DEVELOPING A MODEL FOR INVESTIGATING THE IMPACT OF LANGUAGE ASSESSMENT WITHIN EDUCATIONAL CONTEