
<td>.68</td>
<td>.61</td>
<td>.71</td>
<td>.73</td>
<td>(.89)</td>
<td>(.91)</td>
<td>.64</td>
<td>.64</td>
<td>.89</td>
</tr>
<tr>
<td>Test 2</td>
<td>.61</td>
<td>.44*</td>
<td>.58</td>
<td>.65</td>
<td>.54</td>
<td>.69</td>
<td>(.95)</td>
<td>(.93)</td>
<td>.80</td>
</tr>
<tr>
<td>R Test 3</td>
<td>.51</td>
<td>.54</td>
<td>.64</td>
<td>.57</td>
<td>.37*</td>
<td>.64</td>
<td>.43*</td>
<td>.37*</td>
<td>.69</td>
</tr>
<tr>
<td>Test 4</td>
<td>.35*</td>
<td>.32*</td>
<td>.34*</td>
<td>.69</td>
<td>.49*</td>
<td>.80</td>
<td>.51</td>
<td>.41*</td>
<td>.82</td>
</tr>
<tr>
<td>E Test 5</td>
<td>.66</td>
<td>.66</td>
<td>.68</td>
<td>.67</td>
<td>.56</td>
<td>.75</td>
<td>.60</td>
<td>.57</td>
<td>.84</td>
</tr>
<tr>
<td>Test 1</td>
<td>(.85)</td>
<td>.49</td>
<td>.46*</td>
<td>.17**</td>
<td>.37</td>
<td>.66</td>
<td>.64</td>
<td>.54</td>
<td>.69</td>
</tr>
<tr>
<td>T Notes</td>
<td>.55</td>
<td>.39*</td>
<td>.45*</td>
<td>.74</td>
<td>.84</td>
<td>.67</td>
<td>.61</td>
<td>.93</td>
<td></td>
</tr>
<tr>
<td>E Dictation</td>
<td>(.91)</td>
<td>.55</td>
<td>.39*</td>
<td>.45*</td>
<td>.74</td>
<td>.84</td>
<td>.67</td>
<td>.61</td>
<td>.93</td>
</tr>
<tr>
<td>S Test 2</td>
<td>.53</td>
<td>(.85)</td>
<td>.65</td>
<td>.29**</td>
<td>.53</td>
<td>.37*</td>
<td>.50</td>
<td>.62</td>
<td>.78</td>
</tr>
<tr>
<td>Part A</td>
<td>(.93)</td>
<td>.47*</td>
<td>.26**</td>
<td>.63</td>
<td>.51</td>
<td>.45*</td>
<td>.62</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>T Test 2</td>
<td>.55</td>
<td>(.93)</td>
<td>.47*</td>
<td>.26**</td>
<td>.63</td>
<td>.51</td>
<td>.45*</td>
<td>.62</td>
<td>.73</td>
</tr>
<tr>
<td>Cloze</td>
<td>(.85)</td>
<td>.65</td>
<td>.29**</td>
<td>.53</td>
<td>.37*</td>
<td>.50</td>
<td>.62</td>
<td>.78</td>
<td></td>
</tr>
<tr>
<td>S All Tests</td>
<td>.77</td>
<td>.80</td>
<td>.69</td>
<td>.82</td>
<td>.85</td>
<td>.54</td>
<td>.76</td>
<td>.70</td>
<td>.74</td>
</tr>
</tbody>
</table>

The correlation matrix summarises a considerable amount of information. The following are key points on our present concern of how my test data can validly inform profiles of participants' performance.

1. All the correlations on the matrix are positive, the majority (un-asterisked) at p < .01 or above, a few (*) at p < .1 and three (**) not statistically significant. The part:whole correlations (eg Test 1 notes: Test 1 complete) are bracketed as a reminder that the coefficients concerned are by definition inflated. The general message
from the predominantly moderate, substantial or high correlation indices is that the various competence and performance tests in the batteries tap overlapping combinations of knowledge and abilities, some of the tasks being more obviously closely related than others. Given the fact that all four performance tests and the 'competence' test (Test 2) taken as a whole were designed as integrative rather than discrete-item tasks, the level of positive correlation emerging here is *prima facie* evidence of concurrent validity. In convergent/discriminant validation terms, the substantial correlations between the writing test (Test 5) and the reading and report-writing test (Test 3), at .68 and .57, are evidence of convergent validity, the same trait ('writing') influencing the correlation between two different tests measuring (among other things) the same thing. The relatively low correlations between Test 4 (oral interaction) and the cloze test (Test 2, Part B) can be seen as discriminant validation, a lesser relationship, at .26** and .41*, because of the greater independence of the two sets of abilities measured.

2. Consistently the highest correlations on the matrix are between individual tests and the rank-order of participants on the battery as a whole. Given that 'each part contributes the same amount of variance to the total .... so that its proportion of contribution is relatively small' (Guilford and Fruchter 1978 p.331), the coefficients in these cases are not unduly inflated. The implication is still that each test is a fairly good predictor of concurrent overall performance. None of the tests shows immediate signs of contributing irrelevantly or disruptively to profiles of individuals or of the group.

3. The competence-oriented test (Test 2) emerges as a consistently high predictor of overall performance, though on the Pre-Test it
is no better at predicting oral interview performance than Tests 1 or 3 are and in the Post-Test correlates unimpressively with Test 3. We have already seen evidence of the reliability and economy of Test 2 as well as the diagnostic potential it possesses because of its heterogeneity of coverage. Note that Test 2 as a whole is a better predictor all round than either of its two parts. Also that Part A (the structure, spoken and written use section) is a rather more consistent predictor than Part B (the cloze passage), though this may be partly explained by the time factor which, statistically at least, distorted pre-test cloze results. Certainly Test 2 correlates positively enough with the performance tests for the competence:performance relationship in individual profiles to bear fruitful scrutiny in Section 6 below.

4. A further scan of the matrix for high correlations brings additional insights into concurrent validity. Test 5 (free writing) correlates strongly enough with all other tests, as well as with overall performance, to suggest that the task did tap the wide range of linguistic, cognitive and expressive skills that it should have done. It also suggests that the evaluational criteria used and the way they were interpreted by raters, made the test a valid communicative measure as well as a reliable one. Test 1 also shows the substantial correlations with other tests that we would expect it to, given the range of communicative activities it involved. The only inconsistency here is that pre- and post-test correlations between Test 1 and Test 4 differ so significantly (.35 vs. .71). This hints at a validity problem. It will be investigated immediately below.

5. A search for low correlation figures provides insight into validity, either in a positive discriminant validation sense or
negatively if it suggests the improbable. Pre-test 4 shows low correlations with all other tests except free writing yet only one of these low correlations (with the cloze test) is paralleled in the post-test battery. It is feasible to envisage a weak relationship between interaction in an interview and cloze-test completion (which is most traditionally associated with reading ability but (see Davies 1978 pp.128/219) which is still of uncertain status with regard to what it does measure and may even most closely relate, as Alderson (1978) suggests, to formal linguistic skills). But it is rather more counter-intuitive that participants' aural/oral performance should not correlate more highly with listening, note-taking and dictation (Test 1) and, to a lesser extent perhaps, with the test of underlying competence (Test 2). The most likely explanation of erratic performance by some participants on Pre-Test 4 is the ecological validity problems caused by facing a video-recorded interview so soon after their arrival in the UK. It is in this most 'public' of the performance tasks that arrival shock is most likely to tell. In fact, its effect in some of the interviews was remarked on by the outside expert testers. Individuals' Pre-Test 4 performance must therefore be viewed with this factor in mind.

We have a statistical reminder of a pragmatic reality here, which is typical of the way this kind of study has to relate psychometric fact and real-life value judgments. But such a balancing act would not, in this case, lead us to declare Pre-Test 4 invalid. In a multi-dimensional framework such as this, shyness and shock are real communicative factors for learners in a new culture (see Chapter Eight below) and if participants' varying ability to overcome them explains some of the gains made between Pre- and Post-Test 4 performance, we are acknowledging a relevant ability.
6. **Test 3**, as might be expected in the light of findings from the reliability investigations in Section 4 above, also seems asterisk prone. In both pre- and post-test versions it correlates lowest of all the complete tests with overall performance on the batteries. It also correlates with significant inconsistency with Test 4 (.34 pre, and .64 post) and with dictation (.39 and .64 respectively). Figure 6.27 below is taken straight from the computer's answer to my request for corroboration of just how low the shared variance between Pre-Test 3 and Pre-Test 4 was, as well as for the usual check that Pearson r statistics would not significantly change the picture given by Spearman rho. The answer to the first question from an analysis of variance (adjusted for degrees of freedom) is that there is just 13.8% of shared variance between the pre-tests of reading/reporting (Test 3) and participating in an interview (Test 4). The answer to the second question is that the Pearson correlation is .42, the expected slightly higher figure than the Spearman .34. Figure 6.27 itself is the relevant scatter diagram for the two tests, graphically confirming that those who perform well or badly on the intensive reading and reporting task are not all that likely to be the same individuals as those doing well or badly in the interview. But, as the competence measures will constantly remind us, the correlations are still positive rather than negative; there is still a TL threshold below which individual performances on any test will correlate very highly because low linguistic competence will ensure they are poor. As a group, my participants were not below that level, so the fairly broad scatter in Figure 6.27 is suggesting a genuine difference between the skills tapped in these two tests.
Scores on Pre-Test 4

Figure 6.27: Scatter diagram of scores on Pre-Test 3 and Pre-Test 4

The relevant correlation figures in 6.26 above mean we can also assume this kind of regression pattern between Test 3 and the dictation. Test 3 was given a further validation-seeking check against participants pre- and post-course performance on the SRA Reading for Understanding (RFU) Senior Placement Test (Thurstone 1965), relevantly administrable on the ELTI programme as the SRA RFU Kit was one of the self-access options. Correlations between Test 3 and the RFU Test were inconsistent; .73 at the beginning but .44 on post-tests, when the groups average gain in RFU performance was significant at p< .01 and where pre- post-test correlation was .84. Yet it would be premature to claim falsification for a hypothesis that Test 3 is a valid predictor of communicative performance on any of this evidence. We already suspect that Pre-Test 4 performance was affected by the 'shock' factor; the
dictation test is similarly a more reliable predictor of overall performance at the end than at the beginning of the course (see below) and the RFU test itself has been heavily criticised for its emphasis on vocabulary to the neglect of broader skills such as inference, interpretation and comparison (Turnbull 1968), which are all crucial in my Test 3, and for 'yielding scores of indeterminate reliability and suspect validity' (Turnbull 1968 again, p.178). Besides, Test 3 performs consistently and positively in relation to Test 1 and the very reliable Test 5, both of which share with it the 'broader skills' of note-taking and writing respectively. Its correlation with 'all tests' is also consistent (at .69) on both batteries. When the focus is on prognostic validity, then on case-study profiling below, Test 3 needs further special attention.

Where Figure 6.26 gives the general picture of how my tests correlated concurrently in two separate administrations, Figure 6.28 below summarises the prognostic power of the pre-tests for performances at the end of Phase One. It ignores correlations between part pre-tests and the post-tests containing those parts as likely to be statistically distorted as well as some of those between whole pre-tests and part post-tests; (to see how well the listening and note-taking test (Test 1) predicts later scores on dictation, for example, seems trivial).

<table>
<thead>
<tr>
<th>P</th>
<th>Pre-Test 1</th>
<th>Pre-Test 2</th>
<th>Pre-Test 3</th>
<th>Pre-Test 4</th>
<th>Pre-Test 5</th>
<th>(Dict)</th>
<th>(A) (Cloze)</th>
</tr>
</thead>
<tbody>
<tr>
<td>P</td>
<td>Pre-Test 1</td>
<td>.67</td>
<td>.63</td>
<td>.65</td>
<td>.42*</td>
<td>.43*</td>
<td>-</td>
</tr>
<tr>
<td>R</td>
<td>Pre-Test 2</td>
<td>.55</td>
<td>.77</td>
<td>.49*</td>
<td>.40*</td>
<td>.61</td>
<td>.49*</td>
</tr>
<tr>
<td>E</td>
<td>Pre-Test 3</td>
<td>.43*</td>
<td>.63</td>
<td>.63</td>
<td>.45*</td>
<td>.59</td>
<td>.28**</td>
</tr>
<tr>
<td>T</td>
<td>Pre-Test 4</td>
<td>.48*</td>
<td>.44*</td>
<td>.42*</td>
<td>.90</td>
<td>.49*</td>
<td>.28**</td>
</tr>
<tr>
<td>E</td>
<td>Pre-Test 5</td>
<td>.45*</td>
<td>.65</td>
<td>.62</td>
<td>.80</td>
<td>.81</td>
<td>.47*</td>
</tr>
<tr>
<td>(notes)</td>
<td>Pre-Test 1</td>
<td>-</td>
<td>.46*</td>
<td>.54</td>
<td>.22**</td>
<td>.25**</td>
<td>.45</td>
</tr>
<tr>
<td>T</td>
<td>Pre-Test 1</td>
<td>-</td>
<td>.48*</td>
<td>.53</td>
<td>.52</td>
<td>.57</td>
<td>.36*</td>
</tr>
<tr>
<td>S</td>
<td>(Dictation)</td>
<td>-</td>
<td>.46*</td>
<td>-</td>
<td>.44*</td>
<td>.38*</td>
<td>.58</td>
</tr>
<tr>
<td>S</td>
<td>Pre-Test 2</td>
<td>-</td>
<td>.46*</td>
<td>-</td>
<td>.44*</td>
<td>.38*</td>
<td>.58</td>
</tr>
<tr>
<td>(Part A)</td>
<td>Pre-Test 2</td>
<td>-</td>
<td>.46*</td>
<td>-</td>
<td>.44*</td>
<td>.38*</td>
<td>.58</td>
</tr>
<tr>
<td>(Cloze)</td>
<td>All Pre-Tests</td>
<td>.72</td>
<td>.74</td>
<td>.75</td>
<td>.76</td>
<td>.78</td>
<td>.52</td>
</tr>
</tbody>
</table>

Figure 6.28: Correlation Matrix for prognostic validity of Pre-Tests for Post-Test performance.
The following are key points emerging from this matrix which may be taken along with points 1 to 6 above as correlational evidence for and against various aspects of test validity.

7. In all cases (except Pre-Test 3) the complete tests correlate more highly with their post-test selves than with any other test (see the underlined Spearman rho coefficients in the matrix). This is positive evidence that the tests measure the combination of traits designed into them rather than those in other tests. (It is not, however, evidence that those traits are necessarily what we intended them to be). In the case of Test 4, shock factor and all, the relationship between pre- and post-test assessments is especially marked. The computed Pearson r coefficient is .90 and the ANOVA shows an 80.5% shared variance across Pre- and Post-Test 4, even allowing for the variability of TL acquisition events during Phase One. The scatter diagram in Figure 6.29 underlines how closely scores cluster round the regression line, how closely we could predict end-of-course oral interview performance from the initial interview.

Figure 6.29: Scatter diagram of pre- and post-test scores on Test 4
The consistently rather low correlations between Pre-Test 4 and the other tests suggests that a measure of students' ability to communicate orally does not predict their later performance in note-taking, reading, report-writing or even free writing very strongly. Perhaps it also suggests the existence of a fairly independent 'speaking skill' (see, for example, Hinofotis 1976) or supports Cummins' (1980) BICS/CALP distinction, see Chapter Two, Section 3.4.

8. Pre-Test 1 is a reasonably powerful predictor of performance on Post-Tests 2 and 3 but not of the more personal communication required by Tests 4 and 5. The predictions of post-test note-taking are low all round. This needs further investigation in Phase Two (see Chapter Eight, Section 3 below). The 'competence' test (Test 2) is the second best predictor of overall post-test performance but correlates no better prognostically with oral interview performance than it did concurrently. Its two sub-tests (Part A and Part B) predict marginally less strongly than the whole test, the cloze this time performing a little better than the mixed-focus Part A, though not correlating positively enough with any of the other tests to give clear evidence of what area of communicative performance it really represents.

9. The combined insights from both correlation matrices have interesting general implications for the construction of test batteries seeking maximum concurrent and prognostic validity. Clearly a comprehensive battery such as this is more effective than any single test or smaller set of tests when the investigatory purposes are really multi-dimensional. But if placement and diagnostics rather than detailed profiling are the aim, my evidence suggests the following kind of battery would be useful:
An analytically rated free-writing test

A 'competence test' such as Test 2 but incorporating a dictation, which, in communication terms, seems to belong at the competence rather than performance end of the test construct continuum and, an important practical point with mixed-background groups, is very susceptible to significant gain as EFL learners quickly make up ground on ESL learners in the intensive TL-hearing context of C2.

An analytic criterion scored oral interview.

6. Individual Profiling: some case studies

The analysis of test performances so far gives a clear indication that while general levels of competence influence performance on all the various test tasks, there is enough individual variability in participants' performances on them for each test to be said to be contributing specific, apparently reliable and valid information, to learner profiles. Of Oller's three hypotheses, it is the partial divisibility hypothesis that is supported, i.e:

"there will be a large chunk of reliable variance shared by all the tests, plus small amounts of reliable variance shared by only some of the tests" (Oller 1979;1 p.425).

And a key paragraph from elsewhere in Oller 1979;1 makes an apt introduction to this section of my chapter, where the focus is on individual differences in the context of overall trends:

"If an individual happens to be much better at listening tasks than at speaking tasks or at reading and writing tasks than at speaking and listening tasks, we would be much more apt to discover this fact with valid language tests than with non-valid ones. However, the case of a particular individual, who may show marked differences in ability to perform different language tasks, is not an argument against the possibility of a very high correlation between those same tasks for an entire population of subjects, or for subjects in general" (p.194).
In this section, I shall make brief case studies of individual participants to see how validly interpretable the test data are, checked against and complemented by other independent criteria. The participants selected are interesting from the point of view of important considerations such as the competence/performance relationship, variability across tasks and generalisability.

Profiles of all 27 participants were written at the end of Phase One and sent to TCTD programme officers for onward transmission to tutors at receiving institutions (see pro forma in Appendix 2.4). A powerful practical argument in favour of the profiling validity of test data is the fact that on both 1980 and 1981 ELTI programmes:

1. pre-test results and profiles were used to group participants into two general ability groups (1980) and two-skill ability 'sets' (1981), without the staff feeling there was evidence of wrong placements.

2. pre- and post-test results were interpreted privately to individual participants by course teachers with general agreement on both sides that the resultant TL profiles were accurate, procedurally convenient and acceptable.

Participant C's pre-test battery performance placed her in the middle of the group (15th out of 27) overall with an all-test average suitably close to the mean (50.6% against 50.1%). However, a closer analysis of individual test performances reveals the interesting variability in performance and competence summarised here:

<table>
<thead>
<tr>
<th>PARTICIPANT C</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Note:</strong> r/o = rank order</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Listening Notes (T1)</th>
<th>Reading &amp; Rpts (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Dialogue</td>
<td>Discourse</td>
<td>Close</td>
<td>Overall</td>
</tr>
<tr>
<td>(15)r/o</td>
<td>(10)r/o</td>
<td>(10)r/o</td>
<td>(25)r/o</td>
<td>73 3=</td>
</tr>
<tr>
<td>13 2=</td>
<td>6 16=</td>
<td>7 4=</td>
<td>18 2=</td>
<td></td>
</tr>
</tbody>
</table>
POST-TESTS:

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Listening Notes (T1)</th>
<th>Reading &amp; Rpts (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Cloze</td>
<td>Overall</td>
<td>Dictation</td>
<td>% r/o</td>
</tr>
<tr>
<td>(15)r/o (10)r/o</td>
<td>(25)r/o % r/o</td>
<td>14</td>
<td>70</td>
<td>41 % r/o</td>
</tr>
<tr>
<td>1=</td>
<td>9 2=</td>
<td>7</td>
<td>78</td>
<td>15.5</td>
</tr>
<tr>
<td>18 6=</td>
<td>80 5</td>
<td>13</td>
<td>57</td>
<td>64 8</td>
</tr>
</tbody>
</table>

The pattern here cannot be unfamiliar to most ESOL teachers. C has a reasonable underlying competence as shown by scores and rank orders on Test 2. But her pre-test competence profile becomes less and less predictive of her performance the more performative, active and expressive the demands made by the other tests are, as seen by her results on Test 3, 5 and 4, in that rising order of interpretational and expressive pressure. This profile is supported by her EPTB (Form D) overall score of 44.5 (the second highest of the 16 previously Davies-tested participants in the group) and by relevant sub-test scores in the Davies Test. Her biggest problem was oral interaction (rated as 'unsatisfactory' on the EPTB subjective oral rating) and reflected also, perhaps, by her low mark and rank on the spoken use section of her competence test. C came to us, in fact, as a participant not 'performing her competence' (cf Spolsky 1968).

C's post-test profile shows signs that some of the gaps had been filled by the ELTI course. Her score and rank-order improvement on lecture note-taking and free-writing is significant, the former clearly supported by continuous assessment tutor comment (eg "Listening skills improved a lot") the latter tying in with a tutor's compliment on 'good planning, organisation and transcoding ability'. Interesting too, is the fact that the three raters on C's Pre-Test 5 gave her an average of only 4.3 on the compositional organisation criterion compared with
7.3 on the post-test of free writing. Although C makes progress (+14%) on the oral interview post-test, her rank-order does not change significantly. The continuous assessment comment here is suitably mixed: one tutor notes 'signs of a greater willingness to talk', another pinpoints referential inadequacy as responsible for communicative breakdowns along with a lack of 'strategies to overcome this'. He feels C is 'a message reducer rather than a risk taker'. Certainly, scores assigned for referential adequacy do not show improvement according to the raters of C's Post-Test 4. In fact most of C's oral interview improvement is accounted for by much better scores on the listening comprehension and formal accuracy criteria, gains which are also validated by continuous assessment comment. It is relevant in addition to note that C's own record of her individual self-study activities (designed into the ELTI programme) shows that she concentrated almost all her attention on formal/functional, aural/oral language laboratory exercises. Her own view (as recorded on our course feedback pro-forma (see Appendix 2.1)) was that listening and speaking were her main TL problem areas both at the beginning and by the end of the course.

On the problematic Test 3, C shows a reasonable gain. A check for concurrent validation of reading/reporting progress through the continuous assessment cards for C (and then the other participants), however, reveals a possible reason for the lack of significant improvement in group scores on Test 3. For C the only tutor comment on the 'Study Skills (3) Reading' card is that she is 'quick'. On no fewer than 14 of the equivalent cards for other participants, there is no comment at all. Does this suggest, perhaps, that the course did not in fact give much emphasis to the kind of global information processing and
reporting activities tapped by Test 3? If this was the case, the fact that Test 3 behaves more like an alternate form without an intervening treatment may be more explicable. Certainly a study of the programme schedule indicates an emphasis on aural/oral skills and the more controlled micro-level approach to reading that is exemplified by the core course reading book, *Think and Link* (Cooper 1979).

On the identical form re-administration of Test 2 (the competence-oriented test) C's 7% gain is accounted for entirely by a marginal improvement on the structure sub-test and a relatively significant one in the spoken use section of Part A, the latter reflecting her performance improvement in spoken interaction. She does not improve her well above average cloze score; no evidence here, therefore, of what the essentially non-authentic activity of cloze text completion predicts about performance unless C's equally static post-test SRA RFU score is an indication of a relationship between that test and cloze. (Following up this possibility though, the Spearman rho cloze:RFU correlations on pre- and post-test were found to be an unremarkable .42 and .64 for the group as a whole.)

So, our first case study, Participant C, may be regarded as typical of learners who have not yet fully realised their communicative potential but are helped to get nearer to performing their competence by initial learning and living experience in the new culture. My Phase One evaluation seems capable of producing TL profiles of such students without any obviously inexplicable inconsistencies or invalidities as this case-study check with various external criteria has shown. In Chapter Seven the profile can be extended and refined in the light of cognitive/affective and social data ready for longer-term predictive
validation over Phase Two.

In Participant L, we have someone who in some respects performs in almost the opposite way from C, but who is again probably recognisable to ESOL teachers as fairly typical of students who do not seem to display a competence worthy of their performances. The summary of L's pre- and post-test scores and ranks may begin to suggest how this phenomenon is evidenced:

**PARTICIPANT L**

**PRE-TESTS:**

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Listening &amp; Notes (T1)</th>
<th>Reading &amp; Rpts. (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Dialogue</td>
<td>Discourse</td>
<td>Cloze</td>
<td>Overall</td>
</tr>
<tr>
<td>(15)r/o</td>
<td>(10)r/o</td>
<td>(10)r/o</td>
<td>% r/o</td>
<td></td>
</tr>
<tr>
<td>13 2=</td>
<td>5 20</td>
<td>5 12=</td>
<td>8 17</td>
<td>52 14</td>
</tr>
<tr>
<td>Notes (T6)</td>
<td></td>
<td>Dictation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(25)r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
</tr>
<tr>
<td>34 21 86 2</td>
<td>45 12.5</td>
<td>81 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**POST-TESTS:**

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Listening &amp; Notes (T1)</th>
<th>Reading &amp; Rpts. (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Dialogue</td>
<td>Discourse</td>
<td>Cloze</td>
<td>Overall</td>
</tr>
<tr>
<td>(15)r/o</td>
<td>(10)r/o</td>
<td>(10)r/o</td>
<td>% r/o</td>
<td></td>
</tr>
<tr>
<td>14 1=</td>
<td>7 14</td>
<td>3 20=</td>
<td>11 22</td>
<td>58 18.5</td>
</tr>
<tr>
<td>Notes (T6)</td>
<td></td>
<td>Dictation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(25)r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
</tr>
<tr>
<td>62 12 86 5=</td>
<td>46 8</td>
<td>84 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

L placed 4th in the group overall with a mean score of 61.4% in the pre-tests and 8th in spite of a gain to 64.4% in the post-tests. But his profile from individual tests and sub-tests reveals an apparent mismatch between competence and performance that underlines just how complicated the relationship between those two notions can be. L's performance on the structure sub-test of Test 2, Part A is strong and predicts his strong, genuinely ESL oral interaction and free-writing performances: (on the former he was one of the very few participants to be given maximum native-speaker level criterion scores - and by several raters, especially the outside testing experts). Dictation
is also no problem, with errors only of the kind that could safely be attributed to 'carelessness' preventing maximum scores. Yet his scores on the spoken and written use sub-tests and his cloze test are low in percentage and rank-order terms resulting in overall 'competence' (Test 2) percentages of 52 and 58, ranks of 14 and 18. Thus the participant whom almost everyone connected with the course considered nearest to native-speaker TL competence is ranked below the mean on my competence test. Is this evaluation, then, invalid for L (and thus, logically,) as a test? Not necessarily. Where Test 2 focuses most narrowly on linguistic competence and TL receptive familiarity (ie in the structure sub-test and in the dictation) L emerges somewhere near where his speaking, free writing, outsider assessment and insider comment (eg "Formal Accuracy (spoken): 'Almost native'") say he should be. But as we have seen, my notion of competence is a broad one, extending along EAP-related dimensions similar to Cummins' (1980) 'cognitive/academic language proficiency' (CALP). The written use discourse skills tapped in Test 2 Part A sub-test 3 and the cloze test were expected to reveal things about this, and L's results on both are not very good. They can be taken to predict his uncertain performances in note-taking and reading/report writing, both supported by some interesting continuous assessment card entries:

eg STUDY SKILLS (1) - LECTURES
"Note-taking has been slow with far too many non-essentials included."
"Still inclined to be seduced by his 100% comprehension into writing down too much."

STUDY SKILLS (3) - READING
"Some problems noticed with speed of reading and finding main points."

STUDY SKILLS (4) - WRITING
"2 pieces of homework reveal fluency but lack of organisation."

SELF-ACCESS, INDIVIDUAL 'PROJECT' WORK
"Hard to keep him on a logical path."
"Very slow initially in referencing skills."

EXTRA-LINGUISTIC FACTORS
"Study skills are going to be a problem."
Now the test profile looks altogether more valid. L does indeed seem
typical of learners with good 'basic interpersonal communicative skills'
(BICS (Cummins 1980)) but with more questionable CALP. Interesting
too, that his self-assessment of his own needs puts 'study skills'
followed by reading as top priorities, pre- and post-course. We shall
see how Phase Two developments bear out this diagnosis in Chapter Eight.

A final point of testing relevance with L is that he arrived with the
lowest-but-two EPTB objective score of 33 interpreted as showing
'Insufficient English to follow a course. Minimum of six months full-
time English will be needed' (Davies and Alderson 1977), though he was
rated as satisfactory on the subjective essay and oral assessments.
It is difficult to agree comprehensively with the low Davies Test
assessment but it certainly sounds the kind of EAP warning a global
screening test is designed to. There is another inconvenient but
undeniable fact of testing life underlying L's test profile, too,
captured perhaps in a tutor comment that he 'seems to find it difficult
to follow instructions'. This was partly responsible for his poor
cloze-test performance and may well have contributed to his poor scores
on the fairly instruction-intensive Davies Test.

With Participant L, then, the Phase One evaluation system produces an
interesting, but not atypical profile of a learner strong in TL
communication when the activities concerned are open-ended and
expressive but less secure in its study-oriented use.

Participant 0 is a relevant case study because she was considered by
most of those connected with the ELTI course as one of the students
who showed most TL improvement during Phase One. It will be important
from the validation point of view to examine the extent to which her test profiles reflect and explain this. O began with an overall score just above the mean (at 53.4%) and ranked 12 out of 27. In detail, these were her pre-test assessments:

PARTICIPANT O
PRE-TESTS:

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Listening &amp; Notes (T1)</th>
<th>Reading &amp; Notes (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Dialogue</td>
<td>Dict.-Notes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(15) r/o</td>
<td>(10) r/o</td>
<td>(10) r/o</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8 20=</td>
<td>5 20=</td>
<td>3 22=</td>
<td></td>
<td>52 9= 58 19</td>
</tr>
<tr>
<td>Dis-course</td>
<td>Cloze</td>
<td>Over-all</td>
<td></td>
<td>51 6= 59 13</td>
</tr>
<tr>
<td>(15) r/o</td>
<td>(25) r/o</td>
<td>(% r/o % r/o % r/o)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11 11=</td>
<td>45 17=</td>
<td>52 9= 58 19</td>
<td></td>
<td>51 6= 59 13</td>
</tr>
</tbody>
</table>

The obvious interpretation is that O's underlying competence is restricting potential performance strength; those sub-tests where the focus is most clearly on TL knowledge and familiarity (ie Test 2 Part A and the dictation) are her weakest spots. Where CALP is most at issue the picture is more encouraging. A finer analysis of her reasonable reading/reporting and free writing performances reveals that it is the formal accuracy criteria that are scored low on all ratings. Early continuous assessment comments validate this diagnosis:

**FORMAL ACCURACY (Spoken):**
"Fairly frequent inaccuracies eg concord "we was" etc."
"Slips show because her fluency and desire to communicate outstrip her formal grasp ...."

**FORMAL ACCURACY (Written)**
"Poor grammar and spelling in an excellent creative description of her home town. Odd expressions - translations from Spanish?"
"Odd little areas of ignorance eg over 'from', 'to', 'at' time expressions."

Her lowish EPTB score (36.2 overall) seems an accurate enough assessment for O at the beginning Phase One. So does her own assessment of her needs; she rates 'study skills' as her lowest priority, listening and speaking highest.
By the end of the ELTI course, 0's test profile looks like this:

**POST-TESTS:**

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Listening &amp; Notes (T1)</th>
<th>Reading &amp; Notes (T2)</th>
<th>Dict-Notes (T3)</th>
<th>Oral Rpts (T4)</th>
<th>Free Writ- (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structure</td>
<td>Dialogue</td>
<td>Discourse</td>
<td>Close</td>
<td>Overall</td>
<td>Structure</td>
</tr>
<tr>
<td>(15)r/o</td>
<td>(10)r/o</td>
<td>(10)r/o</td>
<td>(25)r/o</td>
<td>% r/o</td>
<td>% r/o</td>
</tr>
<tr>
<td>13</td>
<td>7=</td>
<td>8</td>
<td>8</td>
<td>3=</td>
<td>15</td>
</tr>
</tbody>
</table>

The improvement is most marked on Test 2, Part A and the dictation, the areas of competence that were weakest to start with. And on the related formal criteria in the performance tests her scores are consistently higher, except in the oral interview, where it is better scores on fluency and flexibility that account for most of her improvement. Improved competence certainly seems to allow 0 to realise her oral and written potential as her percentage and rank gains on Post-Tests 4 and 5 show. Test 3, again, slightly spoils the picture, though 0's more or less static performance on the reading/reporting test is more in line with a lack of practice activities in the skills than with test invalidity. Interestingly enough her own self-access record shows she did no optional reading practice except for a couple of sessions with SRA cards - and on the RFU post-test her score was actually down, marginally, compared with the pre-test.

The tests profile Participant 0 as a quick learner, improving on a fairly weak, strictly EFL competence with a complementary general improvement in performance in self-expressive and study-skill tasks. By the end of Phase One her all-test percentage is 65 and, even more significantly her overall rank up from 12th to 6th. Interesting prospects, then, to be checked against her cognitive/affective and social profile in Chapter Seven, then, in Chapter Eight, the learning
and living events of her later C2 life.

As a final case study, I shall consider Participant T, this time as a student who, unfortunately, was not considered by the ELTI team to have made much progress over Phase One and seemed likely to have difficulty coping with her main course.

The first point that strikes one is that T's all-test percentage rises from 34.8 on the pre-tests to 43.4 on the post-test battery, though with her rank constant at 24th out of the 27. This gain, however, has to be seen in the light of the low starting point which, in the logic of the law of diminishing returns mentioned in Chapter Two, Section 3.1, makes it easier to show gain. This is probably even more the case if, like T, there is a deal of ESL experience still inhibited from showing itself fully at arrival shock stage. T's detailed pre- and post-test profiles show more clearly the lack of significant progress in competence as well as in key study-related areas of communication.

### PARTICIPANT T

#### PRE-TESTS:

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Structure</th>
<th>Dialogue</th>
<th>Discourse</th>
<th>Cloze</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(15)r/o</td>
<td>(10)r/o</td>
<td>(10)r/o</td>
<td>(25)r/o</td>
<td>% r/o</td>
<td>7 23=</td>
</tr>
<tr>
<td>5 20=</td>
<td>4 17=</td>
<td>3 22=</td>
<td>32 24</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Listening &amp; Notes (T1)</th>
<th>Dict-Notesation</th>
<th>Reading &amp; Rpts. (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td></td>
</tr>
<tr>
<td>26 24=</td>
<td>42 22</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39 14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37 24=</td>
<td>32 24=</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### POST-TESTS:

<table>
<thead>
<tr>
<th>Competence Test (T2)</th>
<th>Structure</th>
<th>Dialogue</th>
<th>Discourse</th>
<th>Cloze</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>(15)r/o</td>
<td>(10)r/o</td>
<td>(10)r/o</td>
<td>(25)r/o</td>
<td>% r/o</td>
<td>4 26</td>
</tr>
<tr>
<td>7 14=</td>
<td>1 27=</td>
<td>7 25=</td>
<td>32 26</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Listening &amp; Notes (T1)</th>
<th>Dict-Notesation</th>
<th>Reading &amp; Rpts. (T3)</th>
<th>Oral (T4)</th>
<th>Free Writing (T5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td>% r/o</td>
<td></td>
</tr>
<tr>
<td>28 25=</td>
<td>58 20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40 17=</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>56 22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>47 23=</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
T's competence, as measured by Test 2, stays static at 32% with a significant regression in the structure and written use sections of Part A that would raise questions about the stability of the test if the reliability analyses above had not indicated otherwise. The continuous assessment comments on T's linguistic competence are uniformly pessimistic, all of them suggesting a low level of formal and referential adequacy, some that the problem is serious enough to strongly inhibit fluency and flexibility. It is noticeable that even on the tests where T does show significant gain, the oral (Test 4) and free writing (Test 5), the formal accuracy criterion is still rated low. Still, T's gains on these two tests are statistically very significant. It is the competence test scores, the lack of rank-order improvements and the teacher comment that persuade one to look more critically into these two pockets of apparent improvement and to see the low starting point and arrival shock (the latter supported by evidence from her pre-/post-test dictation scores) as key factors in T's Test 4 and 5 gains. On the study-skills oriented performance tests (1 and 3), scores are static, rank-orders marginally down, and at the bottom of the group this may mean more, since quite marginal improvements against the weakest 'competition' can cause relatively greater change. The continuous assessment verdict here is again discouraging: 'Very basic and showing few signs of improvement' on listening comprehension and note-taking; 'Does not participate' under 'Seminars/tutorials'; an ominous, though as we have already seen, typical blank under 'Reading', and 'slow' at writing.

There are two sets of outside validation evidence for T. On a January
1979 subjective assessment in Cl, she was rated A on 'ability to understand spoken English'. This rating is glossed, on the level descriptions used, as follows:

"When addressed in normal to fast English with no concessions made to the fact that (s)he is a non-native listener, and with only very occasional rephrasing or repletion, (s)he understands everything."

The straight B's that T was given on the three other macro-skills are normally interpreted as indicating satisfactory performance. Now these ratings clearly do not tally with my test profiles or with ELTI tutor assessments. However, in May 1980, two months before our course, T took the new English Language Testing Service battery (eg ELTS 1979). On this, her overall rating was just below Band 4 ie.

"MARGINAL COMMUNICATOR. Lacking in style, fluency and accuracy, is not easy to communicate with, accent and usage cause misunderstandings. Generally can get by without serious breakdowns."

Constituents of this overall banding were bands rating T as 'extremely limited' (Band 3) on listening comprehension, 'a marginal communicator' (4) on reading, a 'modest communicator' (5) on her interview, 'marginal' (4) again on writing and 'extremely limited' (3) on study skills. With these assessments both my test profile and ELTI teacher opinion would be in broad agreement. (General concurrent validation exercises with ELTS battery was not, however, possible with my group since only 4 of the 27 had taken the new test. The 1981 ELTI group provides useful information for comparisons between ELTS and my battery since 23 out of the 29 had taken the former test in their home countries and retook it at the end of their ELTI programme.)

For Participant T, then, the Phase One evaluation system informs a profile that sounds definite warnings. Again, their significance will
be put into personal and social perspective then checked against subsequent reality below.

One important area where further investigation has been promised but where the four case-study participants are not representative of the group, is note-taking from a lecture as sampled in Test 1. C, L and O showed substantial progress on this, based on pre-/post-test results whereas, it will be remembered, Figure 6.1 (a) indicated a statistically non-significant gain by the group overall. I needed to investigate various external criteria for insights into whether the explanation lay in the invalidity of the test or in other factors. The following relevant points emerged.

1. That, unlike the case of Test 3, there were a fair number of entries on the appropriate continuous assessment cards ('Study Skills - (1) Lectures (Listening comprehension, note-taking)'), suggesting the abilities tested in Test 1 had received attention during the programme. However, tutor comments on individual participants' relevant abilities and progress, do not, this time, in general match the verdict of the post-test. So, a closer look was taken at the actual programme activities on which the tutors were basing their comments.

In summary, these activities were:

.1 controlled, step-by-step approaches to lecture and note-taking eg (see James et al 1979) a short dictation of the gist of lecture followed by a slightly expanded version with specific information - and/or skill-based exercises, then the lecture itself.

.2 'talking head' video lectures, usually with preparatory orientation work and group follow-up discussion or exercises.

.3 occasional visiting speaker orientation talks (eg on banking,
health) with question and answer sessions. The immediate course-design sense of all these activities is clear. Their relationship with real-life lectures is not, however, so obvious. In the real EAP world there is not usually much advance guidance on what kind of information you will have to process or how, and often your note-taking task is unguided by tutors, unaided by the group. Thus, tutors were commenting on participant abilities and progress in activities by no means the same as those required by the unhelpful, normal time-constrained conditions of Test 1.

2. Tutor comments were also based on the listening practice activities most participants selected during some of their self-access sessions. These were almost exclusively the ELTI Listening Library tapes, which are mostly of authentic oral lectures on a variety of topics. These should surely have helped with the communicative abilities tapped by Test 1, even if, as most of the individual study record cards indicate, most participants had only three or four such sessions during the course. But again, if we look more closely at the actual routine involved, there are factors that reduce the direct relevance of practice to test. Listening to taped talks gives one the chance to pause or replay as often as one wants; indeed the accompanying instructions to the Listening Library tapes encourages this - and for perfectly understandable pedagogic reasons. In terms of full genuineness of activity, however, the luxury of time to reflect and repeat is non-authentic. The fact that connections between quantity of self-access listening practice and Test 1 results are even more tenuous than those between the test and teacher comments becomes more explicable.
On the evidence so far, we can neither exonerate nor discount Test 1. A final verdict on its validity will have to wait until Phase Two. If participants' performance on the alternate form of the test used at the May 1981 reunion has not improved, after eight months of real-life lecture/note-taking events, than Test 1 can surely be discounted as a construct-valid measure.

The four students case studied here were not, as should have become clear, selected because they show my Phase One evaluation system in its conveniently best light but because they are four individuals with differing yet generalisable profiles that search out the capacity of test and other TL data to describe and predict abilities with a degree of sensitivity. In fact, rather 'easier' participants to evaluate are those scoring highest and lowest on tests. It is a reality of testing that most language tests, competence- or performance-oriented, will sort out the two extremes of proficiency fairly comfortably. In my group, W's TL level was so obviously weak that he averaged 14 and 11% overall below the next weakest student in pre- and post-tests respectively. Similarly, E and D were not too difficult to profile as communicatively effective (nearly) all round with their firm underlying competence and strong, though individually variable, performances. Participants like A and B also seem amenable to confident profiling through my test evidence supported by other concurrent validation data, the former erratic in the CALP areas, the latter weak in BICS (in L1, he would, very pertinently, claim, as well as in TL), but both showing a competence that keeps them in the top 4 all round. Whereas U and V remain in the lowest 4 at well below the 50% competence-test level and even though the former performed adequately on Pre-Test 3, the
latter on Pre-Test 1, both raised doubts about their EAP future. In fact, a close study of all the TL evidence up to the end of Phase One left one with a feeling, verging on a hypothesis, that a reasonable competence level as shown by a score above the mean (ie 63.8%) on Post-Test 2 was reason for confidence about the EAP future of participants, scores below possible reason for concern. But this would be to hypothesise from unnecessarily limited insights. In the next Chapter I shall try to flesh out the profiles with cognitive/affective and social findings so that I can hypothesise with deeper insights.

Research Hypothesis 3 has been tested in this chapter in the way promised in the thesis outline in Chapter One and the empirical research design in Chapter Five, that is through the combined quantitative and qualitative investigation of narrower, testable hypotheses. On the quantitative side every statistically significant finding has implied the rejection of a relevant null hypothesis requiring refutation on my way towards establishing reasonable reliability and validity for my tests. Alongside throughout, and centrally in Section 6, more qualitative investigations have been made, sometimes tending to falsify specific hypotheses, sometimes in their support. Given the acknowledged constraints, for example on sampling, the non-feasibility of a strictly experimental methodology, and my acknowledged preferences, for example for a mixed-method individual difference-oriented approach, Research Hypothesis 3 is considered sufficiently well supported not to invalidate the pursuit of corroboration of my two other Research Hypotheses. We have elicited the kind of data that usefully informs participant TL profiles. We can now test specific hypotheses germane to RH2, the hypothesis that certain cognitive/affective and social insights will validly refine the profiles. This is the task in Chapter Seven.
CHAPTER SEVEN

THE EMPirical STUDY: PHASE ONE - THE EVALUATION OF COGNITIVE/AFFECTIVE AND SOCIAL FACTORS
1. Introduction

This chapter has the following main purposes in Phase One of my study:

1. to validate and interpret findings on personality factors.
2. to summarise test and experimental results from investigations into cognitive style.
3. to describe, analyse and interpret findings on motivation and attitude.
4. to assess, through selected case studies, the use of cognitive/affective data to supplement and refine profiles of participants' communicative competence.

The chapter is the second of those reporting on my empirical study. Its main focus is the testing of specific hypotheses implied by Research Hypothesis2, which posits relationships between individual and combined cognitive/affective and social factors and TL learning and use. Discussion of the social dimension will be included in the cognitive/affective investigations rather than under a separate heading.

2. Personality Factors

Chapter Two Section 3.4 sought to justify the investigation of personality factors as part of the cognitive/affective and social profiling of learners. It also explained why I chose the Cattell 16PF inventory as a starting point for the investigation, though to be subjected to close checks on reliability and validity.

One of the reasons for the use of the Cattell 16PF was an interesting finding during my pilot work with the 1979 ODA Group. This was the fact that no fewer than 19 students out of the sample of 26 scored high on Factor Q1, a personality trait that seemed potentially interesting
in a study of overseas EAP students. The following summary description of tendencies among high $Q_1$ scorers, taken from Working Guide for 16PF Interpretation (IARC 1978), may begin to suggest why:

**FACTOR $Q_1$**

**Experimenting, Critical, Liberal, Analytical, Free-thinking.**

(Radicalism)

The person who scores high on Factor $Q_1$ tends to be interested in intellectual matters and has doubts on fundamental issues. He is sceptical and inquiring regarding ideas, either old or new. He tends to be more well informed, less inclined to moralise, more inclined to experiment in life generally, and more tolerant of inconvenience and change. Open to new ideas, shows initiative in introducing change ...." (p.13).

Cattell's own slightly fuller description of the trait (1970 p.104/5) mentions more interest 'in reading as opposed to class instruction'

and that:

"In group dynamics the $Q_1$ + person contributes significantly more remarks to discussion, a high percentage being of a critical nature" (p.105).

The trait seemed to have considerable potential. Was it not a plausible independent presage variable, likely to characterise overseas EAP student groups, people prepared to take the radical step of training and living abroad, with the healthy capacity to tolerate 'inconvenience and change'? And would not their openness to new ideas, willingness to take initiatives, their questioning prominence in discussion, all be relevant to their TL and specialist studies? Especially so perhaps, if, as is indicated in the 1978 working guide, the $Q_1$ factor relates to Cattell's second order factors of independence and creativity, to 'values' such as achievement, goal orientation and to 'self-sentiment/career sentiment' in the area of motivation.

But even with the pilot group I felt the need to examine the statistical
evidence further, given the question marks over the reliability of the 16PF (see Chapter Two, 3.4 above). In all my sub-group comparisons, however, the mean on Q₁ never fell below a sten score of 7, the level we are enjoined to regard as 'definitely "departing from the average"' (Cattell 1970 p.63). Both 'arts' (n=12) and 'science' (n=14); male (n=20) and female (n=6); Latin American (n=15) and 'others' (n=11); those aged 30 or under (n=14) and the over thirties (n=12); those taking Form A of the test in L₁ (n=16) and Form E in fairly 'easy' TL (n=10), all these sub groups averaged over 7. Yet on none of the other 16PF factors did the 1979 group, as a group, produce scores outside the 4 to 7 sten range. At a mean of 7.65, the Q₁ factor was uniquely significant nomothetically.

I thus approached the more serious re-use of the Cattell 16PF Test with the 1980 participant group with the intention of checking its potential as a general profiling aid, and with a specific interest in exploring the Q₁ factor, in both, constrained by warnings from previous research that the Cattell instrument required careful handling. As the chronological diagram of my study (see Chapter Five, Figure 5.1) shows, the Test was administered for the first time in the middle of Week 1 of the ELTI programme, this administration following the 1979 pattern very closely. Participants for whom the test was already available in mother-tongue translation, were given Form A, the rest Form E, the version designed for candidates with a less than sophisticated level of English. In effect this meant that the seven Latin American participants, plus participant K from Mali (who preferred the French version) took Form A, everyone else Form E. As before, the Test was independently administered and computer-scored by the IARC.
Two immediately interesting statistics emerged.

1. Factor Q₁ was again the highest scoring factor, at a mean sten of 8.15, with 22 of the 26 participants who took the test scoring 7 or more.

2. None of the other factors (with the exception of Factor B the 'intelligence' scale which even Cattell himself advises against interpreting without supplementing it with other IQ tests or 'special aptitude measures' (1970 p.83)) 'definitely departed from the average' for the group as a whole.

With the Q₁ group mean even higher than in 1979, it was not even necessary to check various sub-groups to know that whichever way you divided the participants, scores on the 'experimenting' trait would be uniformly high. But this time the participants were the real focus of serious longitudinal research, rather than a pilot group. The Popperian ideal, my own focus on individual differences rather than group stereotypes and the fact that such a focus had not up till now led me to expect clearcut findings such as this (cf Chapter Six), urged further investigation. Since the participants themselves were interested in the research and claimed to have enjoyed the test, I did not feel it was an imposition on them to take up another 40-60 minutes of their time in a second administration of the 16PF questionnaire.

If there was a flaw in the strikingly neat findings so far, previous research (eg Savile and Blinkhorn 1976) indicated it might come from the lack of stability across the different forms of the 16PF. And Cattell's own equivalence coefficients for Q₁ (.34 for Form A:B, .26 for C:D and .51 for A+C:B+D (1970 p.33)) did seem rather low. I needed
to try a different form of the test. Three weeks after the first Test, IARC agreed to administer Form D of the 16PF, a rather shorter test than Forms A or E but not available in non-English translations. They agreed that ELTI course teachers should be permitted to aid their own staff in clarifying any language problems arising, at any time during the test. Answer sheets would again be computer scored.

The IARC computer produced contradictory results. They can be relevantly exemplified by just four summary statistical points.

1. The mean score for my group on Q₁ was just 4.74, the twelfth highest instead of the highest of the 16 factor scores. Only 5 of the 27 participants had a sten score of 7 or more.

2. The only factor with a sten mean above 7 was Q₂ [+ self-sufficient, + resourceful], a factor averaging 5.7, the eleventh highest sten on the first administration of the 16PF test.

3. The Spearman rank correlation coefficient of factors on first and second administrations of the 16PF was −.35; the extreme change in prominence of Q₁ in other words, was reflected by changes in the ranks of other factors.

4. The Pearson r correlation coefficient for Q₁ scores on the first and second administrations was a low .17.

I now had to decide whether the unstable scores on the 16PF inventory meant that the test was too unreliable (and thus invalid) to be of any use to my study or whether there was another explanation, possibly the contaminating influence of the TL (see Ollier (1979;1) in Chapter Two above), which, if it could be controlled, might still make further investigation worthwhile. Since the construct of the Q₁ factor itself was still of interest, I decided to make a final attempt to pursue it,
this time with all participants able to respond in their $L_1$. With IARC permission, I designed a short inventory containing all the $Q_1$ items from Forms D and E of the 16PF and a selection of other items from both forms to reduce a possible 'over-focusing' effect of a set of items all tapping a single trait. (After all, it is part of the rationale of such 'indirect' personality measures as Cattell's that people should not know what trait any item actually pertains to.) All items were translated into the 9 first languages required, the resulting inventory of 23 items presented in bilingual format, TL and $L_n$. The results on this third $Q_1$ measure were as follows:

1. A mean sten score of 5.3 (SD 2.0) on the $Q_1$ items from Form D.
2. A mean sten score of 6.7 (SD 2.23) on the $Q_1$ items from Form E.
3. A mean sten score of 6.0 (SD 1.9) on the $Q_1$ items from Forms D and E combined.

One immediate effect of the attempt to control any extraneous variance contributed by the TL factor seems to have been to bring $Q_1$ mean scores on both Form D and E closer to the centre, on Form D the score now being higher, on Form E, lower than before. In neither case now, would the group as whole be characterised as particularly high on $Q_1$. But some more statistical checks were needed if data from the 16PF were to be used for individual profiling. Hence the following analyses and interpretations.

1. $T$-tests for differences between scores on different administrations of the test revealed that the Form D version showed the highest level of stability with a $t$-value of 1.6 with 26 degrees
of freedom, showing a statistically non-significant difference. The product-moment correlation for the two Form D tests was .63, compared with figures ranging from .66 to .83 quoted by Cattell (1970 p.30) as 'dependability coefficients' 'between two administrations of the same test when the lapse of time is insufficient for people themselves to change' (his emphasis), in my case just under four weeks. These figures could be interpreted as indicating that the language factor contributed towards instability but not to a very significant extent. Similar figures comparing scores on the TL and L₁ versions of Form E showed less stability and a lower product-moment correlation. (p<.01; Pearson r = .47). Since there is no reason to believe that the translation factor should influence Form E any differently from Form D, the implication is that Form E itself is less stable.

2. There is certainly a difference in the way the two versions perform as can be seen from comparing the scores on Forms D and E in their mother-tongue translations. Here a t-test shows a difference significant at p<.0003, though the reasonable Pearson r correlation of .65 and a sign test show most of the difference explained by Form E's tendency to give consistently higher scores.

3. Most researchers would back Form D over Form E on this evidence especially as its format offers three alternative responses on each item as opposed to the either/or choice offered by Form E.

4. Most researchers would also set more store by the findings of the last 16PF administration (the combined D+E items in translation). This is not only because the chances of extraneous TL variance were reduced, but also because participants should have been benefiting from the repeated administration. Cattell (1970)
gives theoretical and independent empirical support for this notion, suggesting that 'extending questions over several equivalent forms, and the practice of not doing all the testing in one session' brings psychometric gains such as greater 'differentiating power' and stability, perhaps because individuals' 'repeated contact with the questions helps them to get to know themselves better' and 'to decide more definitely' (pp.33-34).

There is no doubt that my hopes for broad-ranging personality data from the Cattell 16PF were severely damaged by its unreliable performance with the group. It is good for the validity of the study, however, that problems caused by the lack of stability across forms and those caused by TL interference were discovered. The falsification of the early hypothesis about Q₁ as a group factor led to a re-exploration made possible by the flexibility of the research design. As a result, statistically significant data about Q₁ as an individual factor can be submitted to further validation procedures.

10 of the participants had a sten score of 7 or above on the Q₁ factor based on the combined L₁ test already described. 7 of these scored 7 or above on both sets of items (i.e. those from Form D as well as Form E). Indeed, 8 of the 10 averaged 7 or more over all four measurements of Q₁, which, even though we have raised serious doubts about the validity of the first two administrations of the test, could suggest their response to the items measuring the factor was so clearcut that it overrode any language or test-stability problems. I shall now, therefore, look at other Phase One information to see if it supports or refutes what the 16PF seems to suggest about participants with Q₁ scores that differ significantly from the mean. This subjective
information is seen as complementary rather than alternative, in line with the approach usually advocated by personality assessment experts (eg Ghiselli et al. 1970) and in line with the combined quantitative/qualitative methology of my research.

The most obvious sources of comparison are the 'extra-linguistic' factors card in the continuous assessment system, responses by participants to the feedback questionnaire (all referred to in Chapter Six) and the end-of-programme group discussions (see Chapter Five, Figure 5.1). Not that I expected Cattell's own terms - 'experimenting', 'sceptical', 'inquiring', 'tolerant of inconvenience and change', 'critical' etc, or their antonyms, to appear explicitly from any of these sources. Neither teachers nor students had been oriented towards Cattell's Q1 factor; the continuous assessment and feedback system was primarily part of the course design and only secondarily for the benefit of my empirical study.

The following relevant points emerged from a detailed analysis of data from these sources.

1. On the 'extra-linguistic factors' cards in the continuous assessment system (the cards on which, interestingly enough, teachers chose to write most comments) reference to participant awareness is noticeably more frequent with the Q1+ group than with the other 17 participants. Participant D is described as 'very aware', G as 'well-organised and aware'. O 'is able to decide what her problems are' and comes across to another teacher as a 'concerned' individual, like J, who is described as 'lively and concerned' then as 'very conscious of her rights'. E is 'a good organiser' according to two different teacher comments, 'perceptive' and someone who 'susses his audience out'
according to a third. Y is 'shrewd' and 'good at bringing discussion back on course without offending ....'. C (as we already know) has 'clear organising skills' but about F there is disagreement with one teacher not certain of her clarity of thinking where another considers her 'well-organised'. The remaining two Q₁⁺ participants, I and U, are not given any continuous assessment comment that could be construed as explicitly indicating 'awareness'. Of the participants not scoring above 7 on the factor, K is certainly described in terms that are in tune with those used with the high Q₁ group: 'very sharp and precise-minded' and 'good at seeing wood as well as trees'. But otherwise there does seem to be a clear differentiation by the teachers broadly following the Q₁ pattern in the group.

If this pattern does in fact relate to the conservative ....

experimental construct among Cattell's factors, what are the possible connections? The ELTI course teachers seem to be characterising the behaviour of at least 8 of the 10 Q₁⁺ participants as evidence that they are particularly aware, committed, thinking individuals. It does not seem too fanciful, pending further criterion validation, to link such traits with Cattell's own use of descriptors such as 'tendencies to .... inquiry', 'well-informed', 'less unquestioning about views generally', 'more interest in analytical thought' or 'free-thinking'. This potential link can be followed up immediately from a second Phase One source.

2. Reference to the end-of-course feedback questionnaire was made in Chapter Six Section 6. The final part of this (see Appendix 2.1) invited critical comment and recommendations from participants. Bearing in mind Cattell's conviction that Q₁⁺ individuals tend to favour
critical comment, to contribute 'significantly more remarks to discussion, a high percentage being of a critical nature', I felt it was worth analysing this part of the feedback sheets to see if there was evidence of a $Q_1$ connection. The answer was yes. A count of the number of critical points made by each participant (validated by a second reader) showed a mean of 3.4 in the $Q_1+$ group compared with 1.8 among the remaining 17 participants. (The variance ratio $F$ on the two sets of scores is significant at $p < .05$). A similar analysis of recommendations for future courses revealed no significant difference between the two groups. Nor did my attempt to discover whether $Q_1$ appeared to relate to the subject matter of the comments and proposals made. No, it was the explicit exercise of the critical faculty itself that the $Q_1+$ participants emerged predominant in.

3. My empirical study design in Chapter Five, Figure 5.1 shows that there were also discussion group feedback sessions at the end of the ELTI course. Since Cattell's 1970 suggestion about the $Q_1+$ inclination towards critical comment was originally made in connection with 'group dynamics', it is relevant to analyse this event with reference to the behaviour of the two sub-groups, the high and non-high $Q_1$ scorers.

The group discussion I analyse, 'live' and from video-recording, involved participants who were in Group S on the ELTI course, ie assessed on the pre-test battery as the 14 most proficient TL learner/users overall. It took place at the end of the actual teaching part of the course in round-table discussion form without teacher intervention. Unsurprisingly there is no clearcut $Q_1$-related distinction in the tallies of substantial discussion contributions by participants. Where 'substantial' is taken to mean contributions of
an initiating or responsive kind making one or more definite points in the discussion, the high \( Q_1 \) group (\( n=8 \)) has a mean of 8.25 compared to the low or average \( Q_1 \) group's (\( n=6 \)) 7.33. The quantity of contributions on this basis is more strongly connected with other factors, especially the positive communicative inclinations of some participants.

I cannot now use 16PF results on Cattell's [+ outgoing] Factor A or his [+ venturesome, socially bold] Factor H because of the reliability problems analysed above. However, if the 'extra-linguistic' criterion as evaluated by all assessors of Pre- and Post-Test 4 is analysed, a clear relationship between references to individual participants' 'extroversion', outgoingness, lack of shyness, self-confidence and their number of contributions in this group discussion emerges. Easily the most frequent contributors are Participants A, E, L, 0 and X, all of whom are described in Test 4 assessments by at least two assessors as 'extrovert', 'outgoing', 'talkative' 'socially without problems' or simply as 'good communicators'. Infrequent contributors like D, F, I and K are characterised equally often as 'shy', 'retiring', 'reserved', 'low key'. Only Y, who contributes but three times to the discussion has Test 4 comments (eg 'likes to talk' and 'confident') which seem to belie his performance on the day. With this exception, however, it is teacher assessment of participants' interactional tendencies, rather than Cattell's \( Q_1 \) Factor that is predictive of individuals' frequency or oral involvement in the feedback session.

Yet if the more difficult task of evaluating the discussion qualitatively is attempted, a somewhat different picture emerges. As two other observers of the discussion agree, the session is actually orchestrated by E and 0, the former with the highest number of
initiating comments and managerial suggestions, the latter selected unopposed as chairperson, serious about her responsibilities, steering the discussion towards 'propositions' and insistent on an analytic summary of conclusions. It is from E that the most radical suggestion comes, namely that pre-sessional EAP students would be better served by 'guest' participation in native-speaker training programmes, where the unhelpful prominence of TL communication with other foreign students would be avoided and learning by and for survival would be enforced. J, who asks for British students as official contacts and informants in the hostel where all participants stayed during Phase One; D who stresses that:

"It's not enough to learn English just with 6 hours a day. It's up to you to be and talk with people and try to understand them ...." or that "Sometimes we are talking English between ourselves and British people cannot understand us."

and I, also initiate on the importance of acquisition direct from C2 sources. In fact, I's comment, his only utterance during the whole discussion, is critical and searching in its implications:

"My opinion is that we spend too much time here speaking to ourselves. Because I'm not interested in understanding people from other countries. I'm interested in understanding my teachers who are going to be English. My .... most of my classmates are going to be English and I'd like to spend more time listening to English .... correct and speaking to somebody who could speak more about my correct pronunciation because I'm going to deal with the English."

G, 'retiring' according to his teachers and high on Cattell's Q1, contributes with only average frequency but is uniquely concerned with questioning the whole format of the discussion. He is against the idea of having anyone in the chair, asks several times for the discussion to be 'free', yet is sceptical of some of what has been said:

"I mean if you have some opinion about some subject, then you can explain - but if you haven't any opinions about the subject then it is rather better to hear about other's opinions."
These high $Q_1$ participants are again showing awareness, the kind of 'awareness of the factors affecting their study' that Brew (1981 p.2) sees as crucial to the development of perceptions of underlying frameworks or Entwistle, Ramsden and Burkenshaw (1981) to the 'deep approach' or 'meaning orientation' to academic study which, they find, relates to personality traits such as 'thinking introversion' and 'theoretical orientation'.

And there may be discriminant validation evidence for the $Q_1$ connection here, too, if we analyse the contributions of other participants in the discussion. A is a frequent contributer (in line with Test 4 and continuous assessment views on his willingness to talk) but his comments are not analytical or critical. Rather they are descriptive, generalised or even, in the terms of Entwistle et al. (1981), of a 'strategic' orientation, that is tending towards 'cue-seeking' about the pragmatic rules of the study or assessment game (see also James' Type 1 student in Figure 2.3 above). For example:

"This language teaching course is useful to a great extent .... makes me able to adapt to the new environments in the university."

or his initial suggestion during the group's self-organising phase of the discussion:

"And I would think that the course is especially for our advantage and not for any disadvantage .... some disadvantages .... I think should not be brought in our purview."

H and L are two other lowish $Q_1$ participants who make frequent contributions, the former anecdotally with some criticisms that suggest a somewhat inaccurate view of the course, for example, that its emphasis was on writing (a view that is not supported by the rationale, timetable or, as far as teachers were concerned, actual practice); Participant L playing an interesting TL informing role, but sometimes making
points that the others consider have already been covered or are already accounted for, eg

"0: The course contains some points: training, academical things, social things, introduction to the life of English people. We can begin with this order, you know and after that to get questions, to get conclusions, to get ....

L: I would propose a second one: What is the relationship between the English teaching at present and our placement course in the university?

0: Yes but maybe this question is contained in one of the points .... the first one is academical points ...."

Participant X is a counter-example to the hypothesised Q1+: 'deep', analytic, critical relationship as his comments, in spite of his average rather than high score on Q1, are percipient and searching. K, too, who it will be remembered, was described by teachers in Q1+ terms not reflected by his 16PF score, exhibits 'thinking introversion'. His only contribution to the discussion is certainly a fundamental one based on incontrovertible critical logic:

"In my case, I think I can't give any opinion about this course before I begin my main course. After my main course I can see if it is suitable or not, not before."

4. It is probably necessary at this stage in my qualitative investigation of hypotheses based on a quantitative measure to pause to check on the possible intrusion of uncontrolled variables. The 10 Q1+ participants are representative of the group as a whole in terms of sex, age, the 'arts' vs. 'science' parameters. They include members of the three major regional areas (Near/Middle East, S E Asia and Far East, Latin America) but not from the two lesser represented regions (S Asia or Africa). Perhaps A's reluctance to comment critically in the group discussion is an instance of variance contributed by cross-cultural factors, as some teacher comment suggests it was. More significant,
perhaps, is the fact that the 10 Q\textsubscript{1}+ participants in the total group have a higher TL proficiency level than the remaining participants, with mean scores of 55.5\% and 62.3\% on my pre- and post-test batteries compared with 47\% and 53.3\%. This should not significantly affect inferences based on the extra-linguistic observations from the continuous assessment cards (see .1 above) since teachers were, in theory at least, using the card concerned for the very factors not covered by the other TL-related cards. The same goes for the 'extra-linguistic' criterion on Test 4, used for validation in .3 above. And if the TL proficiency factor influences the amount of critical feedback points on the questionnaire (see .2 above), it is not immediately clear why it had no such influence on the recommendation section, which may be seen, in experimental design terms, as a control. Data from the group discussion (analysed in .3) should be free from non-systematic TL variance since the 14 participants were all members of the higher TL group according to overall pre-test placement and with the Q\textsubscript{1}+ sub-group (n=8) and Q\textsubscript{1}- sub-group (n=6) as nearly matched as possible on my oral interaction post-test scores (x=70.25\% (SD=8.7) and x=70.0\% (SD=9.9) respectively). This attempt to control the TL factor variance is necessary because the null hypothesis positing no relationship between TL and Q\textsubscript{1}+ is not proven. Computer analysis shows a Pearson r of .23, the kind of weak but positive relationship revealed by the scattergram in Figure 7.1 below, where mean post-test scores on my complete battery are plotted against the combined Q\textsubscript{1} measures on the L\textsubscript{1} version of Forms D and E for all 27 participants. But the relationship is only a weak one. Hypotheses testing Q\textsubscript{1} as independent variable with TL as measured by improvement between pre- and post-tests prove null whether scores or ranks are used. If we used pre-test scores on the vertical axis, Figure 7.1 would not look significantly different.
It is to be hoped that the procedures I have used to salvage valid data from the problems encountered in the use of the Cattell personality test with my group are instructive in themselves from a research point of view. Also that the investigation of the $Q_1$ factor has provided profile information that will prove valuable when it is related to Phase Two developments. The consideration of personality variables does not end here, though, since, as the discussion of the inevitable overlap between them and more cognitive and motivational factors in Chapter Two shows, they are likely to re-emerge as we move on to new hypothesis-testing areas.
3. Cognitive Style

My source of objective observations on cognitive style is the Group Embedded Figures Test (Witkin et al. 1971). The test itself and its cognitive, affective and social interpretations are described in detail in Chapter Two, Section 3.5. It was administered under strict test conditions on July 31 1980 with the following results.

1. The mean score for my group was 10.63 compared with the 11.3 given as a relevant norm 'based on men and women college students' (Witkin et al. 1971 p.28). The slightly lower mean score expected from female testees (10.8 compared with 12.0 for men) is also reflected in my population, where the scores are 9.75 and 11.0 respectively.

2. The spread of scores in my group was extremely broad, ranging in fact from the minimum (zero) to the maximum (18). The 10 participants with scores from 0 to 7 were characterised as field-dependent, since their scores put them in the first quartile according to the selected norm. The 9 participants with scores between 14 and 18 were considered as field-independent according to the test, given that their scores put them in or near the fourth quartile.

As the test is language neutral, performances on it did not reveal any obvious ecological or administrative problems and scores with my control group (n=9) (which included ELTI course teachers) were a comparable $\bar{x}=12.7$, I had no immediate reason to carry out the kind of statistical re-validation procedures forced on me by the 16PF results. In any case, the experiments planned into my design were themselves aimed at checking connections between field dependency and criteria relevant to the profiling of EAP students.
Field dependency is the key independent variable in these experiments and, as it happened, most other important sources of variance were reasonably well controlled. Both the field dependent (FD) and the field-independent (FI) sub-groups were similar as regards male/female, 'science'/'arts' and regional background distributions except that all three participants from South Asia were scored by the GEFT as field dependent. Using the overall pre-test mean scores as a measure of TL level (since the experiments were carried out before the post-test measures) the two groups were fairly well matched at 49.8% for the FD's and 51.8% for the FI's, a statistically non-significant difference. On the Q1 factor, however, there is a weak positive relationship with field independence; the FI group averages 6.9 on my combined L1 measure of Q1, compared to the 5.9 of the FD group. The scatter diagram in Figure 7.2 illustrates the nature of the Q1:field-dependency relationship for all 27 participants; it is positive overall but not powerfully or consistently so.

![Scatter diagram of scores on field dependent/independent and conservative/experimenting tests](image)
In the first set of experiments I set out to test hypothesised relationships between field dependency and the processing of written text, using insights from work carried out by researchers such as Satterly and Telfer (1979) and Annis (1979), mentioned in Chapter Two, 3.5 above.

In Experiment 1 (see Appendix 2.2) the focus was on participants' ability to sequence disorganised information, with the expectation that field-independent learners should be better at it, given their analytic tendency, their characteristically articulated perception of experience, their experience of parts of a 'field' as discrete, with the 'field' as a whole seen as structured (see Chapter Two, 3.5 above). Participants were asked to sequence three separate sets of information, differing in type, length and complexity. The first was a short text with the original sentences presented out of sequence with no paragraphing. Participants were to put the sentences 'into logical order' using three paragraphs. Assessment of the sequencing was based on the original text, and versions suggested by three non-naive native speakers, who also gave their views on which combinations of sentences would not conventionally be considered logically acceptable. Measured on the number of acceptable sentence sequences, scores by the FD and FI subgroups on this very straightforward, 'familiarising' text were not significantly different.

The second task presented an out-of-sequence set of notes for a short essay. The problem posed was again logical ordering. With the same evaluation criteria using insights from the same informants the null hypothesis of no significant difference in FD:FI performance was once more not refuted.
In the third task, a longer (16-sentence) text was broken down into disordered sequence with participants asked to suggest a reconstitution by putting the letters designating each sentence into a preferred order. They were also asked to suggest inter-paragraph divisions and to select the sentences they considered covered key points for a summary of the text. Assessment was again based on the sequence in the original text and the native speaker versions and views. This time, according to the same formula used before for small-sample independent groups, the FI/FD difference is significant at $p < .001$, the FI sub-group averaging 15 acceptable sequences, the FD's 11.7. This last experiment suggests that with a greater amount of disorganised information to process and re-organise, the claimed ability of field independent participants to discriminate, articulate and impose a structure on information experienced does help with logical sequencing. There was no significant difference in the number of paragraphs suggested by the two sub-groups but the FI group made fewer unacceptable paragraph divisions and selected a higher proportion of 'key points' that were in agreement with the native speaker informants. (The latter two inferences are not tested statistically, the first because it would be tautologously affected by the sentence sequencing factor, the second because the informants themselves disagreed fairly significantly over the quantity and nature of the 'key points'.)

In Experiment 2 (see Appendix 2.2) I tried to get away from an emphasis on the linear organisation of text to focus on aspects of non-linear information structure (Johns (1978) see Chapter Two, Figure 2.6). Participants were first asked to match three short texts with three diagrammatic information structure patterns (flow chart, tree diagram
and matrix). The information from the texts was then to be put into the diagrams in summary form. The second task asked them to summarise the main points of two more short texts using information structure diagrams if they thought them helpful. The first task was scored according to the number of points accurately assigned to the appropriate part of the given information structure diagram. On this basis \( H^0 = \text{FD and FI participants will process information through a given structure with equal efficiency,} \) was refuted with a statistical significance of \( p < .02 \).

Although all participants except one matched texts and structure diagrams correctly, the FI sub-group were significantly better at selecting suitable information to fit the requirements of the given structures.

The second task was scored on two criteria, the quality and quantity of information processed (the facts or the 'message content' (Holes 1972)) and the use of information structure diagrams for their presentation. On the first criterion there was no significant difference between the FD and FI sub-groups. On the information structure criterion, however, the FI's again performed better with a difference on the appropriate t-test significant at \( p < .05 \). I also considered it relevant to compare overall performances on both tasks, that is measures of participants' efficiency in matching texts to structures, selecting key points for summarising and organising them appropriately. On this more general information processing measure, the FI group also performed significantly better (\( p < .05 \)).

However, it would not be wise to infer too broadly from Experiment 2 and
for reasons some of the participants themselves imply in the written comment on information structure diagrams I invited at the end of the experiment. Participant L writes:

".... it is easy to use a diagram if the passage is descriptive and proceeding according to stages while if it is narrative it is almost impossible."

And H:

".... it's suitable for me when I take notes like process or flow chart .... but general notes, I always use letters like a, l, to identify in my note-taking."

While I comments:

"For me it has been a useful way of studying and teaching, and I think is a very good tool for learning."

My experiment has focused on the structure of the subject matter of the texts rather than Johns' (1978) more ambitious probes into writers' logical structuring of discourse itself (though of course the two are often related). My participants see the experiment as an exercise in matching note-taking techniques to subject matter rather than in tracing the patterns of developments of arguments. And a final reservation on too broad an interpretation of my findings is that I have restricted my tasks to very short texts, in the interests of objective measurement of performances. I probably have valid findings, but only on fairly narrow text processing skills. (Rank order correlations between overall performance on Experiment 2 and my broader but problematic Test 3 task (see Chapter Six) are .39 and .5 on pre- and post-test respectively).

In Experiment 3 (Appendix 2.2), I set out to investigate some of the more affective and social implications of field dependency as well as its cognitive features. My experimental design was as follows:

1. The 27 participants were divided into 4 sub-groups with field
dependency as the key independent variable. Experimental Group 1 (XG1) was the field-dependent group with a mean score on the GEFT of 4.9. XG2 and XG3 were field-independent as their respective GEFT mean scores of 15.3 and 15.5 indicate. The fourth group (CG) was not a control group in the strict psychometric sense of a group undergoing no 'treatment' but nevertheless being measured on the dependent variable. Since my focus was on process more than product in this experiment this kind of control group was not so appropriate. CG did, however, serve certain control purposes since it was a mixed FD/FI group (GEFT $\bar{x}$=8.9 SD 6.9) and underwent a 'normal' teacher-conducted 'treatment' instead of the independent learning procedures followed by the three XGs.

2. With only 27 participants, in 4 sub-groups, it is impossible to control all 'extraneous' variance (and with my quasi-experimental, ethnographic approach here it is not absolutely essential to do so). Although the sub-groups were matched as far as they could be in terms of regional background and 'arts' vs. 'science' specialisms, the primacy of field-dependency measures did not allow me to balance the male/female distribution (XG2 was all-male) or TL level. Inter-group variation on the latter, however, was not extreme, with mean all-pre-test scores as follows:

XG1 = 48.3%; XG2 = 53.1%; XG3 = 55.3%; CG = 45.6%

3. The hypothetical area in which I expected my analysis to give insight is defined from claimed characteristics of field dependency such as: degree of facility in balancing overall structures and their parts, degree of analytic tendency, degree of self-directedness, degree of attention to human content of the social setting (see my summary of field dependent/field independent styles, Chapter Two, 3.5).

4. A learning activity that gave scope for the investigation
of these characteristics and that fitted in with the design of the ELTI programme was devised. Members of all three experimental groups were given copies of a 2800-word text, *Urbanisation and Political Protest* (Open University 1970). In a very brief plenary session they were asked to move to separate rooms to study the material (in any way they liked) and to produce a 'group report' for submission at the end of the same morning (ie at 12.20 pm after a 9.45 am start and with a coffee break from 11.00 to 11.30). The control group would meanwhile be spending the same time working on the same material in a normal classroom context. All activities in all four venues would be tape-recorded. There would be no teacher presence with the three experimental groups.

5. Data analysis and interpretation for Experiment 3 will again combine quantitative and qualitative approaches. I have the text, the written reports from each sub-group and scores on a cloze-passage recall test administered four days after the event. These are all susceptible to various forms of quantitative analysis. More important for my current purposes, though, is the interpretation of events and processes as they occur during the activities recorded. My qualitative approach to these will learn from the ethnomethodologists in its micro-sociological focus (Garfinkel 1967), and its concern with the 'process of making meaning' (Primmer 1979), except that my interpretive framework will be the cognitive/affective and social dimensions of my hypothesis rather than interactional norms. It will not be a matter of:

"In this utterance Toby indicates that he has accepted membership by giving a response to the teacher indicating his membership" (Primmer 1979 p.122).

but rather:
"Participants D and K are using evidence critically here, arguing about logic seeking a balance between underlying structure and discrete parts, which could reflect their field independent tendency" (an interpretation typical of some of what happens in XG3).

Figure 7.3 summarises the chronology and modes of activity selected by the three experimental groups. The transcriptions and interpretations that follow it attempt to draw inferences about the independent variable of cognitive style.

The chart opposite immediately reveals interesting differences in approaches to the task. I shall compare first of all the activities of XG1 and XG2, then see if the hypothetical inferences drawn are supported or not in an examination of what happened with XG3.

There is evidence in the FD group (XG1) of the difficulties in sorting out woods and trees that is sometimes associated with the field dependent cognitive style. Although the originally declared plan is for all participants to begin by reading (or at least skimming) the whole text, the actual general reading period lasts only two or three minutes before there is an apparent change of plan, with participants deciding to assign themselves only specified parts of the text to read. (We are reminded that no fewer than 8 of the 10 high field-dependent members of my whole group are described explicitly as having problems with organisation or selection on the 'extra-linguistic factor' cards in the continuous assessment system). Perhaps XG1 never really recovers from the decision to focus on isolated parts of the text at the expense of the whole; in any event there is remarkably little discussion of how the eventual pair or individual précis of the parts fit into the inherent (and quite explicit) logical argument of the text.
<table>
<thead>
<tr>
<th>Activities</th>
<th>XG1 (field dependent)</th>
<th>XG2 (field independent)</th>
<th>XG3 (field independent)</th>
<th>Time (mins)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Organise: all read all</td>
<td>1. Organise: all read all</td>
<td>1. Organise: all read all</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>2. Silent skim-read</td>
<td>2. Silent reading and individual note-taking.</td>
<td>2. Silent reading and individual note-taking.</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>3. Reorganise: pairs to read specified parts</td>
<td></td>
<td></td>
<td>2. Silent reading and individual note-taking.</td>
<td>30</td>
</tr>
<tr>
<td>4. Silent reading of specified parts of text; note-taking</td>
<td></td>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>5. Individuals read notes on their parts of text.</td>
<td></td>
<td></td>
<td></td>
<td>50</td>
</tr>
<tr>
<td>6. Writing up of report by 1 participant from others' notes and comments.</td>
<td></td>
<td></td>
<td></td>
<td>60</td>
</tr>
<tr>
<td>7. Report read out to group.</td>
<td></td>
<td></td>
<td></td>
<td>70</td>
</tr>
<tr>
<td>8. Discussion of migrant question in own country.</td>
<td></td>
<td></td>
<td></td>
<td>80</td>
</tr>
</tbody>
</table>

**Plenary report-back session**

Figure 7.3: Summary of group activities during Experiment 3
Most of the discussion is section-specific, even serialistic and the final report, written up by one participant as dictated by someone who has read and taken notes on each section, is a rather disjointed and unbalanced affair. It seems justifiable to make a connection here with Ramsden's 'surface approach' where 'unrelatedness' is a characteristic described as follows:

"Defining the task as separate or its parts as discrete .... treat the task as an isolated phenomenon: confront the material as separate from other ideas or materials, or from the general purpose of the task to which it relates: focus on the elements of the task rather than the whole" (Ramsden 1981 p.145).

Entwistle et al. (1981) relate this to their 'strategic orientation', a concern with what is required rather than what you feel intrinsically motivated to do. The following transcriptions from XG1's first and third activities provides evidence of unrelatedness and a strategic approach caused, perhaps, by problems sorting out the cognitive field and its constituent areas:

Transcript 1:

X: Who will take the ... the first?
J: Or let us have ... er
X: One ... two ... three ... four ...
J: And we are ... er ... we are seven so we can divide, everyone has ...
U: We can read all ...
J: All? No it would take a long time.
X: Long time. Divide the job.
J: Yes, I think we ...
U: But if ... I know this week ... take a test I know. We ... we didn't (laughs) ... only this part we can't make anything.
J: Er...but we have [no time...] [so we have to read it very quickly]
X: ...............[ no time] [who will take the part one?.......]
J: Now ... read our part, yes, so ...
P: But I think the first thing is to check the arrange in the properly order to check if it is and then to divide the work.
U: Yes.
J: Quickly, quickly, have an idea about five minutes to ... everyone to read this.
A: Throughout ... er all through the paper?
J: These ... all the papers, yes.
A: All of our ... first we will divide these into different ...
J: We are going to divide them because this is the quickest way I think. But five minutes I think to have all of us have an idea about the topics.
A: Yes.
N: That's right.
P: And ... if anyone can to check the clock .... the clock would be a good idea ...

But the read-through is aborted and P certainly has not had his way on 'checking the order'. The pressure for discrete tasks based on surface format divisions of the text soon builds again:

**Transcript 2**
X: OK we can divide now.
J: We haven't finished yet, (X).
X: I'm going .... I'm going to read the headings.
J: To read what?
X: The heading of this.
J: Ah to read ... to write them.
X: Yes. (X is at the blackboard)
J: Yes.
(a silence)
J: I think we can divide us [to seven ]
P: ........................[by numbers]
P: By numbers ....... [you know we are veinte-y-tres, twenty-three ]
J: By numbers, yes... [because we are seven. We are not responsible]
P: [Divided]
J: [All ...] every to ... to have er ... everyone has a part.
P: Yes.
A: Yeah, yeah.
J: You see, it's more use ... so we can divide it that's what (P) said to numbers ... Right? ...[because this is ] twenty-one.
P: ........................[but realise that]
N: Twenty-three.
J: Twenty-three, yes, if we divide us we [can take..er] everyone take four.
P: ........................[but take....] ...
P: No, but take care that here there are some subjects that are close united then it would be better ... uno ... one, two, three, four, five, six...seven, eight, nine... for two persons and to ... work together with the first topic.
J: It's OK, it's OK (P), yeah.
P: Because if we divide for instance we shall have ...
J: But [er ... there is ]
P: ... [some of them who] is beginning here and then beginning and beginning and finishing.
X: (still at BB) (P), (P) er ... 23 points.
J: 23 ... er points, yes.
P: One, two, three, four, five, six.
N: They are divided in two sets.
J: Now you can ...
X: No, no (J) ... two to three ...
J: But is necessary not to cut the subject, then ... it's a good idea to have the 23 points but there are three subjects.

N: Four ... four subjects (P).

P: Four subjects?

N: Yes ... with this ...

P: One ...

X: Oh yes.

N: Two.

P: Two.

N: Three and the last one.

P: Four.

J: But we have quickly to get the report you see. So we are going ...

P: If you like it I take for instance ... but there are four subjects and then it ...

N: Something like ... er ... general information ... the second one ...

P: It's quite difficult to divide it!

And it continues to be so. This re-organising phase lasts ten hectic minutes, going on to suggest division of labour by topics, (the 'subjects' referred to in the transcription are main headings in the text) by pages, by 'points' (the 23 numbered paragraphs) but never in terms of the actual theme or logic of the text itself. The reading phase that ensues is much shorter than in the other sub-groups, with all pairs and the one individual reader (P) restricting themselves to their own section, the pairs dividing up their assignment even further into two separate sub-tasks. The writing up phase is brief, too, with no inter-topic discussion. As a result, XG1 has twenty minutes left at the end when they willingly accept a passing teacher's suggestion to discuss the question of migrant/refugee matters in their own countries. Again the interaction is lively but, as throughout with this group, very good natured. Field dependency revealing its attentiveness to the 'human content of the environment' perhaps. This latter characteristic is suggested by the fact that there is more 'chat' in XG1 than in the other group activities sometimes, like the following vignette, only very loosely related to the task:
Transcript 3

U: Calcutta (the case study in the text) where is Calcutta?
A: The migrants of the city of Calcutta ...
U: Where is Calcutta ... Pakistan?
N: India.
U: India ... India ...
A: The fourth ... er ...
J: Before the division, I think so.
A: The fourth city of the world. The fourth city of the world.
U: Oh.
J: The fourth.
U: Oh! ... First?
A: Tokyo ... of Japan.
U: Two ... New York?
A: And second New York.
J: In population you mean?
A: In population and area.
J: And the area.
U: Ooh!
J: Ah.
U: I don't know.
J: Me too.
U: Heard the first time.
A: Third is ... third London, you know this?
U: Third London? Oh.
U: Big.
J: England is ... er.
U: Third. London is third ... city. Third big city ... in the world.
J: London? The third big city?
A: Of the world.
U: Fourth ... Calcutta.
A: Fourth is Calcutta.
U: In India.
J: Oh. I don't know!

This is not only an interesting instance of learning in social interaction: it also has pre-sessional pedagogic implications I shall return to in the report on the experimental control group below.

While XG2 were handling their task in a fashion variously reflecting cognitive/affective and social aspects of field dependence, the scene with the high field-independent XG2 was very different. As Figure 7.3 shows, they quickly decide on everyone reading the whole text then settle down to individual silent reading/note-taking for fifty
uninterrupted minutes. In this beginning and the discussion, then group report writing activities that follow, XG2 seem to display some of the analytic, self-directed, problem-solving traits associated with their observed cognitive style. Their approach could also be interpreted according to Ramsden's 1981 categories of the 'deep' task approach, eg

"integrating the parts into a whole ... desire to relate parts of the task to each other or the task to other relevant knowledge ... active attempts to think about the relationships between different parts of the material eg relate evidence to conclusion ... integrating the whole with its purpose. Indicate intention to impose meaning: think about the underlying structure ...; impose a pattern on the whole task" (Ramsden 1981 p.145).

And there could be complementary evidence of a more intrinsically motivated meaning orientation, less of the strategic kind.

Entwistle et al. (1981) also have field independence loading on a factor alongside 'reflectiveness', verbal reasoning, use of evidence. Yet this particular experimental group does not contain high Q1 scorers (at $\bar{x} = 5.0$ it is the lowest Q1 group of all the 4) so we might expect, perhaps, a lesser degree of scepticism in their reflections, reasoning and sifting of evidence.

In his short briefing before the experimental task, the facilitator (this researcher), refers rather frivolously to 'chalk' and 'cheese' with reference to the way sub-groups had reacted to previous learning tasks on the ELTI course. The comparison is apt, however, with regard to XG1 and XG2 in that almost any selected transcriptions of events from the two groups' approaches and activities in Experiment 3 would underline differences of style and substance. Since XG2 were so quick and decisive in their self-organisation for the task, there is, literally, nothing to compare with the excerpts from XG1's lively and
lengthy attempt to decide who was going to read what and how (see Transcripts 1 and 2 above). Nor is there any substantial social 'chat' during the 140 minutes during which the group were working together. And apart from Participant B's reminders of the time factor during the last 20 minutes of the task, there is almost no sign of the concern with perceived external constraints so evident with XG1. The text, its topic and its logic, rather than what anyone else supposedly wanted done with it, was the interest and concern for nearly everyone in XG2, nearly all the time. The excerpt transcriptions below catch the tone of the events, but they could easily be substituted by others.

The topic and text discussion phase begins in the analytic, meaning (and form)-oriented way in which it is to continue:

Transcript 4
K: Are we going to follow the plan here or are we going to make our own plan?
H: We ... because from this I think we can separate into parts, one we talk about ... er ... urban ... urbanisation ... er ... we can talk how it occur.
D: 'Urbanisation and protest political behaviour'.
H: Because ... er ... we can think why is ... er ... urban ... urbanisation happen. You know there are many factor like as ... er ... we need education ... er ... we need ... er ... because ... er ... education, communication and another material to force the people in the rural go to the city.
D: I'm sorry but I don't think so. Because ... er ... this, this is almost a survey.
H: This is survey.
D: Survey, a case study. A survey ... after that ... I mean we cannot separate urbanisation and political protest ... I think we cannot separate this.
K: It is not possible to separate because ... mm ...
B: It [is ... er ... a] study to test the hypothesis about rural
D: [He says ........]
migrants ... political ...
H: ... political protest.
D: This is a hypothesis that is ... he wants to test.
B: And here is ... er ... he use ... er ... Calcutta to test ... er.
H: For example.
(looking to text).
H: Because ... er ... it is as Calcutta it happen caused by by the urban...isation.
K: Let us see the pronunciation of this word! ... I think it is urbanisation [ˈɔːbrənaɪzəʃən] maybe ....
H: Urbanisation (əˈbʌrniˌzaʃən).
K: Urbanisation. (əˈbʌrniˌzaʃən).
z: Urbanise ... nise.
(Dictionary consulted)
Z: It doesn't matter now ... the pronunciation.
K: Ah ... it is important.
Z: You think?
K: Yes ... You're talking about this word.
B: Urbanisation . (əˈbʌrniˌzaʃən] ... 'usually ... accompanied with ... modernisation'.
K: Urbane ... urbane ... there is not urbanisation here.
H: Urban...ization I think urbanise.
K: It is not there.
D: There is a survey: 'Urbanisation and political protest' ... a case study. He's saying that urban growth is something creates violence and political extremism. There is a portrait in here. 'Protest political behaviour in urban areas is a product of rural immigration ... rural migration'. He tests this ... in Calcutta ... because in Calcutta he could separate... it was a good case study because he could separate refugees, migrants and residents ... and he found that refugees' and residents' votes against the government ... er ... urbanisation.
K: I don't know.
z: I follow only
D: Refugees' and residents' votes against the governing party.
K: Mm.
D: But not the migrants and this is ... he thinks that this is ... surprising because residents and refugees ... er ... they have much more conditions.
z: About the character ... characteristics of these people ... yes for example and residents and refugees ... educated people ...
D: Yes.
z: But migration it means only the worker.
D: Can you explain?
z: I mean ... er ... er ... the ... the major ... er ... the major vote from resident and refugee ... refugee because he has a good ... and the own history and he has ... a ... educated people.
D: Yes.

The writing-up phase is similarly analytic and questioning, again with the logical argument intense yet without talk for its own sake, fear of silence or an unwillingness to let the rapporteur, the Cartesian K, actually write in peace. It is a serious-minded, relation-seeking, text-focused process:

Transcript 5
D: And you can follow 'and': "And the growth of urban centres as migrants move from rural countryside to the cities da-da-da-da-da-da-da-da".
K: Yes ... 'and' ... '(Pause for reference to text)
Z: I think his test for his hypothesis is fine ... His only prediction that it is the migrants rather than the native born who are responsible for urban political extremism.

D: Yes ... and he summarises ...

Z: It is his thesis.

D: He summarises the protest here.

Z: Had he tried ... or attempt to test his hypothesis ... this is ...

D: Hypothesis is this.

Z: Mm?

D: Hypothesis is that: hypothesis: 'protest political behaviour in urban areas is a product of rural migration' ... Here he explained the whole thing ...

Z: Mm.

D: And here he summarises ... the hypothesis ... He puts just 'hypothesis' ... Because if he had to write this he's gonna be half a page in ...

(Pause)

K: We can say ... in a short ...

D: Mm?

K: This is ... what we have said here is said in another way here: 'It is logically assumed that it is the migrants rather than native-born city dwellers who are responsible for urban political extremism and violence'.

H: He says 'effect'...

Z: No ... here another idea ... he related ... this problem to the ... the migrants.

H: It is clear enough because he said ...

Z: but not to the residents ... he related this problem ...

K: Yes ...

Z: to the migrants and not to the residents.

K: Yes, this is the same.

H: This is a hypothesis.

K: Migrants, 'as migrants move to the cities results in increasing the violence and political extreme'.

H: Mm. I think this a hypothesis ... He said ...

D: We can follow the hypothesis...

H: 'Typical ... migrants ... living in urban ...' ... this a hypothesis ... I think we start with ... OK ... the next we give ... er

Z: Then in this case before we ... we have to speak about ...

K: No we must be care because we ... are already talking about the hypothesis ... the hypothesis here ... er ... the fact is there is urbanis .. urbanisation process the growth of ... er urbanised areas, there is one fact. Another fact is ... the increase of violence. But the relation is not obvious. The relation is in the hypothesis.

D: Yeah.

K: So we must show only the two facts here. One fact: the urbanization, second fact the increase of violence, violence and political extreme; and then in the second, the anti-social behaviour of migrants. I don't know if you follow my idea ... and the hypotese gives the relation between the migration and the development.
This is an hypothese which must be ... verified in ... by the illustration. But if we say already here that er ... the migration increases the violence ... er ... we give already the ... the hypothesis.

Z: Yes, he gives his hypothesis and then he try here to give some reasons.
K: Yes, the rea...
Z: Unsocial behaviour of the migrants. Why? It was here described.
B: I think your ...
Z: He tried to describe ...
K: Yes ... you, we are not going to follow the plan here. We have done our own plan.
Z: Mm.
K: And we can take the informations here to ... to make our text ... Yes.
Z: I see.
K: You see.
Z: So you will speak now about the anti-social ...
K: Yes ...
Z: behaviour ...
K: Yeah.
Z: and this will be the relation or the connection ...
K: Yes.
Z: ... between the two factors.
K: Yes, exactly. That's right.

In spite of, or because of, the detailed analysis of the text, XG2 produce a lean, logical written report with clearcut sections following the structure they imposed on the text not the headings of the original. They completed their report precisely on time, leaving no gaps during or after the intensive study process for recapitulation or, again unlike XG1, for the socially sensitive 'bringing in' of participants who had not said much.

But as if to remind us that no independent variable will cause incontrovertibly predictable behaviour when people are involved, XG3, a group with an even higher field-independent score on the GEFT (x = 15.5) did not handle the experimental task quite so clearly in line with expectations. In the task organising phase they neither take the immediate strategic surface-study decision to read the text according to its original sub-heading or paragraph divisions (as XG1 did), nor the
immediate integrated-meaning, deep-study approach of everyone reading everything as did XG2. Instead, they compromise on a plan for pairs or individuals to read the whole text 'if they like' but note-taking on specific sections, though not four sections as the text headings superficially indicate but three because:

"G: But I think ... er ... I think ... er ... 'How migrates ... er ... migrants voted' and 'How refugees voted' is the same subject."

And as Y indicates, there is concern about splitting up the reading task:

"Y: What ... er ... what will be the combin ... er ... what will be the ... er ... completion of the ... er ... er ... of this study? How we will complete, how we will ... link with ... er ... with the topics I mean?"

In the event, it is clear that some participants do read the whole text but their focus on their particular sections and close analytic approach to them during the reading, note-taking and pair discussion phases, means that the group never really re-integrates for consideration of the total text or their whole report. The latter is thus long, fragmented, with the quality of each part reflecting the success of the pairs or individuals writing it rather than, as was the case with XG2, the collective views and skills of the group. (Perhaps their highish $Q_1$ mean ($\bar{x} = 6.6$) intrudes here. There is evidence that E and 0, for example, want to experiment with the part-reading approach, because they have used the alternative approach in previous course activities).

In the pair discussions, however, the strong analytic concern with relationships, underlying structure and patterns associated with field independence and a meaning orientation does emerge. The following excerpt transcriptions illustrate the point.
Transcript 6

"G: Migrants causes violence and extremism.
O: Yes.
G: This is the main point only.
O: Mm, Mm.
G: This is main point because ... er ... first of ... in the first paragraph he explain the urbanisation ...
O: Problem, yes, G, but I think it is important because these ... these things: they are after ... after doing, these, these people... this is ... they are talking about ... er ... the proportion of people that is vote against ...
G: Yes ...
O: that is vote in favour of the ... of the Congress, so they are analysing the people, they are dividing the people in refugees, migrants then ... and residents, so it's important this point, I think it is important.
G: Hypothesis: the migrants are responsible for ...
O: ... increase in violence.
G: Yes, yes it's the same ...

and

Transcript 7

"Y: (reading from his report) 'Although compared to residents the migrants are poorly clothed, housed and educated and away from their families, the study on the simplistic theory of material deprivation will ...'
E: No 'will' ... er 'leads us'
Y: 'Would lead us'.
E: No ... it's a theory that we just touched ... that this lady had done ... no 'will' no ... 'deprivations lead us to expect the reverse of our ...'
Y: O.K..

or

Transcript 8

"R: (reading S's summary report) 'they had to ... be migrated by ...
S: 'Communal'
R: Comm ... mm?
S: 'Communal' ... 'communal disturbances'
R: 'communal'
S: 'communal'
R: Mm.
S: 'communal disturbances' means ... the
R: Mm mm.
S: the ...
R: 'communal disturbances' Mm mm [noise from other pairs]
S: ... means the different races or different religions.
R: ... different religions mm mm.
S: Yes.
R: Mm... mm.
S: and ... disturbances coming from different races or different rel...
R: 'They had to be migrated by communal disturbances' ... ah their religions. Mm mm.
S: Because the Hindus ...
R: Yes, yes.
Perhaps this last rather groping interaction should lead us on to the events of the control group lesson where the teacher concerned organised the following approach to the task:

1. Brief teacher introduction to the concepts of hypothesis, evidence and conclusions in the methodology of normal science 'to get the students to recognise the organisation of a text' (teacher's own taped report back on the activity). Aim to 'move the students away from the idea of burying their heads in a text and not standing back and looking at it'.

2. Students encouraged to exemplify the concepts in their own terms and discuss.

3. Connection made with the text itself; students to read whole text then work in two pairs and a threesome on
   .1 the writer's purpose (hypothesis-testing)
   .2 evidence
   .3 conclusions
   this phase monitored informally by the teacher.

4. Leading to three overlapping reports under the headings in 3.

In his very frank and useful report back, the teacher concerned discusses constraints which, he felt, caused the lesson to fail to achieve its objectives. The group was too mixed in TL level with I, 'extremely advanced in comparison with someone like W', three people in the group with 'very weak English' and the others 'ranging from somewhere between intermediate and advanced'. The text itself was 'extremely long' though of realistic length 'in that this is the kind
of thing that they would be expected to do at university or college' but not the kind of text 'one would normally use in the classroom'. The result was 'far too much of me and far too little of them'. Only I, C, F and L definitely 'got something out of the lesson'; T, W and V 'got nothing out of it'. For the latter three, the teacher felt that to talk about the organisation of the text was pointless, given their TL level. There was 'very little participation, very little chat'.

As indicated above, two of the transcriptions (Transcripts 3 and 8) from XG1 and XG3, may seem to be evidence that the task was beyond some of the participants in the experimental groups, too, though given the teacher-less context of their activities they have no choice but to participate and learn. But then, the control group teacher is probably, like most good teachers, rather too self-critical in his assessment of the learning that went on in his lesson.

The cloze test, with carefully selected gaps to test recall of key points in the text as a whole is not, on its own, a measure of the 'dependent variable' (learning), since the emphasis in this experiment was on the process rather than a product. The results should thus be evaluated taking account of my suggestions above that XG2 (FI+) seemed to get most out of the task through their successful organisation, deep-meaning orientation and the balance they achieved between analysis and integration; that XG3 also delved deep, though seeing more of the trees than the woods and that XG1 were more strategically or socially than meaning oriented, producing a fragmented report but interacting interestingly. Test scores were as follows:
Maximum = 15

<table>
<thead>
<tr>
<th>Group</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>XG1 (FD+)</td>
<td>9.1</td>
<td>3.9</td>
</tr>
<tr>
<td>XG2 (FI+)</td>
<td>11.6</td>
<td>1.1</td>
</tr>
<tr>
<td>XG3 (FI+)</td>
<td>10.9</td>
<td>2.3</td>
</tr>
<tr>
<td>CG (Mixed)</td>
<td>8.4</td>
<td>3.9</td>
</tr>
</tbody>
</table>

These performance measures permit the following interpretations.

1. That XG2, a field-independent group, apparently approaching the task as its independent variable predicted it would, not only performed best on the recall test but helped individual participants to perform better than their TL level (pre- and post-test scores combined) would have predicted. 5 out of the 6 ranked higher on the X3 test than on my TL test batteries; the somewhat frustrated D (see Transcripts 4 and 5 above) being the only exception.

2. That XG3 averaged next best on the recall test but paid a price for their decision to fragment the reading and report-writing task. One of their members (S) certainly 'missed' the general structure and meaning of the text because of this, and in fact four out of the six in XG3 ranked lower on the recall test than their overall TL ranking.

3. That the field-dependent XG1 performed erratically on the recall test (note the high standard deviation of scores). The evidence is that performance was more influenced by the degree and depth of members' participation in the ad hoc discussion of the migrant issue in their own countries than by their rather disorganised, strategically-motivated handling of the set task itself. Certainly J, who was so concerned to hurry completion of the task and U, who predicted that there would be a test, did worse on the test than their TL level would have predicted. P, the one member who asked for a focus on the thesis and hypothesis of the text, A and Q who also got very involved in the C1
discussion, did well. And X, who actually wrote the group report, scored highest of all 27 participants.

4. That the mixed field dependency control group scores lowest overall on the recall test, though not entirely in line with their teacher's prediction, which was mainly on the basis of his perception of their TL level. T, highly field-dependent on the GEFT and thought by the teacher to have 'got nothing out of' the lesson, in fact ranked rather higher than her TL order whereas C, a maximum FI+ scorer on the GEFT and identified as one of those benefiting from the lesson, ranked rather lower in terms of a TL level-based prediction.

5. That cognitive stylists (eg Pask and Scott) who advocate the matching of teaching approach and cognitive style may well be right in principle. However, as my experiment shows, other variables, other constraints so easily mitigate against learners being able to study in the way that best suits them. XG2 managed to set about the task in a way that worked for them in process and measured by product. Yet XG3, under similar conditions, did not and of XG1 it can only be said the some of the field-dependent members would have been better with a teacher to guide them.

The notion of field dependency emerges from my experiments as an interesting variable in a study of EAP learners. The advantage certainly seems to be with the participants rated as field-independent. If we consider the original 9 FI+ learners (see Experiments 1 and 2) we find that they cope consistently well with the various experimental tasks. The implication is not so much that field independence can be equated with a preference for self-directed modes of learning but rather that, faced with a variety of EAP learning conditions, the FI+ students
are more likely to have the analytic judgment to make sound independent decisions on how best to handle them.

4. Motivation and Attitude

McDonough (1981) provides a useful reminder of the broad, sometimes unclearly delineated concepts covered by the term 'motivation' and which I attempted to clarify in Chapter Two.

"... one should distinguish at least between (a) energy (b) willingness to learn (c) perseverance (d) interest (e) enjoyment of lessons (f) incentives (g) benefits of knowing the language" (McDonough 1981 p.143).

My main explicit, 'objective' elicitation device on attitude/motivation in Phase One is the Motivation/Attitude Questionnaire (MAQ/1) (Appendix 2.3) adapted from Gardner and Lambert (1972) and Hartley et al. (1971). But this is only a starting point, inviting comparisons of findings with teacher comment, the participants' own more freely expressed views (especially from their Test 5 essays) and, most important, the data collected during Phase Two, which includes MAQ/2 and MAQ/3 for longitudinal follow up to the first questionnaire.

Results from my adaptation of Gardner and Lambert's orientation indices should be seen in the light of my participants' compulsory presence on the ELTI pre-sessional course prior to the specialist programmes for which they held ODA awards. The situation itself is inherently 'instrumental' so we could expect my participants to rank my added orientation reason, 'I am studying English because it is important for my specialist studies' top (I use rank ordering in place of Gardner and Lambert's original 2 points for an 'integrative' response and 1 for an instrumental). The group was unambiguous on this point, with the EAP reason (mean rank order = 1.7), the first choice for 20
out of the 27. The rankings thereafter, however, remind us that, used with overseas students in C₂ at least, the instrumental/integrative distinction is rather gross even if the elicitation device itself is useful. The other reasons for learning English were ranked as follows:

2. for 'contact with more and varied people' ($\bar{x} = 3.0$)  
3. 'for a better understanding of British people and their way of life' ($\bar{x} = 3.37$)  
4. 'to make me a better educated person' ($\bar{x} = 3.44$)  
5. 'to get a better job' ($\bar{x} = 3.56$).

The immediate C₂ pre-sessional context probably turns the 'integrative' reasons (2. and 3.) into instrumental ones; the group's rank ordering could be seen as their chronological sequencing of priorities, invited by the aspect of the verb on the questionnaire 'I am studying English because ....'. This makes Gardner and Lambert's next set of statements revealing, too. Here students are asked to rate rather more forward-looking reasons for learning English e.g. 'A good knowledge of English will be useful in getting me a better job ... '. Again they respond, as a group, with encouraging logic. The rank-ordered mean scale scores on the values of a good knowledge of English are:

1. to complete specialist studies (6.26)  
2. 'getting a better job sometime in the future' (5.15)  
3. making 'friends more easily among English-speaking people' (4.63)  
4. contact with non-native English users (4.41)  
5. understanding the British and their way of life (4.26)  
6. as a foreign language helping one to 'be more highly regarded socially' (4.19)  
7. 'to begin to think and behave' like British people (2.56)  
8. because 'no one is really educated unless he is fluent in English' (2.0).

The order and the scores are interesting again for their 'instrumental'/ 'integrative' mix as they reflect longer-term orientations. Hence, for
example, the climb to second place of the EOP interest. For the
crucial longitudinal interpretation, however, it is necessary to
disembib the double statements that the Gardner and Lambert format
seems to entail. If you give a positive rating to a statement like
"A good knowledge of English will help me to understand better the
British people and their way of life"
bys responding that this is 'definitely my feeling' are you not being
forced to imply that you definitely feel you want to understand the
British way of life? In MAQ/2 and /3 (see Chapter Eight), I split such
statements, first, for example, inviting a response on whether partic-
ipants feel it is important to understand the British way of life well,
then, separately, to their view on the importance of English for the
purpose. This seems particularly relevant once they are out into their
real C2 EAP lives. At this Phase One stage, I can leave the question
of TL orientation priorities at the group analysis level. The data
on individual differences of orientation is there, however, for later
longitudinal comparisons.

The description of plans elicited in Pre- and Post-Test 5 provides more
individual evidence on orientational priorities. The standard response,
eg from Participant F, talks of specialist training hopes and needs,
the TL question, C2 and, finally, spare time plans. For some partic-
ipants however, especially those from countries with historical
connections with Britain, (eg A, L, Q and R) there is an emphasis on
the belief in the quality of the education expected. The Phase Two
study will reveal whether these high expectations prove to be a positive
motivating force or increase the likelihood of disappointment. Some
participants (eg V and X) talk almost exclusively of the TL question,
but not so much as a sign of an 'integrative' orientation as of a fear
of failure caused by their TL level. D and U, on the other hand, only discuss their specific training needs. Again, it will be interesting to see if having a very clear view of what you want may increase the possibility of disappointment if your precise needs are not precisely met. And is J's concern with 'the relationship with others', with 'being always as a group' to prove a helpful drive to get into C₂ or is it a sign of the wrong sense of priorities? Phase Two essays should provide useful insight.

Section Two of MAQ/l attempts to go beyond the orientational aspect of motivation, that is participants' beliefs about reality and their values, what they would like to be true (cf Ajzen and Fishbein, Chapter Two above), towards their attitudes, their behavioural intensities, in the first, still largely anticipatory stages of achieving what they want (see Chapter Two, Figure 2.2). I accept the potential influences of social acceptability, self-ignorance, response set and so on (Oller 1979;1 discussed in Chapter Two) on responses to questions on habits and feelings. But with my small sample of intimately known participants, these influences should be both detectable and interpretable. Also, the potential 'threat' of MAQ/l was reduced in that it was administered as part of a group project where the participants themselves had chosen UK overseas students as the topic. (Participant X actually administered the questionnaire with his project-group colleagues briefed to give L₁ help where it was required and with the confidentiality of individual responses assured). Section Two asked for 1 to 5 scale responses on (items A to 0) ELTI course reactions and habits and (items P to Z) on extra-curricular TL-related habits.
The items in both sub-sections allow composite motivational measures
to be formed to give an indication of the varying levels of active
interest in the course and the broader TL context. The top 9 positive
scorers on the ELTI course items are Participants A, O, Q, L, U, X,
Y, M and P. On O, Q, X, Y and P, teacher comment is in agreement with
student self-assessment (though, interestingly, the teachers very
seldom use the term 'motivation', preferring terms such as
'hard-working', 'tries hard', 'ambitious' etc). On A, L, U and M,
however, the continuous assessment verdict is contradictory. It is
not so much that the teachers question the conscientiousness or interest
of the four participants, but that they are, understandably in this
context, looking for indications of self-directed learning approaches
and not finding them in these participants. Whether this is a bad sign
in terms of future perseverance, we shall discover later, perhaps. It
is interesting to note that 5 of the 9 here are high field-dependent
scorers, which might relate to a preference for teacher direction or,
even, to a conforming attitude to questionnaire items. Of the 10 part-
icipants rating their ELTI course approaches most negatively, 5 are
in my high field-independent group against only 2 FD+ students, and
none of those 5 are considered unconscientious by the teachers. And
since all but one of the highly serious, analytic XG2 (see Transcripts
3 and 4 above) happen to find themselves in this relatively low self-
rating group, a connection between field independence and a
self-critical image might be inferred. On the items relating to
extra-curricular TL activities and interest, Participants E, I, J and
F replace M, U, X and Y in the high self-rating group, all the
newcomers being QI+ and all having expressed elsewhere, eg in the
group discussion and in Test 5, an interest in 'real' C2 contact.
Perhaps the Q1+ characteristic of experimenting 'in life generally' and showing 'initiative in introducing change' is emerging here. There is only one Q1+ participant in the group rating itself low on extra-curricular C2 habits but again 4 of the 9 are highly field-independent, 3 of the others quite highly so. But perhaps the most important finding is that on both parts of Section Two and on the overall self-rating assessment, the high raters have a significantly higher TL proficiency according to my post-test battery (administered very shortly after MAQ/l) than the low. Using the appropriate t-test on Section Two as a whole, the difference is significant at p<.05. And I prefer the cyclical interpretation of causal relationships between the motivational ratings and TL performance, ie that a reasonable TL level promotes active involvement in the learning process which in turn helps TL performance.

The final section of MAQ/l uses some of Gardner and Lambert's 'attitudinal ratings', asking participants to rate the British and their own nationals on seven-point semantic differential scales eg

friendly ___: ___: ___: ___: ___: ___: ___: unfriendly

My primary interest here was in possible changes of attitudes during the longitudinal study; MAQ/l findings are thus mainly for later comparisons. The following points, however, summarise what emerged at Phase One:

1. The group as a whole reacted without any signs of extreme ethnocentrism or C2 awe, apparently seeing themselves and the British reasonably positively on all items where there is a 'good' and a 'bad' end to the continuum. One can assume that the third world:old world relationship is implied in the group's rating of the British as
significantly higher than the various C₁s (p .01 or more) on 'successful' (p<.001), 'secure', 'stable', 'mature' and 'leader'. Less situation-bound, perhaps, are the significantly high ratings on 'polite' and 'hard-working'. The only significant difference in the other direction is the higher rating for C₁s on 'friendly' (p<.05), though the difference on 'prejudiced' comes worryingly near to statistical significance.

2. 8 of the group: D, G, I, N, Q, T, W and z, give mean favourable ratings to C₂ which are statistically, significantly higher than their C₁ ratings at p<.02 or above. In all these cases, except perhaps T and z, it is their high rating of C₂ rather than a low rating of C₁ that stands out. It will be noticed that only two of these participants, I and Q, are among the group with the highest self-rated level of C₂ involvement in Section Two. And one of that group, F, actually gives consistent, though not statistically significant, lower C₂ ratings than C₁, as do V and X. There is clearly no simple causal relationship yet between favourable or relatively unfavourable assessments of national characteristics and participants' desire for cross-cultural contacts. Such relationships may emerge as time passes, or perhaps B, C, H, K, P, Y and Z (all bar P in the high field-independent twelve in the group), whose perceptions of C₁ and C₂ characteristics show very little distinction between the two, will find less difference, and thus less difficulty, in adjusting.

5. **The Contribution of Cognitive/affective and Social Variables to Participant Profiles: case studies**

The reports and discussion of my cognitive/affective and social factor investigations have already indicated in the text itself the way in
which such data can be used to supplement or refine TL profiles of participants. On the actual profile forms sent out to receiving institutions (see Appendix 2.4) the non-TL information was restricted only to what we were already almost sure of before there had been time for the analyses of data described here. Thus, Participant O’s TL profile is supplemented under 'Attitude' with comments on her liveliness, commonsense and independence; her positive attitude to the course; her awareness of 'what she needs to concentrate on' and her optimistic view of her 'academic and social future'. Nothing was allowed into these initial profiles unless it had been corroborated by more than one of our assessment sources. O’s performance on the GEFT (maximum FI+) was also included, with an interpretation based on 'official' Witkin et al. lines, not on my own subsequent experimental findings. The numerous hypotheses tested or explored in this chapter now allow me to go much further on individual participant profiles; in the terms of the research design in Figure 5.1 above, it is possible to construct Profile 1 for comparison with Profile 2, based on how participants are at the end of their training period or the first year of it. Taking the same four case studies as in Chapter Six it is possible to illustrate the extra insights now available.

1. Participant C

C, it will be remembered, was a TL learner with her communicative potential not fully realised though with the competence-performance gap closing. We can now also note that she belongs to both the high Q₁ and high field-independent groups with the inference therefore, that she is a questioning, aware person, welcoming new experience, able to analyse and put her own perspective on things. On MAQ/1 she shows a
future EOP as well as a present EAP orientation and she belongs to the group of participants with a balanced view of C₁ and C₂ characteristics. She gives rather negative assessments of her own course and C₂ involvement which ties in with certain teacher comment on her tendency to prefer a 'back seat', especially in oral interaction. This is really the only doubtful note, though, in a generally optimistic profile.

2. Participant L

L's TL profile portrayed him as strong in TL communicative competence and performance but insecure in its study-oriented use. We also know that his Q₁ rating is average, his cognitive style highly field dependent, the latter reflected in his poor to average performance on the experiments. MAQ/1 and related sources suggest a strong orientation towards study in the UK per se, but his own high ratings on active involvement are not reflected in ELTI teacher comment. There is some concern about L's level of awareness of the overall EAP structure and its constituent parts as he leaves for Phase Two.

Participant O

One of the most improved participants in terms of TL now confirmed as high on both Q₁ and FI, demonstrating the inquiring, initiative-taking independence expected on nearly all experiments. MAQ/1 shows her orientation as typical of the group's combination of the 'instrumental' and the 'integrative', which is supported by her clear essay accounts of her desire for academic and C₂ satisfaction. O gives the highest self-assessments of ELTI course and C₂ involvement but this time the teachers generally agree. She is favourable, though selectively so, in her view of the characteristics of the British and her own compatriots. O's full profile is even more optimistic than her original TL one.
Participant T

The TL profile for Participant T was not encouraging, with low test scores all round supported by teacher comment. T is scored as low on Q1 and as extremely field dependent. Her performances on the experiments bore the latter measurement out and there is unfortunately little from other sources to contradict indications of a non-analytic, rather passive and disorganised approach to things. MAQ/l shows an unexplained contradiction on her EAP orientation and her fairly high self-ratings on ELTI and C2 participation are not corroborated by teacher comment. The wide gap between T's ratings of C2 and C1 characteristics is not encouraging especially as her post-test essay already suggests homesickness and settling-in problems. The profile does not support an optimistic prediction.

This chapter has not gone against expectations and produced clearcut findings on neat hypotheses with transparent causal relationships. Where quantitative and qualitative data have been used in their promised combination, a feeling that such an approach may pre-empt falsification has sometimes arisen. Still, the various sources of information and their varying elicitation methodologies have provided insights relevant to my profiling purposes. And we still have Phase Two where, away from the narrower teaching, testing and experimental context, the events of real life can exercise their full falsification potential. So far, the many specific hypotheses constituting Research Hypothesis 1, formulated as relating individual and combined cognitive/affective and social factors to TL learning and use (Chapter Two, Section 6) indicate two-way and cyclical relationships. Cattell's Q1 factor is suggested as causally related to critical awareness as manifested in TL but
not significantly to TL level or progress. It also seems to relate positively to active interest in C₂ (and thus TL) contact, which may be expected to have a longer-term effect on communicative competence (see Chapter Eight). Field independence is indicated as a significant independent variable in narrow TL text-processing tasks and also in the organisation and deep-meaning orientated handling of longer text, at least in an experimental setting. Field dependence, on the other hand, appears to relate causally to a less analytic, more strategically-oriented, if more socially-conscious, approach to the same task. The conventional *intrinsic* interpretation of integrative or instrumental orientation is revealed with my group as potentially misleading, since both concepts emerge here as situational rather than personal influences. My attempts to elicit measures of motivational intensities point to a cyclical connection with TL, but like other attitudinal observations in Phase One these early indications need to be reviewed in the light of later (Phase Two) development if their suitability, their dependence on the vital expectation:satisfaction relationship is to be meaningfully investigated.

But this is to state the obvious in a longitudinal study. The main practical purpose that my two Phase One chapters have fulfilled is to provide sufficient systematised data for the follow-up hypothesis-testing needs of Phase Two.
CHAPTER EIGHT

PHASE TWO AND BEYOND
1. **Introduction**

The purposes of this chapter are the following:

1. to summarise the Phase Two data collection design
2. to analyse and interpret the TL data
3. to analyse and interpret cognitive/affective and social data
4. to compare Phase One and Phase Two participant profiles, drawing some conclusions on the hypotheses they inform
5. to re-assess the status and implications of my three general Research Hypotheses.
6. to summarise the educational implications of my main thesis.

This final chapter has to perform two distinct hypothesis-testing functions demanded by the overall design (see Chapter One, Section 5). It has to test 'new' hypotheses specific to Phase Two of my study: it also has to re-assess Phase One hypotheses in the light of Phase Two findings.

2. **Phase Two Data Collection: an overview**

Figure 8.1 below represents in more explanatory form the chronological description of data collection events shown in Figure 5.1 above. The complex relationships illustrated suggest that this chapter should follow a thematic rather than a chronological pattern. If data were analysed for each collection occasion in turn, too many of the lines of inquiry would have to be left open pending the findings of subsequent occasions. I shall thus structure the chapter according to the earlier TL, cognitive/affective and social categories, focusing on the new academic dimension whenever it is appropriate.
Figure 8.1: Sources and types of Phase Two data and their connections with final Participant Profiles
Figure 8.1 reminds us of several features of the Phase Two design.

1. There are a variety of types of data, 'hard' (see the circular frames), 'soft' (square-framed) and combined quantitative and qualitative data (septagonal).

2. These feed into the Phase Two Profiles by design (see the solid connecting lines) or less directly or intentionally (---). 

3. The expected number and complexity of the connections between the data elicitation devices and the multi-dimensional profiles are emphasised by the web of inter-connecting lines.

4. The academic context dimension now comes into its own as contact with participants' receiving institutions is established.

3. Phase Two Target Language Data

The most detailed TL data for Phase Two are from the part administration of my test battery at the reunion of participants in December 1980 and the complete re-administration in May 1981. Since the fairly comprehensive procedures carried out in Chapter Six have already established the tests' status with regard to reliability and various forms of statistical validity, the follow-up tests will not be re-submitted to such procedures except where a completely new form of a test is introduced. Whenever possible, however, questions raised by the Phase One use and analyses of the tests will be pursued in the light of new criterion-validating data. Throughout this section, in fact, test and other TL information are juxtaposed as I check the prognostic validity of the tests and the validity of the criteria to which real life tends to relate measures of TL proficiency and progress. As Figure 8.1 indicates, real-life sources include tutor and participant interviews, group discussions and receiving institution reports.
The most difficult technical problem faced in Phase Two TL test use was the development of an alternate form of Test 2 (the competence-oriented test). The original version had already been used twice in identical form, so a third and fourth administration risked too much contamination from practice effect. Since I had no opportunity to pilot the new version, I could not be sure of equivalence even with careful construction of items apparently closely paralleling the originals. I thus adopted the following post-facto validating procedures:

1. Test 2A was used at both reunions, scored and given the same item analyses as the original version.

2. Items showing deviant performance in item difficulty or discrimination terms were noted and included in Test 2 along with the original parallel items when the test was used with the ELTI/ODA 1981 group (see further references below).

3. Mean differences in level of difficulty between the two sets of items were then computed and the Test 2A scores of my 1980 participants on the items concerned adjusted, using the following formula for each of the three parts (structure, spoken use, written use) separately:

\[
\text{Adjusted sub-test score} = \frac{\text{Original score} \times \text{Mean 'new' item difficulty} + \text{Total No. of items} \times \text{Mean 'old' item difficulty}}{\text{No. of deviant items}}
\]

eg Participant A on Test 2A:

\[
\text{Part Al Score} = 12 + \left( \frac{8}{15} \times \frac{15.5}{12.5} \right)
\]

\[
= 12.66
\]

Interestingly enough, the adjustments required on the first two sub-tests were relatively minor, whereas attempts to produce equivalent written use items had been less successful. The task of finding
parallel excerpts of authentic discourse with equivalent cohesion and coherence features is a difficult one.

A different cloze passage was also used as Part B of the test. Again, it required a check on difficulty level if the new test was to have the equivalence required for longitudinal comparisons. In fact, the two cloze texts performed very similarly when used together (with my own group in May 1981 and the 1981 ELTI group in September 1981), probably because the second text was the authentic continuation of the first. However, a marginal adjustment of scores based on a percentage error analysis was made (ie \( \text{Cloze B score} = \text{Cloze A score} \times 1.01 \)).

Figure 8.2 summarises performance on Test 2 by the 18 participants who were in UK for all four administrations.

<table>
<thead>
<tr>
<th>TEST 2</th>
<th>7/80</th>
<th>8/80</th>
<th>12/80</th>
<th>5/81</th>
<th>Significance of gain, 7/80 to 5/81</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parts A x SD</td>
<td>x SD</td>
<td>x SD</td>
<td>x SD</td>
<td>x SD</td>
<td>p ≤</td>
</tr>
<tr>
<td>and B</td>
<td>33.7 8.3</td>
<td>39.8 8.4</td>
<td>43.8 5.1</td>
<td>46.0 6.3</td>
<td></td>
</tr>
<tr>
<td>Average %</td>
<td>56.2</td>
<td>66.3</td>
<td>73</td>
<td>76.7</td>
<td>.01</td>
</tr>
</tbody>
</table>

Figure 8.2: Mean scores, standard deviations, percentages and gain significance for four administrations of Test 2

The following are key points about Test 2 based on this and other relevant information.

1. The gain in the aspects of participants' competence tapped by Test 2 is educationally as well as statistically significant. What started life as an initial placement and diagnostic test for students already classified as needing remedial English (see Chapter Five above) and scored the whole group at an average of 51.4% is, by May the following year, stretched to near its limits in terms of level of
difficulty and discriminating power. (Percentage error in Part A, sub-tests 1 and 2, for example, was as low as 17.5 and 20 respectively with correspondingly weakish ID figures at 29.5 and 22.8). Yet the test has traced a consistent pattern of improvement and, even at the end (May internal consistency index by K-R 21 formula = .75), seems to be measuring and discriminating some form of underlying competence that relates to performance criteria, in a way examined more fully in the correlational analyses below. Suffice it to note at this stage that the test makes individual distinctions in line with other evidence. Among the big gainers is Participant 0, who we already know (see Chapters Six and Seven) has made outstanding TL progress in Phase One, and is now confirmed as doing so by reports from her tutor and the chance that she will stay to study further. It also picks out L (see above) as getting much closer to his genuine ESL competence (up 30% to 82%), though with question marks, still, over the most study-related section of the Test (discourse cohesion and coherence) where he actually regresses between 12/80 and 5/81 to a 5/10 score. The lowest gain of all is recorded by A, whose competence level started off high (again genuinely ESL) but whose Test 2 scores have reverted almost to their original 77% after reaching 85% at the end of the ELTI course. Has he reached the stage (and age?) where his TL competence is adequate for his C2 studies and social life and thus not felt worth the extra effort to overcome the law of diminishing returns?

2. Rank order correlations on all four administrations have interesting implications for discussion (eg Morrison 1974) on the varying significance of starting level and later TL competence. Inter-Test 2 correlation coefficients are given in Figure 8.3 here:
Given the disparity of intervening events (and the small sample size) it is surprising that these figures are as high and close as they are. The test seems to indicate that most participants gain steadily yet without much variation in their TL level relative to others, in spite of the differences in time between tests, or learning/living circumstances. Taken with the figures in 8.2, the evidence is that the pre-sessional training did give a significant boost to TL level, did sort things out so that the post-test is a better predictor than the pre-test, but that neither pre-sessional learning nor subsequent TL acquisition overrides, in most cases, the strong influence of TL level on arrival. (If the group of 18 were to be 'placed' on the results of the May 81 Test 2, as they were on the results of the whole pre-test battery in July 1980, 15 out of the 18 would be in the same ELTI course group as originally.)

The four performance tests (Tests 1, 3, 4 and 5) stay close to their original construct in Phase Two. Their communicative validity will depend on their relationships with real-life criteria, given the satisfactory level of reliability established in Chapter Six and learnt from at their re-administrations. The performance tasks, then, are allowed to change in tune with the time and occasion, criteria for their evaluation are not.
Test 1 (Lecture note-taking and dictation) achieved, it will be remembered, a high level of inter-rater reliability in Phase One but revealed very little gain in performance. Perhaps this reflected the rather limited amount of realistic lecture practice on the ELTI course (see Chapter Six, Section 4.1 above). My visits to receiving institutions showed an average of 15 hours per week of lectures across the group as a whole, so by the time their lecture and note-taking abilities were tested again, such practice was no longer lacking.

The May 1981 Test (see Appendix 1.1) was devised and administered by an ELTI teacher who had been present at and scored Pre- and Post-Tests 1. His lecture was on the subject of overseas students in four different host countries; his scoring scheme followed the previous pattern and was validated by three native-speaker informants. A dictation was included, as before. Figure 8.4 summarises Test 1 performances on the lecture and note-taking, all scores based on three separate ratings. (Inter-rater reliability coefficients were .81, .93 and .87; p<.001). N=the 19 students who took the test in May.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \bar{X} )</td>
<td>SD</td>
<td>( \bar{X} )</td>
<td>SD</td>
</tr>
<tr>
<td>Lecture</td>
<td>50.4</td>
<td>15.4</td>
<td>53.9</td>
<td>12.3</td>
</tr>
<tr>
<td>Note-taking</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.4: Mean percentage scores, standard deviations and significance of gain on Lecture and Note-Taking

Key features here are as follows:

1. The gain is statistically significant and, as agreed by all raters, educationally so. The amount, accuracy and organisation of information processed were probably the clearest evidence of real
learning seen in any of the final tests. In fact, Participants E, O and X were given scores very near to those achieved by the native-speaker teachers who took the test. Tutor ratings using the English Ability Rating sheet (Moller forthcoming, see Appendix 3) as completed during my receiving institute interviews confirm the top third performers on Test 1 as at least 'adequate' in listening comprehension, though it is noticeable that tutors rate this covert ability, as they do reading comprehension, generally higher than the more overt skills of speaking and writing. All three participants rated as less than adequate in listening comprehension by their tutors, do poorly in the test, including L and A, whose overall level of TL competence is high (see above). However, if we look for a clearcut relationship between the test scores and the amount of lecturing experienced by the participants or their self-assessment of their adequacy in understanding and taking notes, it is not to be found. Data from the follow-up interview forms (see Appendix 3) shows participants who have attended a lot of lectures performing poorly on the test (eg R, N) as well as some with a relatively light lecture load doing well (eg E, B or F). Similarly O, X and B are among the few participants who assess their own listening comprehension as inadequate, yet come top, third and fourth respectively in the test! (The three weakest performers in the test all, incidentally, rate their listening level as adequate.) Among the possible explanations of the first phenomenon is the obvious fact that some lecture-mode activities are more demanding than others (R's 28 hours of 'lectures' are teacher-directed, supportive and oriented towards his laboratory technician certificate examinations; E attends only occasional visiting lecturer talks, but has the regular pressure of ward round reports, case account discussions and doctor-
patient interviews to sharpen his listening and note-taking skills). And the second phenomenon, regarding self-assessments, we have already encountered (see Chapter Seven, Section 4 above). No fewer than 6 of the 8 participants giving themselves a negative rating on listening comprehension are in the field-independent sub-group, further evidence that this cognitive style tends towards a self-critical view of ability yet towards effective use of the analytic skills under fire. All of the top seven performers on Test 1 have successfully completed the first year of their training, two (O and F) with the chance of further UK studies, one (z) in spite of considerable TL writing difficulties. Of the lower seven, however, two have failed their end-of-year examinations and one has been asked to attend extra remedial English classes.

2. But although the May/81 Test 1 emerges well as a predictor of its real-life criteria, there are some anomalies still in the test taken over its three administrations. Why does Pre-Test 1 predict the final version at a Spearman rho of .7 for the 19 participants yet correlate at only .34 with Post Test 1, which itself predicts performance on the May/81 version at only .4? Were the skills involved so little practised on the ELTI course, so little acquired through the six weeks of C2 contact, that the two tests performed merely as two separate measurements of Performance 2? It was not that the post-test was significantly harder; there was, after all, a slight overall gain. But why should X place second then twelfth, Y twelfth then first (of the 19) in Phase One, only for the former to emerge third in May, the latter tenth, both after an academic year with a heavy lecture component at MSc level? I suspect we had a problem of ecological validity. Tying a lecture test in with the orientation, then the round-up talk
on a course has a lot to recommend it, communicatively as well as practically. But there are significant differences in social relationships and attitudinal tone between the beginning and the end of a friendly pre-sessional programme. These pedagogically and socially welcome differences may not be so helpful from a performance testing point of view. Further research (appropriately signalled now I have reached a late stage in my own) might investigate the advantages of a more formally topic-oriented lecture such as the May/81 version as more likely to elicit EAP-relevant performance. The format and means of assessment have worked well; topic selection and psycho-social setting features need checking.

The dictation, apparently emerging on Phase One evidence (see Chapter Six, Section 5) as a competence- rather than performance-oriented measure, continues to lose its discriminating power during Phase Two. By May 1981 the group averages 84.4% (SD 7.8), up from the initial 63.8% (SD 19.8). The influence of the shock factor evident on the Pre-Test has gone and the ESL advantage has receded even further. The pre-test dictation correlates at a low .24 with the final test; the post-test at a significantly higher .62. The most vivid illustration of the absence of a relationship between dictation and note-taking performance is my lowest rank-order correlation yet, a mere .08 between the two on the final test. There is certainly a role for dictation in any TL test battery (see the correlation matrix at the end of this section). But it certainly performs as a linguistic competence measure and as such does not belong, in assessment terms, in a performance test. This is why I have separated the two parts of Test 1 (note-taking, dictation) in my statistical analyses.
Before we look at the performance of the problematical Test 3 (Reading and Report Writing, see Chapter Six, Sections 4 and 5 above), it will be helpful to check what is known about participants' relevant study activities as revealed by receiving institution interview, tutor report and group discussion sources. At the times they were interviewed, the group as a whole were averaging around 20 hours a week of academic reading, though as the standard deviation figures from the coded interview forms show (SD = 14.5), there was considerable variation, with participants on non-academic training (eg K, W, E or R) doing much less reading (x̄ = 9 hours a week) than those on Diploma or Master's courses (x̄ = 25). An average of around 11 hours a week were being spent writing, mainly in essay mode but, in at least six cases, in a mode more akin to 'writing up', for example, laboratory reports (as with R or N), clinical reports (E) or literature survey summaries (U). There was certainly an interesting distinction between those whose reading and writing were in tandem, with regular essays being written on the basis of concurrent reading (eg A, B, F or X), those who were reading intensively, note-making constantly, but with their major writing assignments delayed until the end-of-year exams (eg G, Q or Y), and those whose main course emphasis was on writing up, rather than reading (eg N or R, for whom lectures and lab. work were the priorities). Both tutors and participants distinguished between adequacy in reading and adequacy in writing. On the Moller scale, tutors considered the students rather more than adequate on the former set of skills (though, as indicated above, the less overt nature of reading skills make them difficult to evaluate), but only just adequate on average for the latter. The participants themselves, who had more than one opportunity in the semi-structured interview to discuss and assess their own study
skills, also gave an average 'adequate' rating for reading but less than that for writing (14 out of the 25 interviewed actually rating their writing as inadequate). However, adequacy is, literally, what you make it and no fewer than 19 of the 21 participants with a significant reading load, including those who regarded their ability as adequate, identified reading speed as at least a 'slight' problem and, for the majority, quite a serious or a serious one. Now, since the speed problem can be seen as the effect of many 'micro-skill' causes, the participants may well be identifying reading as a source of real difficulty, often only 'overcome' by the 30 or 40 hours a week spent trying to make it seem 'adequate'. When the sharp (Ql+) Participant I mentions at his interview that reading in English is twice as slow as in L₁ he is making a comment of extreme but easily under-rated significance. A underlines the point nicely at his December 1980 group discussion, even though his is a genuinely ESL perspective:

".... simply knowing to talk and how to behave is not enough .... The local native students who are studying with us .. you see .... it's quite easy for them to go through a book once and then have a clear idea of it, but it is difficult for us because it requires one already to understand the books and all these things, to criticise to make our own comments and especially in English because .... as it is not our mother tongue it is now difficult, it will be difficult, it will ever remain difficult even if we obtain first (class) degrees!"

And in the participants' more explicitly negative assessment of their writing abilities, speed is again the most commonly mentioned and biggest problem, though this time most often in association with 'expression', that is communicating their own ideas. Given that the latter ability is easily the most important requirement in written work as identified by tutors when asked about their evaluation criteria for written work (whether or not they felt, as was sometimes the case for
Master's courses, that overseas students' written performance as a whole should equal native speakers'), we again have the makings of a serious problem. What does Test 3 tell us about it?

In criterial terms, the summary of real events above would seem to support the construct of Test 3 (see Chapter Six, Section 3.3 above). It requires students to process a variety of kinds of written information, summarise, evaluate and relate it to their own worlds in various ways with a varyingly explicit degree of direction. All this under quite considerable time pressure. But the test's somewhat inconsistent correlations with other tests in the battery caused some concern (see Chapter Six, Section 5 above) as did its failure to show significant gain by participants over the pre-sessional course. Figure 8.5 summarises performances on Test 3 after its third administration with 18 participants in May 1981. A form of the original tests using Part A of Post-Test 3 and Parts B and C of the pre-test version was used. Scoring criteria and raters were the same as for pre- and post-test versions.

<table>
<thead>
<tr>
<th>TEST 3:</th>
<th>7/80</th>
<th>8/80</th>
<th>5/81</th>
</tr>
</thead>
<tbody>
<tr>
<td>READING AND REPORT-WRITING</td>
<td>x SD</td>
<td>x SD</td>
<td>x SD</td>
</tr>
<tr>
<td>45.0 10.3</td>
<td>43.6  6.4</td>
<td>48.0  8.8</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.5: Mean % scores and standard deviations on three administrations of Test 3

The following are key points from these statistics and related analysis.

1. The lack of significant gain on the May 1981 test is very worrying; the Post-Test:Final Test difference is significant only at a weak p < .1, the Pre-Test:Final Test difference is statistically non-significant. Educationally, neither is at all satisfactory given the
intervening months of intensive EAP study. Is the test really flawed or have the participants really not improved much in reading and reporting? Probably both. But the flaw in the test is not caused by inter-rater unreliability. The Spearman correlation on the two ratings of the final Test 3 is .84 (p<.001).

2. The Pre-Test:Post-Test correlation based on the 18 participants is even lower than for the whole group (see Chapter Six, Figure 6.28) at .5. And since Pre-Test 3 correlates at .56 with the final test compared to .37 for the Post-Test, it is difficult to detect a pattern of improvement related to the pre-sessional course itself. (The 1981 group (N=19) using the same test showed a small 3.1% improvement during their course with pre- and post-test correlating at .61). Even more significant is that, as Figure 8.8 will show, Post-Test 3 has the lowest correlation of all the tests (or part-tests) with the final all-test rank-order, though, as before, (see Chapter Six, Figure 6.26) the final Test 3 correlates quite highly (at .63 and .65 respectively) with the reliable Tests 2 (competence) and Test 5 (Writing).

3. Of course, discriminant validation logic (see Chapter Six, Section 5 above) could suggest that Test 3's relatively low correlations with other tests mean that it is actually assessing something relatively discrete from what the other tests are measuring. What do the criterion data say to this? There certainly seems to be a relationship between the best performers on the reading and reporting test and courses where reading and essay writing are required in tandem. 8 of the best 9 performers on final Test 3 are on academic courses where they have already had to write regularly in connection with what they have read. Of the 9 weaker performers, 2 claimed to be doing little
reading, mainly writing up rather than writing essays (ie N and R), 3 (G,Q and Y) were not called upon to write until their final examinations (after the May reunion) and 2 more had been noticed as weak enough at writing for their tutors to request extra tuition during their courses. C and P, on the other hand, had heavy reading and writing loads on their courses but still performed fairly weakly on final Test 3.

4. These indications of Test 3's predictive validity still do not explain the unimpressive degree of improvement on a test that is partly a test of reading by students who have spent so much of their study time on that activity. Discussions with the teachers involved with Test 3, all of whom agree that final test performances were generally weak, suggest certain explanations. The time factor makes the reading and reporting tasks more of a speed than a power test so that some students lost heavily by not completing, or even starting, certain sections. The opportunity and their willingness to spend so much time on their reading at their receiving institutions may actually have hindered reading speed development. The type of information, particularly the statistical nature of Task A, may have been ecologically invalid for some participants. The influence of task instructions on performance (see J Carroll, Chapter Four, Section 6 above) may have been greater than in real-life academic tasks: certainly a high number of students misread the instructions for Task C, failing to appreciate that the contradictory quotations could not all come from a single speaker but represented a cross section of views. And finally the 'common core' subject matter of the information to be processed may have failed significantly to trigger the strategies of authentic communication. The new British Council/Cambridge University test
battery has 'modular' (i.e., subject area specific) study skills and writing tests on which the 1981 ELTI group made a more significant gain than they did on my Test 3. The difference made by familiarity of subject matter needs further research, however. The attempt to classify my group's courses in Chapter Five, Section 4 above, shows how difficult it is to specify which domains of discourse are actually typical of which courses. But none of these points invalidate the clearest message of all from Test 3 and related data, namely that the difficulties of reading and reporting are considerable for this group of overseas students on their C2 training programmes.

Test 4 (oral interview) was re-administered at both follow-up reunions, with interviewer/raters who had experience from the pre- and post-test versions and with assessment criteria unchanged. Performances were measured using mean scores from the combinations of 'live' and recorded, joint and solo ratings as before. The interviews now, however, were less structured than during Phase One with interviewers given general areas of discussion to focus on with more freedom in the direction and sequencing of questions (see Appendix 1.4). Also, participant-to-participant topic cards were no longer used. Figure 8.6 summarises results using mean scores of all ratings (six ratings each for the two follow-up tests) for the 19 participants, most of whom performed on all four occasions.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( \bar{x} )</td>
<td>( \bar{x} )</td>
<td>( \bar{x} )</td>
<td>( \bar{x} )</td>
<td>( p \leq .001 )</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>7/80</td>
<td>56.5 13.8</td>
<td>65.3 11.7</td>
<td>68.5 7.8</td>
<td>71  8.1</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.6: \( \bar{x} \) means, deviations and gain significance for Test 4
Test 4 and related criterion data reveal the following key points.

1. There is a statistically and educationally significant gain in participants' oral communicative performance according to the Test 4 results, which are unlikely to be unreliable given the mean correlation coefficient of .77 across all ratings on both December and May administrations. The general pattern of improvement when all four tests are compared indicates the steepest gain during the pre-sessional period ($\bar{x} = 7.1\%$ for the 19) followed by a levelling off of the curve (4.5\% between 8/80 and 12/80 and 3.1\% between 12/80 and 5/81) during Phase Two, though much of the initial gain is probably accounted for by the overcoming of TL/C2 shock after Pre-Test 4 (see Chapter Six, Section 5 above).

2. The test continues to be a strong predictor of its later self; correlations between the various versions range from .76 (7/80:5/81) to .96 (7/80 to 8/80). The good and weaker oral communicators improve consistently in relation to each other. The gap, for example, between the talkative, socially bold Participants E or O and the shy, retiring B, C or G is unlikely to close.

3. Test 4 scores and tutor ratings on the Moller scale generally corroborate each other. 15 of the 19 participants are evaluated similarly on both measures though the cases where there is disagreement are also significant. B's tutor describes him as the 'quietest student' he has ever known, yet rates his spoken English as adequate for the demands of the course. My test is less generous since B's extreme reticence in a communication-demanding situation has to be penalised under criteria such as fluency, flexibility or even socio-cultural appropriacy. On the other hand a tutor rating on spoken adequacy
explicitly based on a personal view of one participant's accent is less generous than the test, where accent is but one of 8 criteria.

4. The participants' own analysis of their listening and speaking abilities is revealing. The receiving institution interview data show them rating their listening comprehension as, on average, slightly less than adequate, their speaking as rather further below the adequate level (on a par with their self-assessments of writing). Easily the most significant problems with listening are speed (again implying a variety of fundamental difficulties) and accent, specifically the accents of course colleagues rather than tutors. Equally easily the biggest obstacle with speaking is the fear of making mistakes which, in the interview responses, co-varies strongly with lack of confidence, shyness and the problem of 'entering' interactions. The speed factor in speaking is less than half as prominent in the participants' self-assessments as it is in listening. Very few mention linguistic features such as vocabulary or grammar as problems without mentioning (usually as a bigger problem) the fear of error factor.

The formats and evaluation criteria of Test 4 have proved robust. Further evidence on validity appears in the inter-test correlational discussion below.

Test 5 (free writing) was also administered four times with tasks modified to suit the occasions, scoring procedures following the same pattern throughout. Figure 8.6 summarises the results for the 19 ever-present participants.
The following key points emerge from the test performances and my other data.

1. The improvement between 7/80 and 5/81 is reasonable here though not convincing enough to contradict the concerns about writing as a problem discussed in connection with Test 3 above. The slight regression between the third and fourth administrations of the test is probably not caused by rater unreliability. Inter-rater correlations ranged from .64 to .82, close to those for Post-Test 2 when the same three raters were used, and only a little lower than for December with one different rater. Whichever combination of raters had been used, the December/80 to May/81 difference would probably have been insignificant. There are two inter-related explanations for this, again to do both with the test itself and with real life. Rater 3 (who also scored the Post-Test) points out that the final test instructions infringe my own criteria for communicative validity (see Chapters Four and Six) by inviting participants to 'write ... advice to a friend in your country', an activity they would be unlikely to carry out in TL, which was not the case in the three previous tasks where they were communicating directly with their teacher/evaluators. And this flaw did, in fact, seem to cause extra uncertainty in essay content as well as tone. This is a sound point and could easily be remedied in similar future tests. Then there is evidence in the final essays (eg from H and P) of the
acquisition of new lexis or set phraseology that it is beyond the participant's linguistic competence to integrate coherently into his written discourse. The result of this is a reduced communicative effectiveness in TL use, where less complexity might have clarified the message. This latter interlanguage point underlines a process: product tension which my tests, as opposed to my experiments, do not allow for.

2. A useful criterion for final Test 5 results is the section on my tutor interview schedule where I ask tutors about their criteria for the evaluation of written work and their assessment of the individual participant against these criteria. Easily the most highly rated criterion is 'organisation of ideas' (re-interpreted by tutors on two occasions as 'argument and analysis' and 'conceptual clarity'). In second place is 'fluency and amount'. The more narrowly linguistic categories such as grammar and vocabulary figured much less prominently except at the two Departments where native-speaker performance criteria were claimed as required. The 9 participants described by their tutors as not measuring up to one or more of the criteria had a mean rank-order of 11.7 on final Test 5. The 8 considered up to or above the criteria averaged 5.0. To this extent the test was making the right kind of predictions, though the factor of different course/level demands can crucially affect prognostic validity. F's twelfth position out of the 19 might indicate an adequate TL level on a diploma course specifically designed for overseas students whereas X's sixth place (and consistently higher scores than F on all four administrations) is by no means conclusive evidence that his writing ability is adequate for an MSc at a university with strict regulations on overseas student TL level. There is an extra argument for the inclusion of the
more CALP-oriented Test 3 in the battery here to complement Test 5's BICS orientation. That participants like C, Q and R appear much happier writing friendly advice than they are reading and reporting under pressure is not necessarily a good sign from an academic writing point of view. Test 5 continues to perform quite sensibly. The correlational analysis below will throw further light on its validity.

The correlation matrix in Figure 8.8 looks at the prognostic validity of the August 1980 post-test battery for performance on the May 1981 readministration. It parallels Figure 6.28 above where Phase One pre-: post test correlations are analysed. The relationships here are of special significance since they make connections between Profile 1 and Profile 2 (see Figure 5.1 above).

<table>
<thead>
<tr>
<th>MAY 1981 TESTS</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>1</th>
<th>2A</th>
<th>2</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(notes)</td>
<td>(dict)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A Post-Test 1</td>
<td>.63</td>
<td>.68</td>
<td>.45*</td>
<td>.48*</td>
<td>.67</td>
<td>-</td>
<td>.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U Post-Test 2</td>
<td>.51*</td>
<td>.70</td>
<td>.71</td>
<td>.29*</td>
<td>.62</td>
<td>.47*</td>
<td>.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>G Post-Test 3</td>
<td>.40*</td>
<td>.62</td>
<td>.37</td>
<td><strong>.39</strong></td>
<td>.41*</td>
<td>.42*</td>
<td>.43*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U Post-Test 4</td>
<td>.17**</td>
<td>.32**</td>
<td>.44*</td>
<td>.83</td>
<td>.70</td>
<td>.25**</td>
<td>.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S Post-Test 5</td>
<td>.40*</td>
<td>.48*</td>
<td>.64</td>
<td><strong>.35</strong></td>
<td>.75</td>
<td>.55*</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>T Post-Test 1</td>
<td>-</td>
<td>.60*</td>
<td>.31**</td>
<td>.28**</td>
<td>.57*</td>
<td>.41*</td>
<td>.61*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(notes) 80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T Post-Test 2</td>
<td>-</td>
<td>.63</td>
<td>.59*</td>
<td>.44*</td>
<td>.72</td>
<td>.23**</td>
<td>.62</td>
<td>.75</td>
<td></td>
</tr>
<tr>
<td>(Dictation)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E (Part A)</td>
<td>.56*</td>
<td>-</td>
<td>.57*</td>
<td>.38**</td>
<td>.64</td>
<td>.48*</td>
<td>.74</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>S Post-Test 2</td>
<td>.47*</td>
<td>-</td>
<td>.60*</td>
<td>.11**</td>
<td>.42*</td>
<td>.49*</td>
<td>.56*</td>
<td>.57</td>
<td></td>
</tr>
<tr>
<td>T (Cloze)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S All Post-Tests</td>
<td>.55*</td>
<td>.79</td>
<td>.73</td>
<td>.5*</td>
<td>.84</td>
<td>.40*</td>
<td>.89</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Figure 8.8: Correlation Matrix for prognostic validity of Post-Tests (8/80) for Final Test performance (5/81)**

The following are key points from the matrix with particular reference to the present concern of how the end of pre-sessional course TL profile (part of Profile 1) predicts final test performance (part of Profile 2).
1. The correlations in general are lower than on the pre-post-test correlation matrix (Figure 6.28). A single asterisk now denotes a significance level of between $p < .1$ and .02, a double asterisk the absence of statistical significance. Some of the difference in the indices, in particular their general tendency to be lower will, unfortunately, be explained by the here mainly irrelevant fact of the reduction in the size of the group by May 1981. Some of the variance, however, will be expected, given the differences in intervening TL events for different participants.

2. There are now tests besides Test 3 that predict results on other final tests better than on their own later versions but, significantly, final Test 3 itself is usually involved. The fact that Post-Test 2 and the post-test cloze correlate marginally higher with final Test 3 could be interesting new information about my problem test. Perhaps it was so formalised in task structure and evaluation criteria that it behaved more like a competence measure than the performance test it set out to be. Whether this is necessarily a bad thing becomes more problematic in the light of the matrix in 8.8 since it reveals that competence-oriented tests (eg Test 2, dictation, Test 2 (Part A)) emerge as the three best predictors of overall final test performance (though with the cloze part of Test 2 actually reducing the whole Test's prognostic power to a level below that of the structure, spoken use and written use component alone).

The emerging hypothesis at the end of Chapter Six, that a reasonable Test 2 score (ie above the Post-Test 2 mean of 63.8%) was reason for confidence about participants' EAP future is partially corroborated by final test evidence, though with the suggestion that it could be
further narrowed, to be based on Post-Test 2, Part A only. What is the real-life picture of participants who performed above and below the mean (i.e. 67%) on this sub-test? It can be summarised briefly here on current evidence (as of December 1981), the summaries to provide a reference point for the further analysis and interpretation of data in Sections 3, 4 and 5 below. Of the 14 participants above the Post-Test 2A mean:

A has passed his Diploma in Education and returned home.
B has passed in his MA coursework and exams, written the dissertation this qualifies him for and returned home.
C has passed her Diploma, was supported by her Department for continuation to a Master's, could not, however, be funded further, so returned home.
D left his MSc course in March on the grounds that it lacked specific relevance (see Chapter 7, Section 4 above).
E has moved on from his intern duties in a general hospital to more specialist studies in another, with enthusiastic support from his consultant.
F has gained her post-graduate diploma and been given an award for an extension to Master's level.
G has passed his MSc by examination and dissertation and may return for doctoral research.
I has encouraging reports on his first year progress with predictions that he will gain his 2-year MSc.
K completed his 5-month course with ease and returned home in January 1981.
M completed her 5-month attachment satisfactorily and also left UK in January.
O has excellent reports on her MA progress and is currently writing her dissertation, with encouragement from her receiving institution to study further.

P is in the same successful position as B.

R has completed a satisfactory first year of his 2-year certificate course.

Y has passed his MSc exams, submitted his dissertation and returned home.

None of these participants have failed academically because of their TL level. To that extent the post-test version of the competence-oriented Test 2, Part A seems to be a powerful predictor, a key component of Profile 1. But what of the 13 participants below the mean?

H has reasonable interim academic reports on his 2-year MSc course but is having to take extra TL tuition especially in writing (he placed last on final Test 5, second lowest on final Test 3).

J did not complete her postgraduate certificate because partly unforeseen academic course demands from C1 meant she had to interrupt. Receiving institution reports up till then had been promising.

L (like J, on the mean in Post-Test 2A) has passed his PGD and returned home.

N has passed his MSc course and returned home.

Q has failed his LLM but will re-sit next year (1982).

S (on the mean) has passed his PGD, been recommended for MSc studies but returned home because of lack of funding.

T has failed her PGD and returned home.
U successfully completed an extended post-Doctoral attachment and has returned home.

V did not complete her 5-month training course, returning home early because of close family illness.

W completed his 5-month training course in spite of considerable TL problems and returned home in January 1981.

X was finally allowed on to the preparatory year for his MSc and obtained 'the minimum requirement' (tutor, 7/81) on it. He has been asked to do extra practice in written English. (Final Test 5 rank 6th (of 19), final Test 3 rank 8th of 18).

Z returned home before the beginning of his tailor-made specialist course for personal family reasons.

Z has persevered with his MSc course and practical attachment in spite of considerable TL problems (again mainly in writing (cf final Test 5, rank 16, Test 3 rank 15) but is now returning home with a good chance of having passed his MSc.

7 of these 13 participants have experienced considerable difficulties or study failure explicitly (ie from receiving institution comment) related to TL level.

Post-Test 2A (after the crucial sorting out function of the pre-sessional course) certainly has powerful prognostic validity on this evidence. But it is not a sufficient profiling instrument in itself as anomalous cases like L (not by Post-Test 2A stage revealing his actual SL competence), N (who passes at MSc level in spite of his consistently weak test scores) and cases like D or Z (where any exclusively TL profile is clearly inadequate). We need the cognitive/affective/social/academic dimensions in Section 4 below both for fuller
profiles and to construct a broader, more sensitive model of 'success' and 'failure'.

3. Meanwhile, a brief look back at the 8.8 matrix for further correlational insights. The oral interaction Test (4) continues to correlate relatively weakly with other tests, again suggesting its BICS vs CALP focus, though now it relates more highly with the free writing final test; has the advice to a friend format made that test more social than academic? (Note the discriminant validation evidence for this in the low correlation between the relatively less social Post-Test 5 and the informal final oral interview.) The low correlation between post-test note-taking and its final test equivalent is probably explained by the high level of change in ability in this skill noted above. Cloze is confirmed as more closely related to the competence than the performance tests, given that Test 3 has itself somehow moved back towards competence status. Dictation, in spite of its lack of discriminatory power now participants are so much better at it, is overall a better predictor than before, though not of the essentially performance-based tasks of note-taking or oral interaction. Nevertheless it is the whole post-battery that remains overall the best predictor of specific-test and all-test performance on the final battery (with a mean correlation coefficient of .67 compared with .61 and .60 for Post-Test 1 and Post-Test 2A respectively, the next best predictors).

The general TL picture of the 19, July-to-May participants is summarised in Figure 8.9 below where the pattern of improvement over the three complete batteries is shown.
We now know some of the individual circumstances behind this group TL profile. In Section 4 both group and individual profiles will be refined in the light of Phase Two cognitive/affective, academic and social factors.

4. Phase Two Cognitive/Affective, Academic and Social Data

The amount and detail of these data would make the task of analysis and interpretation impossible (see Malinowski, Chapter Three, Section 3 above) if the research design had not kept a quantitative control on qualitative data. The analysis here should demonstrate how this worked.

4.1 Personality Factors and Cognitive Style

The two receiving institution interview forms asked for participant and tutor ratings of certain personality traits, including categories specifically related to Q₁ and FI, namely conservative/experimenting (on participant interview forms), curiosity (on the tutor forms) and independence (on both). High Q₁ participants like D, E, F, I, J and U rated themselves as above average on experimenting, below on conservative and were considered by their tutors as above average on curiosity (seen as an important positive academic factor). G and O raised an interesting question on whether Cattell's Q₁ trait is actually dichotomous by rating themselves high on both conservative and experimenting. Given that neither of these two participants make a habit
of illogical statements, both often displaying a particularly rigorous analytic tendency (see for example Experiment 3 in Chapter Seven above), one could infer that, in their cases at least, openness to new ideas does not mean closedness to older ones. None of the 10 participants considered average or below on curiosity by their tutors are members of the $Q_{1+}$ group. In the official tutor reports to the sponsoring agency, most of which, incidentally, refer in only fairly general terms to specialist area performance and TL level, there is some corroborating $Q_{1+}$ related comment. Eg of Participant E:

"quickly got the feel of our working methods and attitudes"
O:
"intellectually adventurous .... tendency towards theoretical analysis .... evaluating her explorations faster than most"
U:
"making use of the relevant information"
J:
"quick to grasp ideas .... tolerant"

These compared with:

"would benefit from a more questioning and positive attitude"
"too 'close' an interpretation of lecture notes"

and a reference to a reluctance to infer, to depart from the objective normal science paradigm made of two more non $Q_{1+}$ students (neither on 'science' (vs. 'arts') courses). There are some $Q_{1+}$ connections to be made in Phase Two. More will be mentioned in connection with the findings on motivation, main course evaluation and $C_2$ contact below.

Where the interview asks participants about perceived changes in various personality traits, the overwhelming response (ie in 83% of the responses on 11 traits by 23 participants) is that there has been no change. Certainly no trait shows a statistically significant difference though there is a slight upward tendency on independence and 'worry', a slight downward one on participants' perceptions of their
sociability and trust (see below for possible explanations).

Participant and tutor responses on independence (the latter both study-related and social) also show some agreement with my Phase One test observations. Of the FI+ group, (see Chapter Seven, Section 3 above) E, K, M, O and Y are confirmed as independent by both self- and tutor-ratings, C and I by their tutors but not themselves, though these latter two may well be displaying the FI+ tendency towards negative self-assessment noted in Chapter Seven, Section 4. Participant U is the only field-dependent participant with an unambiguously independent rating. No tutor at the interviews or in the official reports sees independence as anything but an advantage to overseas students unless, as in the case of D, it is associated with the inclination to reject everything in a course except what is perceived as his personal professional requirement from it. (But this tendency is not necessarily in line with the woods/trees balance implied by field independence; nor is D in the really high FI group).

There is nothing in the evidence so far to suggest that the Phase One indications that $Q_{1+}$ and FI+ participants are not, on balance, at an advantage in an overseas training context. Perhaps O, high in both, and, typically, writing almost the whole of her May/81 'advice to a friend' essay in the form of analytic questions, encapsulates an attitude and approach representative of these two key personality and cognitive style dimensions:

"I suppose you know what you want from England but on the other hand do you know how England is expecting from you? .... among other things you could start thinking in is a require of personal characteristics as self-confidence, independence, strong adaptability to any circumstances, optimism, strong volition for working, respect as an attitude for other people way of life ...."
then, demanding still further analysis and tolerance of change:

"Already defined your objectives, have you thought how your life in England would look like? Are you in disposition of accepting any change from this picture presently drawn in your mind?"

The two dimensions seem likely to figure in Profile 2 as they have in Profile 1 (Chapter Seven above). Participant S, low on Q1 but quite high on FI, suggests so in his May/81 essay:

"Most people think foreign life is only happiness without any suffering or bad things. However, you may meet many different things that you can hardly understand .... You will possibly have to be alone when you must decide very important decision. Actually this is involved in extreme independence, which is sometimes likely to be missed in our lives. But you will have to make very precise decisions whether this is effective on your life or not."

4.2 Motivation, Attitude, Training and Social Factors

The re-assessment of the orientation aspect of motivation was mainly through Motivation/Attitude Questionnaires 2 and 3 (see Chapter Seven, Section 4 above). The decision was made it will be remembered, to separate statements on orientational needs from those on TL relevance to those needs since, combined, they tended towards 'double binds' eg 'A good knowledge of English will help me make contact with more and varied people whose mother tongue is neither English nor the same as mine', being likely to force a respondee into a positive response even if he sees English as important for contact, though he is not particularly interested in such contacts, or if he wants such contacts, but does not see English as particularly important for them. In MAQ 2 and 3 the statement is split:

ie It is important to make contact with people whose mother tongue is neither English nor the same as mine.
(+ Response scale)
A good knowledge of English is important for this.
(+ Response scale)
For my needs for this purpose my knowledge of English is adequate/inadequate.
Perhaps this attempt to sort out orientational priorities from their TL connection is one reason why the relationship between the English language and general educational ambitions is perceived so differently on MAQ/2 and 3 compared with MAQ/1. In December 1980 and May 1981 the English for educational purpose (EEP) orientation rises from its Phase One last place to second place, behind EAP. In fact, on responses to the second part of the split, the importance of English to being 'really educated' is finally rated even more highly than it is 'to complete specialist studies'; this is an indication that by May, the participants have a slightly stronger focus on the future, more general importance of the language than they did with their earlier more exclusively main-course focus. Otherwise the 'integrative'/ 'instrumental' mix is much as before: 1) EAP, 2) EEP, 3) EOP, 4) English for C\textsubscript{2} contacts (these latter two narrowly exchanging rank-order places in May 81), 5) English for insights into the British way of life, 6) English for contact with other non-native speakers (one place higher in May), 7) English for social status and, in last place, 'to think and behave as the British do'.

On the self-assessment of TL adequacy for these various purposes, whether they are perceived as important or not, the 19 May participants see a significant improvement over December (p<.01), in line with the level of significance of difference between the December and May results on the oral Test 4, the most relevant test for at least 6 of the 8 areas of language use being considered here. Whatever else happens to a group like this during their C\textsubscript{2} training period, their orientation towards the TL clarifies, their perception of their adequacy in it strengthens.

Since these orientational data once again underline the primacy of course-specific TL needs, it is appropriate now to look at participant
evaluations of the specialist courses themselves, as a key factor in motivation and attitude. The student questionnaire completed during the receiving institution visits goes into the coverage, quality of content (theoretical and practical), level, up-to-dateness, relevance, flexibility, organisation, teaching expertise, participation potential, feedback/assessment procedures and recommendability of all the programmes concerned. This allows an aggregate evaluation of each participant's view of his course to be made, based on pre-coded quantitative values, checked through participants' qualitative comment and, where appropriate, documentary or tutor interview data. The first inference to be drawn from these indices is the high average rating given by the group to their programmes (n=24, \( \bar{x} = 66.8 \), SD=11.7 with a theoretical maximum of 84). These quotients may, of course, be expected to be influenced by individual tendencies towards scepticism, but in the case of most of the Q1+ group this tendency seems to be over-ridden by their tolerance of and openness to what is new. Of the 8 participants rating their courses below the mean, 3 are from a department where critical attitudes are encouraged as an inherent dimension of the courses (all 3 are, in fact, satisfied overall (see below), fall only slightly below the mean and pass). 4 of the others give significantly low ratings (1.4 to 2.3 SD's below the mean), all of them showing real dissatisfaction and mainly on the criterion of relevance (eg D see above), 3 of them being non-completers. These ratings, like those of the high main course raters, cut across TL proficiency levels; the group of 8, for example, range from 2nd to 24th on post-test battery rank-orders with a mean equal to the median. It is possible to relate these participant evaluations to the tutor questionnaire assessment of the motivation, attitude and specialist
subject expertise of students, using quantified tutors' responses on conscientiousness, interest, co-operativeness and motivation, then (as a second aggregate) on subject area knowledge, work habit organisation and academic potential. The first quotient is very high (\(\bar{x}=10.5/12\)), supporting the impression we already have of how hard the group as a whole works and in line with official receiving institution reports which are peppered with descriptions such as 'conscientious', 'hard-working', 'diligent' etc. The second quotient is lower and more dispersed (\(\bar{x}\) on maximum score of 12) = 6.2, SD=1.5), again cutting across TL proficiency but with the low-scoring group including all those who later failed their courses as well as non-completers like D and J. There is a moderate .4 rank-order correlation between participants' ratings of their courses and tutor-ratings of participants on those courses. This is in fact quite a strong statistical relationship when you consider the intervening variable of individual differences in positive rating propensities and the fact that participants are evaluating mainly the quality of a programme, tutors, mainly, success or failure on it. It should not go unnoticed that 7 out of the 9 Q1+ participants rated by their tutors are above the mean on main course ability, as are 6 of the 7 FI+ group.

These relationships lead logically on to a consideration of the expectation:satisfaction ratio signalled in Chapter One, Section 4 and Chapter Two, Section 3.3 above. Taking participants' rating of their overall satisfaction with their courses first of all, the anticipated relationships emerge with their analytic ratings of courses and their tutors' assessment of their abilities. Participants assessing their satisfaction as high or very high have a mean rank-order of 9.9
on their aggregated main course quotient and of 10.3 as rated by their
tutors. This compares with $\bar{x}=16.4$ and $14.8$ respectively for those des-
cribing their satisfaction as less than high. But using satisfaction
and expectation formulaically i.e. $\text{satisfaction rating}$, the picture
$\text{expectation rating}$
is less clearcut. Where the ratio is 1 or more, ($n=11$), the mean rank-
order on participant course evaluation is $\bar{x}=11.2$, on tutor assessment $\bar{x}=
9.0$. For those whose satisfaction:expectation ratio is less than 1
($n=13$), corresponding mean rank-orders are 12.7 and 13.9 respectively.
The inference here is that some students rate their programmes highly
but without perceiving their satisfaction:expectation ratio as positive.
H, Q, X and z almost certainly feel this way because of the TL problems
which leave their expectations less than fulfilled. E, O and Y (High Q1
and FI all) set themselves particularly high targets (O writes and talks
about her aim of 'a relation between this expectation and what I have
been able to do' of 100:100, Y is described by his tutor as a
perfectionist). The ratio is complex but revealing. It must certainly
inform Profile 2 (see Section 5 below), especially when it is noted
that the 11 participants with a positive ratio include 10 'successes',
the 13 with a negative ratio 3 'failures' and three 'doubtfuls'.

Section 2 of MAQ/2 and 3 parallels MAQ/1 in seeking responses to state-
ments on study habits and reactions which could constitute a further
measure of motivational intensity. On the whole these data are not
as informative as the freer descriptions of study activities given at
the interviews (see Section 3 above where TL test data and course
activities are investigated). There is no significant difference
between the positive and negative sub-groups on these responses in terms
of the expectation:satisfaction ratio, no significant correlations with
'success' or 'failure'. The reason is probably that some of the categories are non-discriminating as regards study habits and reactions. It may, for example, be a good or a bad sign for different individuals whether they 'like being asked questions' during their studies, as Phase One investigations have suggested (see Chapter Seven, Section 2 on the 'interactional tendencies' of different individuals) and is it helpful or unhelpful to worry about 'low marks on a test or exercise'? When all the participants were together on the same (ELTI) course, such items were more relevant and susceptible to validation. Once their paths diverged, this standardised study habit mode of inquiry seems less effective than the on-the-spot interviews in their institutional contexts. Further research is clearly needed here, probably emphasising longer-term participant-observation techniques in the places of study. Nevertheless the general message on motivational drive and perseverance is clear from all my data sources. These excerpts from the May general discussion sum it up:

X: I think that (H) is right .... we came here just to do our duty. We just do our best here whatever the result.
(........)
Y: Actually we were not expecting to play.
(........)
O: I think that depends very much on how you came prepared to deal with this kind of life. If you know that you are coming to study very hard, OK you can accept that and you do your best and you enjoy doing it ....
(........)
P: Yes I think (....) it is a matter of perspective, at least for me. From the beginning when we choose to come here I think it is supposed that we have to work hard (....) When we are finishing our course on Friday we only see our week-ends are the days when we have to work more harder.
(........)
z: ... I want to work hard and I'm happy.
(........)
B: I think that for the course, I enjoy it but you have to compensate by working hard.
B makes the next logical connection for us, from the essentially presage and academic community factors of motivational drive and perseverance to the broader C2 social context, which may be what he is compensating for. Again interview, discussion, essay and MAQ data will all inform this area of inquiry. The general impression of participants' social activities is that they are limited (x = 12 hpw) and more often than not undertaken in the company of other overseas students rather than with British friends. In fact the group as a whole has three times as many foreign as British friends (a ratio by no means adequately explained by the fact that 7 of the 25 were on overseas students only programmes), 13, not necessarily those with fewer friends of course, explicitly stating that 'they did not have enough' friends. The most common reasons given for this were pressure of work and differences in C2/C1 outlook. When the interview discussion moved to more generalised discussion of participants' main perceived problems, however, a more complex picture of the various factors and shares of responsibility in the situation emerges. Fear of failure on their training programmes was a significant problem for 13 of the 25, the TL for 18 (the latter being rated at the times of the interviews as slightly less than adequate), though more than slightly improved for social purposes (almost identical figures as those for their parallel ratings on English for their study purposes). No fewer than 20 of the participants suffered from homesickness to a degree that made it a genuine problem during their C2 stay (the interview questionnaire allowed for events or feelings that occurred but were not perceived as actual problems). It is interesting to note that it was not only family and friends at home that caused the feeling; at least half of the 20 also suffered from missing their jobs, climate, 'customs' or
food. The length of the C₂ period was a problem for 15 participants (all of them on programmes lasting a year or more). And 'loss of individuality' is an explicitly stated problem for 5 of the group, inevitably related to the TL question. Note this excerpt from a December group discussion:

O: But I think that it takes a long time because people understand you from your acts not from your talking .... because ....
Y: From your ....?
O: From your acts.
Y: Ah .... from your acts.
O: So you can .... you can ....
Native-speaker discussion leader: From your .... sorry I ....
O: From your acts, from your behaviour.
Discussion leader: Aah .... behaviour.
O: Yes because .... not because .... how you .... you explain them how you are .... this is very difficult for you to do .... so they need to know you and you need to know them from the behaviour that's all. And this takes a long time.

O's point is made and proved in the process of its being made, simultaneously. Since more practical matters such as the lack of formal arrangements for C₂ contact, accommodation or health do not figure significantly as problems for the group as a whole, and when they do for various individuals, they do so among other more serious problems, the profile of difficulties begins to look more as if it is the product of inherent factors of the overseas student's situation, rather than the responsibility of the host culture. But this interpretation may be a little too convenient given the evidence of lack of C₂ friends above and the worrying statistic that half of the group (though only one high FI member) interviewed admitted to significant problems of personal depression even when it was emphasised that this was to be taken as meaning something deeper and longer lasting than incidental unhappiness. It is worth looking at the participants' own characterisations of the British, in free talk at the interviews then
The balance of responsibility immediately changes. No fewer than 18 of the 23 participants who commented on British characteristics have reservations that could confirm the feeling of this researcher and the validating interviewer of three of the participants that the majority were unhappier than their own study, TL or practical problems alone needed to make them. But, as ever, it is a complicated picture. 10 of the participants characterise their hosts as kind, helpful, friendly or warm, nearly always against their expectations. Yet 6 of these qualify, almost contradict such descriptions: C2 contacts are friendly yet over-formal and rather too insistent on 'by appointment only' meetings; they are not reserved, but are more difficult to approach than other overseas students; they are friendly but often only as far as their strong feelings of class distinction allow; they are kindly but distant; helpful but insincere; tolerant but finally drive you to stop trying to make friends with them; they resist close friendships; they are helpful yet unfriendly; all these verdicts come from participants who do have positive things to say about the more helpful characteristics of C2. 7 others only mention negative contact-discouraging characteristics.

Perhaps we should look at the semantic differential scales in MAQ/2 and 3, 15 of which were re-investigated in exactly the same way as in MAQ/1, at the December and May reunions. The first fact that emerges here is that most characterisations of C1 and C2 are fairly stable over the 10-month period. The factors related in Chapter Six, Section 4, to third world:old world perceptions for example 'success' and
'security', continue to be rated significantly higher for C₂ than for C₁ (at p<.001 and <.01 respectively). The British also maintain their reputation (with this group) for politeness, significantly more highly rated than for C₁ at p<.01 in December, p<.05 in May. So do they on 'honesty', higher than C₁ at p<.01 at the first reunion, and up to p<.001 by the second. But although the perception of the host country's kindness remains quite high, more or less in line with C₁, the same cannot be said for friendliness, which is consistently and significantly rated lower than in C₁, at p<.01 on MAQ/2 then, slightly encouragingly, down to p<.1 on MAQ/3. The odd combination of kindness without complementary friendliness is, then, corroborated by the harder data of the motivation/attitude questionnaire.

There is a fair degree of consistency, too, in the C₂/C₁ perceptions of individual participants. D, G, N, Q, T, W and z continue to rate C₂ characteristics more favourably overall than C₁ (p<.01 (W's very significant distinction between the two) to p<.1) but I has dropped out of this high-distinguishing group by the time MAQ/2 and 3 are completed, and A has joined it. As before (see Chapter Seven, Section 4), T and z continue to rate C₁ characteristics rather negatively as now does G. F again judges in favour of C₁ (p<.1 on MAQ/2) in December, but by May 81 no longer discriminates overall in her evaluation of the two cultures. X, however, is still rather negative about C₂ on MAQ/2 and 3. V, the remaining negative discriminator, left the UK before the December reunion because of a serious family illness at home. There is one possible cross-cultural factor that emerges (as signalled at the end of Chapter Seven, Section 4 above). On MAQ/2 and
3, all three members of the South Asian sub-group rate $C_2$ significantly higher than $C_1$ overall but without their positive verdict being corroborated by other evidence. Q, for example, gives the most explicit mention of racial prejudice in his December 1980 essay:

"In the case of race, they have racial feeling and they don't like to mix up with different colour people coming from different part of the world. This is also one of the important things which is not suitable for British people."

And T seems one of the unhappiest of the group at her interview, with personal depression, homesickness and loneliness mentioned as significant problems not helped by her perceived helpful but unfriendly attitudes of $C_2$. A's reluctance to discuss personal matters at his interview (a reluctance shared by only one other participant of the 25 visited at their receiving institutions) was respected. Both this researcher and the validating interviewer, however, feel that it indicated $C_2$ feelings and experiences not wholly in line with his positive observations on MAQ/2 or 3. The group identified (also at the end of Chapter Seven, Section 4) as showing equally positive attitudes to both cultures on MAQ/1, that is B, C, E, H, K, O, P, S and Y (all bar P, it will be remembered, in the high field-independent half of the group) maintain this stance on the two later uses of the semantic differential section of the MA questionnaire, now joined by I, also FI+. As the summary of actual Phase Two success/failure events shows (see the discussion of the prognostic validity of Test 2, Part A in Section 3 above), 8 of the 9 have been successful on their training programmes, as has H, except for his specific TL writing problem. Participant Z, however, who gave positive $C_1$ and $C_2$ ratings on MAQ/1 was not at either reunion because (also see Section 3 above) he returned home before the beginning of his programme. It is not the
prerogative of this study to discuss the private circumstances of Z's departure. It is, however, legitimate to note the comment on Z's Phase One Profile:

"A question mark over his motivation, though. Very vague, almost dégagé about his specialist studies and not obviously aware of likely difficulties."

This was based on Post-Test 4 (Interview) and on his 22nd and 16th places respectively on the MAQ/1 measures of ELTI course and extra-curricular activities (see Chapter Seven, Section 4 above).

But the general message on social contacts in C₂ is that there are common constraining factors on both sides, both demanding from overseas students awareness, flexibility of personal constructs and determination if they are to be successfully overcome. Again, the last words on this particular issue may be left to the participants themselves. This is K, in his December 1980 essay:

"I think the British are definitely discriminating, I mean by discrimination one's feeling to be different (not necessarily superior) from another person. In my opinion discriminating is one thing, its manifestation is another. It is the manifestation which determines the tolerance and British people are definitely very tolerant. With my experience in UK I have learnt more about social life, how one should take into consideration customs and habit of people he is dealing with."

And this is E on the same occasion:

"At the beginning when you are trying to understand what they are looking for, inside you, when I couldn't approach more to anybody, when I felt loneliness and that I was a foreigner, many questions only one answer: "Think and link" and go ahead, change your attitude, discover again that there are many things to do, show them that you are still there, that also you can be responsible .... get to a point where you feel that you are learning but also you are living. In conclusion, approach them because unlikely, with some exceptions, will approach you. Once you do that they are much more warm than the reputation they have. Then you didn't realize when and where you are there, living and sharing with the English people."
Now, given the complexities of the $C_2$ situation and the variability of individuals' approaches to it I would not have predicted corroboration with my small group of any of the three major broad hypotheses mentioned in Chapter Three, Section 2 above. The national status hypothesis is not supported by my data since none of the interviewed participants detected hostility to their particular countries and only two of the vast majority who noted $C_2$ ignorance of $C_1$ considered it qualified as a problem, most of the group feeling it was compensated for by an interest in $C_1$ which they enjoyed responding to. (It is worth remembering too, in this connection, that the group shows strikingly little evidence of chauvinism when you consider their ratings of national characteristics on MAQ/1, 2 and 3 generally score $C_2$ higher than $C_1$).

The modified culture contact hypothesis has already received sufficient indirect attention here for it to become obvious that my kind of case study research reveals too many intervening variables with too few (and too individually different) people and contexts to support a straight causal relationship between the amount of $C_2$ contact and academic/social success. This is true, to some extent, with the U-curve hypothesis although participants were asked, at their main interviews as well as on MAQ/2 and 3 to generalise about their state of 'happiness' during their stay. Taking the mean ratings of all participants for each specified time (eg on arrival, after 1 month, after 2 months etc) and plotting them on a graph there is, somewhat surprisingly, a mild but unmistakable tendency towards a higher-lower-higher pattern of morale, both for the 25 participants responding in December and for the 19 still here in May/81. Figure 8.10 illustrates this:
The distribution (from a full scale ranging from 1 (very unhappy) to 5 (very happy)) is very tight and the averaging hides the many individual differences in level and patterns of contentment that we now know exist. The U-Curve Hypothesis tends to emphasise satisfaction without explicit reference to expectations. My concern with the relationship between the two requires a re-assessment now.

5. Profile 1 and Profile 1 Connections: towards some conclusions

The potentially most revealing single index in Phase Two is the participants' response to the request, at their main interview, for a rating of their general expectations and satisfactions from their C2 study period, not restricted as was the similar rating discussed more fully above, to their training programmes alone. The following key points on these data will lead us towards general Profile 1:Profile 2 conclusions.

For the group as a whole, the general expectation:satisfaction ratio is slightly closer to a positive figure (ie one or higher than one) than the similar rating for their training programmes, at .92 as against
419. The implication is that in spite of the socio-cultural difficulties, in spite of the generally high approval of the training programmes, satisfaction with the C2 experience generally is marginally closer to expectations than it is with the study experience alone.

And the explanation is, statistically, at any rate, straightforward. The group's expectations of their training programmes were relatively higher (about halfway between high and very high) than their expectations of the whole C2 experience (just less than quite high). The latter were thus, it may be assumed, more easily satisfied.

Now, it is natural to feel that a key dimension of any profile of success would be a positive satisfaction:expectation ratio from a participant on both his main course and his perception of the C2 experience as a whole. In fact, the follow-up interview data produce only 6 such participants out of the 25 seen at their receiving institutions. Figure 8.11 summarises their profiles in terms of variables germane to the present discussion. (Other presage information is given in Figure 5.2 above).

Key points in Figure 8.11 overleaf, as we build towards final conclusions on the power of Phase One variables to predict Phase Two outcomes are:

1. That positive perceptions and outcomes are more likely for participants on short, 'non-academic' training courses than for those on longer academic ones; K, M and U are three of only five participants in the whole group who are on short 'special' courses with no formal assessment. Yet they make up half of the small sub-set of those expressing all-round satisfaction.
### Phase One Data (Profile 1)  
### Phase Two Data (Profile 2)

<table>
<thead>
<tr>
<th>PARTICIPANT</th>
<th>Programme Type</th>
<th>PF + CS</th>
<th>TL Post-Test %</th>
<th>TL Final Test %</th>
<th>Training Outcome</th>
<th>Training Satisfaction Expectations (High = 4)</th>
<th>General Satisfaction Expectations (High = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>PGD/'Arts'</td>
<td>$Q_1^+$</td>
<td>63.2</td>
<td>71.4</td>
<td>Pass</td>
<td>$\frac{2.5}{2.5} = 1.0$</td>
<td>$\frac{3}{3} = 1.0$</td>
</tr>
<tr>
<td>I</td>
<td>MSc/'Science'</td>
<td>$Q_1^+$</td>
<td>63.6</td>
<td>70.4</td>
<td>Good Yr.1 Reports</td>
<td>$\frac{3}{3} = 1.0$</td>
<td>$\frac{3}{3} = 1.0$</td>
</tr>
<tr>
<td>K</td>
<td>Special/'Sc'</td>
<td>FI+</td>
<td>58.4</td>
<td>-</td>
<td>Good Reports</td>
<td>$\frac{3}{2} = 1.5$</td>
<td>$\frac{4}{4} = 1.0$</td>
</tr>
<tr>
<td>M</td>
<td>Special/'Sc'</td>
<td>FI+</td>
<td>53.8</td>
<td>-</td>
<td>Good Reports</td>
<td>$\frac{3}{2} = 1.5$</td>
<td>$\frac{4}{4} = 1.0$</td>
</tr>
<tr>
<td>P</td>
<td>MA/'Arts'</td>
<td></td>
<td>53.6</td>
<td>64.4</td>
<td>Pass</td>
<td>$\frac{3}{3} = 1.0$</td>
<td>$\frac{4}{4} = 1.0$</td>
</tr>
<tr>
<td>U</td>
<td>Special/'Sc'</td>
<td>$Q_1^+$</td>
<td>41.2</td>
<td>-</td>
<td>Good Reports</td>
<td>$\frac{4}{3} = 1.3$</td>
<td>$\frac{4}{3} = 1.3$</td>
</tr>
</tbody>
</table>

$\bar{x} = 55.6$ ( $\bar{x} = 68.7$ )

$\frac{7}{6} = 1.1$

$\frac{3.1}{2.8} = 1.1$

$\frac{3.7}{3.5} = 1.1$

**Figure 8.11:** Phase One presage and programme variables related to Phase Two outcomes for the sub-group perceiving positive overall and training programme ratios.
2. That the TL variable, as evaluated in Phase One, would not have been a sufficiently powerful predictor for these six participants. The most obvious reason is the fact that the short specially arranged training programmes undertaken by K, M and U either enabled them to adapt their own specialist training activities to their own TL competence and performance levels (as was the case for M and U, who were researching independently using the institutional facilities and expertise available to them) or (as with K) were designed to make allowances for students with particular specialist expertise but, perhaps, with a limited TL level. In these three cases, then, a relatively low TL level (cf U's Post-Test mean of 41.2% and rank-order of 25th out of the 27) was sufficient and, thus, a high level would be unlikely to correlate any more strongly with eventual success. The same might be said for students such as C and P, whose programmes were not short or 'non-academic', but were specially designed for overseas students. However, C's high Phase One Post-Test mean and her above average TL improvement were always likely to prove positive predictors. As for P, below the all-test mean on Phase One, ranked 19th on the post-tests with only an average improvement, one would not have made a positive prediction with much confidence. Except that he is one of the stronger performers on Test 2 Part A (ranking 4th on the pre-test, 9th on the post-test), the competence-oriented sub-test which Section 2 above has shown to be a somewhat unexpectedly powerful predictor. Participant I is the only one of the six on a long, 'non-special', academic course. His consistent, efficient all-round test performances in Phase One gave every TL reason for confidently optimistic prediction.
3. That the generally positive relationships between the $Q_1^+$ personality factor and field independence in cognitive style are corroborated rather than refuted by evidence from this group of six. Three of the group are $Q_1^+$, four FI+ and two are characterised by both traits. These are occurrences well above chance given the size of the $Q_1^+$ (n=10) and FI+ (n=9) sub-sets.

4. On presage factors of regional background the six are mixed. There is a slight balance in favour of 'science' rather than 'arts', female over male.

The following four Figures cover the remaining participants, again grouped according to the satisfaction ratios. They allow a thorough, cumulative pursuit of falsification, that is the testing of hypotheses emerging from Figure 8.11 leading towards valid inferences and conclusions to be described in my final section. Figures 8.12 to 8.14 thus focus on:

1. participants with a positive general but a negative programme ratio
2. participants with a positive programme but a negative general ratio.
3. participants with negative ratios on both.

1. There are no short, 'special' non-assessed training programmes represented in the group in Figure 8.12 opposite. Yet the group, with the exception of Participant Q, is a successful group, given their training outcomes and general satisfaction ratio. You clearly do not have to be on a short 'special' course to be predicted for success.

2. Four of the group have Phase One TL profiles giving reason for confident prediction (ie E, L, O and Y ($\bar{x}$ post-test rank-order = 5th). H, Q and z ($\bar{x}$ rank-order = 21) were identified by their receiving
<table>
<thead>
<tr>
<th>PARTICIPANT</th>
<th>Programme Type</th>
<th>PF + CS</th>
<th>TL Post-Test %</th>
<th>TL Final Test %</th>
<th>Training Outcome</th>
<th>Training Satisfaction Expectations (High = 4)</th>
<th>General Satisfaction Expectations (High = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>Special(Sc)</td>
<td>Q1+</td>
<td>71.0</td>
<td>76.0</td>
<td>Very good Yr.1 reports</td>
<td>3 = .75</td>
<td>4 = 1.0</td>
</tr>
<tr>
<td>H</td>
<td>MSc(Sc)</td>
<td></td>
<td>52.2</td>
<td>56.6</td>
<td>Good Yr.1 reports except TL</td>
<td>3 = .75</td>
<td>3 = 1.0</td>
</tr>
<tr>
<td>L</td>
<td>PGD(Arts)</td>
<td></td>
<td>64.4</td>
<td>72.0</td>
<td>Pass</td>
<td>2 = .66</td>
<td>4 = 1.0</td>
</tr>
<tr>
<td>N</td>
<td>MSc(Sc)</td>
<td></td>
<td>49.4</td>
<td>59.4</td>
<td>Pass</td>
<td>1 = .2</td>
<td>3 = 1.0</td>
</tr>
<tr>
<td>O</td>
<td>MA(Arts)</td>
<td>Q1+</td>
<td>65.0</td>
<td>73.6</td>
<td>Very good Yr.1 report</td>
<td>4 = .89</td>
<td>5 = 1.0</td>
</tr>
<tr>
<td>Q</td>
<td>LLM(Arts)</td>
<td></td>
<td>53.4</td>
<td>55.6</td>
<td>Fail</td>
<td>3 = .75</td>
<td>5 = 1.0</td>
</tr>
<tr>
<td>Y</td>
<td>MSc(Sc)</td>
<td>Q1+</td>
<td>65.6</td>
<td>68.6</td>
<td>Pass</td>
<td>2.5 = .83</td>
<td>4 = 1.33</td>
</tr>
<tr>
<td>Z</td>
<td>MSc(Sc)</td>
<td></td>
<td>44.4</td>
<td>58.2</td>
<td>Pass but TL problem</td>
<td>3 = .75</td>
<td>5 = 2.5</td>
</tr>
</tbody>
</table>

|          |                |        | x=58.2         | x=65           |                  |                                      |                                      |

Figure 8.12: Phase One presage and programme variables related to Phase Two outcomes for the sub-group perceiving positive overall but negative training programme ratios.
institutions as handicapped by TL level, whether they finally passed or failed their main courses. N has described his TL problems as very considerable (in fact requesting from this researcher corroboration from his C1 employer that his non-arrival TL level was 'inadequate'). Thus the hypothesis emerging from Figure 8.11 above, that TL is not a powerful predictor of success needs modifying. It may not be too crucial a factor in short, special, non-assessed programmes but a detailed TL profile can, on the evidence of Figure 8.12, predict the degree of TL-related main course difficulties.

3. Four more of the Fl+ sub-group emerge here as successes, E and 0 despite (or because of?) the constantly self-demanding, self-critical, hyper-analytical pressures they have put on themselves (passim above). And note that 0 insisted on rating her main course expectations at a level actually above the 'permitted' maximum! Even more significant, among the overall successes covered in Figures 8.11 and 12 we already have 4 of the 5 participants characterised as combining the analytic, thinking introversion, meaning-oriented Fl+ qualities (see Chapter Seven above) with the questioning, experimenting tendencies of Q1. The hypothesis suggesting a causal relationship between these traits and C2 training success is not yet falsified.

4. There is again no significant regional-background related factor in this group. The male participants now even out the imbalance in the Figure 8.11 group and the 'science' 'arts' proportions are representative of the group as a whole.

What emerges, now, from the sub-group whose training programme satisfaction ratio is positive but who are less sure about the over-expectation all C2 experience? Figure 8.13 informs.
<table>
<thead>
<tr>
<th>PARTICIPANT</th>
<th>Programme Type</th>
<th>PF + CS</th>
<th>TL Post-Test %</th>
<th>TL Final Test %</th>
<th>Training Outcome</th>
<th>Training Satisfaction Expectations (High = 4)</th>
<th>General Satisfaction Expectations (High = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>PGD(Arts)</td>
<td></td>
<td>66.8</td>
<td>74.5</td>
<td>Pass</td>
<td>$\frac{3}{3} = 1.0$</td>
<td>$\frac{3}{4} = .75$</td>
</tr>
<tr>
<td>B</td>
<td>PGD(Arts)</td>
<td></td>
<td>65.8</td>
<td>71.8</td>
<td>Pass</td>
<td>$\frac{3}{3} = 1.0$</td>
<td>$\frac{2}{4} = .50$</td>
</tr>
<tr>
<td>G</td>
<td>MSc(Sc)</td>
<td>Q₁⁺</td>
<td>63.4</td>
<td>66.2</td>
<td>Pass</td>
<td>$\frac{4}{3} = 1.33$</td>
<td>$\frac{4}{5} = .80$</td>
</tr>
<tr>
<td>S</td>
<td>PGD(Sc)</td>
<td>FI⁺</td>
<td>62.6</td>
<td>64.2</td>
<td>Pass</td>
<td>$\frac{3}{2} = 1.5$</td>
<td>$\frac{4}{5} = .80$</td>
</tr>
<tr>
<td>W</td>
<td>Special(Sc)</td>
<td></td>
<td>24.0</td>
<td>-</td>
<td>Very big TL problem</td>
<td>$\frac{4}{3} = 1.33$</td>
<td>$\frac{3}{4} = .75$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\bar{x}=56.5$</td>
<td>($\bar{x}=69.1$)</td>
<td>$\bar{x}=\frac{3.4}{2.8}$</td>
<td>$\bar{x}=\frac{3.2}{4.4}$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.13: Phase One presage and programme variables related to Phase Two outcomes for the sub-group perceiving positive training programme but negative overall ratios.
1. W is the only 'special' short-course member here and although he could not fail in the formal sense, it was admitted by his tutors that his severe TL problems made it difficult to evaluate how much he really got out of his training, in spite of his conscientiousness and positive attitudes. Short 'special' courses cannot guarantee success.

2. Phase One TL assessments thus again emerge as more powerful predictors of TL-related training programme outcomes here. Apart from W, this sub-group (x rank-order on the post-tests = 7.1) was rated fairly confidently (though with individual FL/SL, CALP/BICS variations described in Chapters Six and Seven above).

3. Our final Q1+, P1+ participant appears here, again as a success. (In fact G has been offered the chance of doctoral research by his receiving institution). A, B, and S show, thankfully, that you do not have to be characterisable in these personality or cognitive style terms in order to succeed, but they do not falsify a hypothesis that these traits may help.

4. This sub-group is all-male but the male/female ratio so far (ie for Figures 8.11 to 8.13) is representative of the group as a whole. National backgrounds are similarly mixed. But taking the three groups so far, there are signs of a higher probability of programme satisfaction at PGD than at Master's level. (The running count is that 4 out of 5 have rated their diploma training positively, compared with 3 out of 8 on Master's courses.) But we need to study the final two sub-groups for further evidence on this and other hypotheses.

Figure 8.14 shows the 5 participants with negative satisfaction ratios on both training programmes and overall C2 experience.
<table>
<thead>
<tr>
<th>PARTICIPANT</th>
<th>Programme Type</th>
<th>PF + CS</th>
<th>TL Post-Test %</th>
<th>TL Final Test %</th>
<th>Training Outcome</th>
<th>Training Satisfaction Expectations (High = 4)</th>
<th>General Satisfaction Expectations (High = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>MSc(Sc)</td>
<td>Q₁⁺</td>
<td>66.0</td>
<td>-</td>
<td>Did not complete</td>
<td>(\frac{1}{4} = .25)</td>
<td>(\frac{3}{4} = .75)</td>
</tr>
<tr>
<td>F</td>
<td>PGD(Arts)</td>
<td>Q₁⁺</td>
<td>59.4</td>
<td>69.6</td>
<td>Pass</td>
<td>(\frac{4}{5} = .80)</td>
<td>(\frac{3}{4} = .75)</td>
</tr>
<tr>
<td>J</td>
<td>PGD(Arts)</td>
<td>Q₁⁺</td>
<td>64.5</td>
<td>-</td>
<td>Did not complete</td>
<td>(\frac{2}{4} = .50)</td>
<td>(\frac{2}{4} = .44)</td>
</tr>
<tr>
<td>R</td>
<td>PGD(Sc)</td>
<td></td>
<td>58.8</td>
<td>61.0</td>
<td>Yr.1 pass</td>
<td>(\frac{3.5}{4} = .88)</td>
<td>(\frac{3.5}{4} = .88)</td>
</tr>
<tr>
<td>X</td>
<td>MSc(Sc)</td>
<td></td>
<td>62.6</td>
<td>70.2</td>
<td>?</td>
<td>(\frac{2.5}{3} = .83)</td>
<td>(\frac{2}{3} = .67)</td>
</tr>
</tbody>
</table>

|           | 3               | 0      | \(\bar{x} = 62.3\) | \(\bar{x} = 66.9\) | \(\bar{x} = 62.6\) | \(\bar{x} = 2.6 = .65\) | \(\bar{x} = 2.7 = .69\) |

Figure 8.14: Phase One presage and programme variables related to Phase Two outcomes for the sub-group perceiving negative ratios on both training programmes and overall.
Points relevant to the cumulative hypothesis-testing argument are:

1. There are no 'special' short-course members in this group, which tends to corroborate the hypothesis with such courses as a positive predictor of success.

2. Phase One TL assessments did not predict the training programme non-completion of D or J but then neither were their non-completions directly related to TL problems. As we have seen in the summaries connected with Test 2A in Section 3 above, D's narrowly specific training requirements were not satisfied and J (in spite of coping with the extreme demands of teaching practice in TL to native speakers) had to return home because of C1 MSc regulations. For F, R and X, TL predictions proved reasonably accurate though (see discussion point 2 on Figure 8.8 above) not in linear ranking terms because the three contrasting levels of their programmes required correspondingly contrasting TL levels. X being on one of the most rigorous MSc courses in the whole group, F on an overseas-student oriented diploma and R on a laboratory assistants' certificate course.

The academic situations of these three participants exemplify the crucial need to investigate TL level in interaction with specific programme requirements.

3. Three of this group (D, F and J) are Q1+, none of them FI+. Now, although their negative ratings on the satisfaction ratios does not make them failures (after all, F not only passed but has gone on to MA studies), the Q1+:successful outcome hypothesis must be questioned. All three Q1+ participants here have been mentioned in Chapter Seven as well as in this chapter as enthusiastic for C2 contact and/or positive on their ratings of C2 characteristics though F, it
may be remembered was, until May 81, the one who perceived her own compatriots in a more favourable light than $C_2$. The consistently high values assigned by these three on their expectations may warrant a check on whether this tendency is a general $Q_1+$ trait in the group. The answer seems to be yes. Expectation ratings by all the $Q_1+$ members ($n=10$) have a mean of 3.9, vs 3.5 for those not scored as high on the personality trait ($n=14$). Does the questioning, critical aspect of the trait, the sceptic's questioning of the ideal of validity (not the cynic's denial of the sincerity of such claims) actually imply high expectations rather than low? After all, a desire for innovation and initiative (other aspects of the trait) suggests you still expect good from them. Nor should we necessarily equate non-completion with 'failure'. In his reply to an informal follow-up letter after his return home D (after the disarming opening: "I'm sorry for the delay in answering your letter/test? test/letter? .....") writes this:

"The course at (....) up to now did not prove useful in my work, but the books and the notes that I brought, yes. My attitudes at work are quite the same but, I did change my attitudes to life for better. And I have good memories of England and of the British people. I believe that you did more than I expected to solve my problems, but the problems that I had did not depend on you .... And finally I do not think my study fellowship as a failure."

$Q_1+$ still seems a positive attribute rather than a negative one provided that the $Q_1$ student and his teachers can live with the awareness and questioning tendency involved. But this is not just a $Q_1$ issue. Higher expectations are, almost by definition, harder to satisfy than lower ones. It will not have escaped the perceptive reader that in all 4 pairs of mean expectation:satisfaction statistics in Figures 8.11 to 8.14, the higher the expectation figure is, the less likely it is to be equalled by the satisfaction index.
4. There is nothing significant about the regional background, male/female or Arts/Science breakdown of this fourth group. The presence in it of three PGD students evens out the PGD/MSc proportions on satisfaction but we do end up with 4 out of 8 participants rating their diploma courses positively on the ratio against only 3 out of 12 at Master's level. Predictions of satisfaction should take account of the likelihood that is somewhat harder to come by at the higher academic level.

Finally, before my general conclusions, I need to present, in Figure 8.15 overleaf, the incomplete data for the remaining three participants, T who did not give satisfaction responses at her interview, then V and Z who returned home too soon to be interviewed at their receiving institutions (see Section 3 above).

The following are key points relating to the current hypotheses.

1. Since two of the three members of this group are on 'special' courses, one on a PGD programme, the emerging hypothesis suggesting a relationship between the presage factor of programme type, and $C_2$ outcome must take account of this group of three even though they do not figure on the satisfaction map. Thus, we finally have 5 'successful' special programme completions against 2 non-completed, which must now serve as a reminder that the presage variable of 'tailor-made' training programmes is a powerful but not a sufficient predictor of $C_2$ success. And since T's unfortunate failure was on a PGD course, the modified final tally relating academic course level and success/failure (as a formal assessment category) shows, at Master's level 8 passes, two successful first years, one doubtful first year, one fail
<table>
<thead>
<tr>
<th>PARTICIPANT</th>
<th>Programme Type</th>
<th>PF + CS</th>
<th>TL Post-Test %</th>
<th>TL Final Test %</th>
<th>Training Outcome</th>
<th>Training Satisfaction Expectations (High = 4)</th>
<th>General Satisfaction Expectations (High = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>PGD(Sc)</td>
<td></td>
<td>43.4</td>
<td>57.8</td>
<td>Failed</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>V</td>
<td>Special(Sc)</td>
<td></td>
<td>35.8</td>
<td>-</td>
<td>Did not complete</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Z</td>
<td>Special(Sc)</td>
<td>FI+</td>
<td>57.2</td>
<td>-</td>
<td>Did not begin</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0 1</td>
<td>$x = 45.5$</td>
<td></td>
<td>$\bar{y} = 5$</td>
<td>$\bar{y} = 3$</td>
<td></td>
</tr>
</tbody>
</table>

Figure 8.15: Phase One presage and programme variables related to incomplete Phase Two outcomes for the sub-group not making expectation/satisfaction ratings
and one non-completion; at PGD level, 4 passes, one successful first year, one fail and one non-completion. Thus, although Master's level courses have caused more TL difficulty, the formal success rate at Master's and PGD level is not significantly different. But it is noted that all participants on academic courses designed for overseas students (PGD and Master's) have been successful.

2. The TL profiles for T and V certainly predicted TL-related problems on their training programmes. In T's case, receiving institution tutor interviews and reports do not explicitly support these predictions, tending rather to ascribe T's problems to subject background, C2 social/academic adjustment problems and homesickness. My validating interviewer, however, suggests an understandable and respected protectiveness here. Certainly T's TL level was a significant factor in her actual main training course performance when her seminar presentation was found 'incoherent'. Certainly all our Phase One and Two TL assessments produced worryingly low scores from T (see case studies in Chapters Six and Seven). The unfortunate Phase Two outcome for T probably corroborates the importance of TL as a predictor along with the cognitive/affective and social factors I have investigated.

V's sad C1 family illness was the reason for her early departure. Given the available comparison with W (whose TL level was without doubt the lowest of the whole group) on the same short special course, it is likely that V would otherwise have completed her programme, though not without TL-related problems. About Z, whose premature departure is discussed in Section 4 above, we can only surmise. His overall post-test score is just above the mean, as it is on the interestingly
predictive competence Test 2A. Since his course was to be tailor-made ('special' though not short) this should have meant that the TL itself would not have been the decisive problem. But then, as we know, it was not, anyway. This brief look at the last group of three does not refute the hypothesis that a mixed competence- and performance-oriented battery such as the one used in this study is a crucial, though not sufficient, predictor of $C_2$ EAP outcomes. (see Section 6 below).

3. Since Z is the only participant in the whole group who is FI+ but without a successful outcome, he would, if my sample were larger, probably have 'exception-that-proves-the-rule' status with regard to the FI+ hypothesis. But the sample is not large so he does not have such a status. At least his case serves to refute a simplistic equation between GEFT field-independence and the strong personal independence that might give overseas students an immunity to all pulls from home. There are no Q+1 participants in this final group. Thus, on balance, positive relationships between the Q+1 and FI+ factors of personality and cognitive style and $C_2$ training outcomes are not yet falsified (see Section 6 below).

4. T, V and Z are from different regional backgrounds. Their presence in this final group makes no difference even to the most tentative cross-cultural inferences regarding success/failure that such a small-scale study might have made (see Chapter Two, Section 3.7) but does not.

Nor are 'arts' vs 'science', 'male' vs 'female' differences in formal success or failure shown to be significant now these final three cases have been considered.
6. The Research Hypotheses Re-visited: some theoretical, practical and methodological implications

This renewal of connection with my general Research Hypotheses represents a return from the narrower level of specific hypothesis testing to the broader level of inference, allowing some general conclusions to be drawn. The task of generalisation is not easy, however. The commitment in the research to a multi-dimensional approach to profiling and the complexities of individual variability that it has inevitably revealed, mean that the inferences drawn must always be seen in the still exploratory and tentative empirical contexts from which they have emerged. The summary here will, in line with its Popperian influence, set up new problems ready to be submitted to further falsification processes.

6.1 Research Hypothesis 3: TL elicitation/assessment for learner profiles

1. The case for the use of a C2 pre-sessional TL test battery to collect finer diagnostic and profiling information on learners is strong, especially when, like my group, they arrive with a variety of TL assessments, made at different times, under different circumstances by different people using different criteria, mainly with the aim of measuring proficiency against a supposed globally valid cut-off point. The value of such a battery is clearly enhanced when its use is designed into the pre-sessional programme itself, with data from continuous assessment or experimental sources used to enrich learner profiles in a mutually validating relationship with test data. The longitudinal use of identical, equivalent or parallel test forms is not only helpful in evaluating changes in TL level but should also have a role to play in the long-term validation procedures for global TL proficiency test
batteries. But follow-up test results must also seek mutual-validating relationships with other data sources. In my case, Phase Two TL profiles would have been considerably less valuable (or valid) without the TL evaluations contributed by receiving institutions and by the participants themselves.

2. My attempts to design tests of Performance informed by the communicative criteria developed in Chapter Four, were not entirely successful as the analyses of Test 3 (Reading/report-writing) in Chapters Six and Seven show. But participant profiles would have been much poorer, less prognostically powerful without the insights provided by the performance-oriented parts of the test battery. The criteria for the construction and evaluation of performance tasks seem to be steps in the right direction and the fairly strict reliability/validity checking procedures probably prevented significant profiling errors caused by performance test invalidities. But more confident solutions to the problem of relating test tasks to actual main course activities are needed. Future research in this area should be informed by more intensive on-site observation and analysis of such activities. It has to be remembered that we can only arrive at the required standardised versions of communicative performance tests by first going through the difficult stages of trying to mirror the complexities of reality, then idealising, from criterion-referenced evidence, on the extent to which tests can predict validly when some of these complexities are stripped away.

3. The indications from my study are not that the 'stripped-away' versions will be like the non-performance elements in my battery (ie Test 2A, Part A, the dictations and cloze passages). These do not earn the label of 'competence tests' simply because they are
indirect modes of assessment. The analyses of their Phase One and Two uses here have indicated that they do tap competence as the underlying knowledge of the linguistic 'what' but only unevenly predict the communicative 'how'. The competence measures have proved essential in the test battery because of their reliability, explicability and easier generalisability. (Test 2, Part A is a surprisingly powerful overall predictor). But as case studies such as those in Chapters Six and Seven have shown, competence tests sometimes fail to predict individual differences in the gaps between knowing and doing, to an extent that could lead to significantly inaccurate profiling. All my correlational analyses have confirmed that a mixed competence/performance TL test battery is more informative and prognostically valid than a battery of tests exclusively of the one kind or the other. And ELTI teacher comment on the relationship between TL evaluation and pre-sessional course design suggests an additional practical advantage of the mixed-construct approach to assessment. The combined focus on competence and performance reflects the pedagogical rationale. If there is going to be a back- (or bow-) wash from tests to course, it is better if it encourages learners to think of their competence and their performance.

4. The crucial role of TL teachers in this research should have emerged in spite of its specific focus on group and individual learner profiling. The invaluable help I have received from particular teachers in test construction, scoring, administration, validating continuous assessment comment, interviewing and so on is clear from my acknowledgements. But this research is teacher-oriented in another important way, too. The level of the statistical analysis of data is essentially practical, feasible for non-academic teaching-
related research, not dependent on (though it has been supported by) computer technology. (The most valuable lesson from the Imperial College, London, computer-programming course I attended in connection with this study was about when not to use the computer unnecessarily.) The scale of my study and the nature of my sample has restricted me, in the main, to a non-parametric statistical approach, without recourse to factor analysis, for example, where the small size of the group would have risked the emergence of 'factors' overly influenced by the performance of individuals. The result should be increased applicability (and accessibility) of the statistical data to practising course designers, teachers and testers.

5. My main perceived errors of omission or commission should also inform future work connected with TL profiling. I needed more time for the piloting of my tests. Although their use with the 1981 ELTI group made it possible to make post-facto adjustments (see Section 3 above) it would clearly have been better to have piloted all the tests with a first group, modified and concentrated their research use on a second then re-checked tests and findings with a third. The temporal implications of this ideal are obvious, of course. They also underline the extra problem caused by group-specific performance tests. By definition these require closely matched groups for piloting, main use and post-checking; almost equally by definition, such groups tend to arrive but once a year.

A more serious deficiency in the study is the lack of observation and analysis of TL learning in progress (or process), especially in Phase Two. The experiments described in Chapter Seven, Section 3 were revealing and, again, practical in the sense that they are adaptable
to other TL training programmes as learning as well as data elicitation exercises. But my design did not allow for close linguistic analysis of the interactions as examples of language acquisition; their potential for such analysis, for example, with an interlanguage or monitor theory perspective, is obvious from the transcripts in Chapter Seven as well as the original recordings.

Still, the product and process of the TL elicitation and assessment data have served their general research hypothesis purposes. They facilitate the investigation of my original problem, helping to reveal new problems for further investigation. Chapters Six, Seven and now this chapter have seen the elicitation and use of TL data to inform learner profiles: thus:

TL elicitation/assessment → Learner profile data

My instrumental RH3 is corroborated.

6.2 Research Hypotheses 1 and 2: cognitive/affective and social factors related to TL in multi-dimensional profiles for prognostic C2 use

1. Even the teaching, testing and experimenting experience of Phase One soon underlined the illogic of expecting learner profiles based exclusively on TL proficiency and progress to adequately predict future success in contexts where specialist subject and academic community factors take over from TL study as the main variables. Phase Two experience in those contexts confirmed this illogic. My attempts to broaden and refine Phase One profiles to take account of personality factors encountered severe problems stemming from the unreliability of the chosen psychometric instrument and the influence of TL. It is hoped that the documentation of these problems and measures to overcome
some of them will prove helpful for future research into TL:PF inter-relationships. Such research may also benefit from the combined quantitative/qualitative approach to data on personality, again taking account of the crucial role of teacher and tutor personality trait evaluations. Although I failed to get my intended 16 personality-factor profiles I did end up with a factor (Q1+) which proved of revealing interest throughout the study as it had promised to in pilot observations with the 1979 group. A further methodological point is worth bearing in mind in research, which, like mine, is not exclusively psychometric. If a researcher has a whole set of personality factors to take account of as different kinds of data from different sources are analysed, the method adopted here of using just two key PF and cognitive style variables as reference points for the cumulative interpretation of new data along with the reinterpretation of the old, would probably not be practicable.

2. The Group Embedded Figures Test leading to TL-free measurements on field dependency generally fulfilled its promise as a source of insight into cognitive styles relevant both to TL learning and academic activities. If the kind of experiments I ran using this cognitive style as the independent variable can be replicated and validated, the implications are considerable. The fact the FI+ and FI− do seem to co-vary with different approaches to, and efficiency in, study-related tasks such as those investigated in Experiments 1 to 3, may be justification for the pedagogically-motivated matching of pre-sessional training tasks to styles. But this would be done not so that learners have the luxury of having their activities tailored to their strengths but rather so that they may develop strategies to
deal with necessary future activities that may otherwise expose their weaknesses.

3. The use of MAQ/1, 2 and 3 to extend and refine profiles in the light of reasonably objective data on various dimensions of attitude and motivation was revealing on perceived TL orientations and comparative $C_1/C_2$ attitudes, but less so on motivational intensity as measured by the items intended to focus on study habits or feelings. My research would have benefited from closer, subjective insights into this dimension; there would be a strong case in future similar longitudinal inquiries for more direct, longer-term observation of students in their receiving institutions with an attempt to get as close as possible to participant-observer status. As it is, the MAQ's are an example of usefully repeatable data elicitation instruments combining the experience and validity of previous researchers' inquiries with innovations such as my re-vamping of the orientation index items and the inclusion of a U-curve related measure. Again, however, MAQ data would not have enriched the profiles much if they had not been used in combination with related information elicited in other ways.

4. Especially from the receiving institutions interviews. These may be considered to have achieved their purpose of providing a useful source of insights on TL, academic, cognitive/affective and social matters in a way which combined the qualitative advantages of free talk with the quantitative advantages of responses susceptible to statistical analysis based on pre-coded scale values. Two practical research points here, though. Lengthy, semi-structured, in-depth interviews requiring the discussion of individual perceptions of personal experiences and problems can probably only be carried out by people who are quite close to the participants. Thus it would be
difficult to handle a large sample in this way. Secondly, a promise of confidentiality is probably necessary to achieve the degree of frankness my validating interview partner and I achieved with all but two of the participants (see Section 4.2 above). Although all participant interviews were tape-recorded, they have not, it may have been noticed, been quoted from and have been listened to, also as promised, only by the validating interviewer, a validating listener and myself.

5. The many specific hypotheses tested and discussed in my empirical study have tended to support a cyclical relationship between TL learning and use and cognitive/affective and social factors. (See Chapter Seven, Section 5 above). Thus, Research Hypothesis 1 can be redrawn as follows:

![Diagram](attachment:diagram.png)

And the richer the detail resulting from this interaction of factors the more likely it is that Research Hypothesis 2 receives corroboration, as the empirical findings earlier in this Chapter have shown. The poorer they are, of course, the less likely. On balance, the profiles have proved rich enough. Thus the implied relationship in Research Hypothesis 1 can be supported:

<table>
<thead>
<tr>
<th>Detailed Learner Profiles</th>
<th>Prediction of C₂ future</th>
</tr>
</thead>
</table>
7. The Main Thesis Revisited

My main thesis (see Chapter One, Section 4) suggests that 'training for EAP learner/users in C2 can be fruitfully informed by the systematic, sensitive, multi-dimensional profiling of individuals'. Now that the three Research Hypotheses have shown that such profiling is, given certain problems requiring further investigation, feasible, what kind of training approaches does this profiling imply? Inferences on this will depend, of course, on the reader's interpretation of my evidence and the degree of conviction it seems to carry, but the following are some of the most strongly supported training arguments.

1. The profiling approach itself, a systematised attempt to understand key presage, process, programme and community factors in individual learners' development involves pre-, post- and continuous evaluation procedures that will lead to training programmes with a special emphasis on close teacher-learner understanding and flexibility to individual needs and preferences. There would be a great deal of personal discussion of personal variables (as there was on the 1980 and 81 ELTI programmes) which should heighten the awareness of both sides of the training partnership of how learners can be helped to learn better. There would also be more activities of the kind described in Experiment 3 above, arising out of mutually agreed individual or group needs. Most of my participants agreed that the Hawthorne and Pygmalion effects they themselves recognised as partners in an individual profiling project, were beneficial and re-assuring rather than an extra burden.

2. And awareness was what Participant z's tutor was so urgently asking us to heighten when he wondered why we could not 'help them to realize what it's all about'. The striking need for a balance between
competence-focused TL remedial work (stressed by the vast majority of participants as not to be neglected, seen as irrelevant by none) and academic orientation training is clear from my empirical evidence from under whichever 'heading' it emerges. Some of my participants arrived with very little documentation on what their specialist programmes really involved in terms of media, modes, work load, assessment procedures or even subject component focus. Even more of them were unaware of the implications of such information. The considerable collection of documentary and tutor description of programmes I amassed during Phase Two emphasised just how much we, as pre-sessional teachers, did not know about these programmes. Our course design would have benefited a great deal if we had made this kind of information available to ourselves at the planning stage and for our individual discussions with students.

3. An even more fundamental inference can be drawn from this, and from my finding that TL level is a powerful but not sufficient predictor of C2 EAP success. Assessments of TL are still the main criteria for decisions on whether overseas students attend pre-sessional courses or not and most pre-sessional courses are for remedial English training with a widely varying emphasis (or lack of it) on study approaches. Yet I have still to meet a study of overseas students that denies the existence of 'academic shock' as a factor in their C2 lives. Most universities with experience of overseas students realize this, of course, and design their own pre-sessional English courses accordingly. But they may still be selecting students for such programmes only if they have been classified as needing remedial English. My evidence is that selection (and, therefore, assessment) should be on the basis of UK academic awareness as well as TL level.
The increased demands on pre-sessional training that this would involve are enormous, of course, but this does not alter the fact of its importance. If the implications were accepted by the increasing number of C1 pre-departure training institutions, however, the burden this end could be lightened.

4. Certainly, the pedagogic implications should be taken seriously by trainers who already receive overseas students pre-sessionally (or in session) on the basis of the TL criterion. There seems to be a strong case for as direct a replication as possible of as many of the future study conditions as possible during pre-sessional programmes. Where we at ELTI had a useful collection of authentic specialist lectures available for listening practice, for example, but allowed learners to listen, play-back and pause at their leisure, we should probably have insisted on a more realistic 'once only' approach and asked learners to try to use the notes they had managed to take for further activities, integrating the notes with data from other sources for a report, say, or an essay. Better still, would be 'live' lectures from their future tutors, who would make only the allowances or follow-up demands that they would normally make, an approach successfully used on the 1981 ELTI programme. And one further example that might help to illustrate the point. On the 1980 programme we did ask participants to take part in group projects (see Chapter Seven, Section 4 above). But their project work was self-selected, tutor-facilitated, on topics of general interest with results to be presented at an informal plenary session in a relaxed end-of-course atmosphere. This was quite nice (for those who chose to take an active part) but it bore very little resemblance, as my Phase Two visits showed, to reality. In 1981 the ELTI course designers took a different
line. Individual students were asked to propose a project from their own specialist field which they then had to pursue, without an atypical degree of tutor assistance, through literature search, note-making, report-writing, supervisor feedback and seminar presentation stages. The experience was, apparently, sobering and informative and, incidentally, provided tutors with such valuable insights that a whole section on individual project work was included in the 1981 version of the profiles sent to receiving institutions. The 1980 group and their teachers mentioned the 'cocoon' effect of their ELTI pre-sessional programme. Sometimes the implication of such comment was that it probably made the emergence into the hard realities of specialist training life more fraught.

5. The C₂ social factor picture presented by my study is complex, as Section 4.2 above, in particular, has shown. Most participants agree that coverage of routine social events (banking, shopping, health problems etc) is important and, in line with the message in 4. above, more effective if it includes practice in the real settings. On the question of C₂ contact, where overseas students' combined client and visitor roles are at issue (see Chapter Three, Section 2 above), pre-sessional programme implications are less clear. Given the complex interaction between C₁ and C₂ characteristics and attitudes the issue would seem to be a more likely subject for discussion than for teaching. My participants have certainly provided interesting starting points for such discussions, for example on the subtle theme of the apparent British ability to combine helpfulness with unfriendliness or the possible link between the dimensions of the Q₁+ trait and openness to C₂ contact. Only a few of the group felt that
more formalised approaches to C2 contact were the answer, but pre-
sessional programme designers might take note of the suggestions in 
Chapter Seven, Section 2 above concerning the involvement of UK students 
in such programmes.

The general training implications here are, of course, extrapolated 
from the many detailed findings in my study. Most of these detailed 
findings have some potential specific implications for course 
designers. The relevance of these implications is for the reader to 
decide - which seems a suitable epilogue to the research as a whole.
BIBLIOGRAPHY


Atkinson, J. (1957) 'Motivational determinants of risk-taking behaviour'. Psychological Review 64.


Bartz, W. (1974) 'A study of the relationship of certain learner factors with the ability to communicate in a second language (German) for the development of measures of communicative competence'. Ohio State University, PhD Thesis.


Bradshaw, J. (1981) Untitled lecture given to overseas students at ELTI.


Davies, A. (1964) English Proficiency Test Battery. Published for closed use by NFER and the British Council. NFER.


Davies, A. (1977;2) 'Do foreign students have problems?' In Cowie and Heaton (eds).


Grice, P. (1957) 'Meaning' in *Philosophical Review* LXVI.


Hancock, A. (1964) Survey. Longman.


Hymes, D. (1979;1) 'Educational Ethnology' in Anthropology and Education Quarterly XI/1.


Independent Assessment and Research Centre (IARC) 1978 Working Guide for 16PF Interpretation.


Jakobson, R. (1953) Chapter Two in *Results of the Conference of Anthropologists and Linguists* Memoir 8, IJAL.


Kaplan, R. (1976) 'A further note on contrastive rhetoric'. *Communication Quarterly* 24/2.


Kline, P. (1966) 'Extraversion, neuroticism and performance in a psychology examination'. British Journal of Educational Psychology 36.


Kluckhohn, F. (1940) 'The Participant-Observer Technique in Small Communities'. American Journal of Sociology 46.


ODA (1980) Evaluation of Training Programme for Swaziland. ODA.


Oller, J. (1979;2) 'Communicative Competence: Can it be Tested?' mimeo.

Oller, J. and Hinofotis, F. (1976) 'Two mutually exclusive hypotheses about second language proficiency: factor analytic studies of a variety of language tests'. Paper read at the Winter Meeting of the Linguistic Society of America.


Overseas Students Trust (OST) (1979) Overseas Students in Britain: Some Facts and Figures. OST.


Rubin, J. (1975) 'What the "good language learner" can teach us'. TESOL Quarterly 9.


Savile, P. and Blinkhorn, S. (1976) Undergraduate Personality by Factored Scales: a large scale study on Cattell's 16PF and the Eysenck Personality Inventory. NFER.


UKCOSA. (1979) *Suffering for Success.* United Kingdom Council for Overseas Student Affairs.

Universities Statistical Record (1980) *Numbers of Overseas Students Accepted for courses in Universities.*


Verner, C. (1962) 'A conceptual scheme for the identification and classification of processes for adult education'. Adult Education Association of the USA.

Vollmer, H. (1979) 'Why are we interested in "General Proficiency"?' *Lingua et Signa* 1.


Wijasuriya, B. (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'. MEd Thesis. University of Manchester.


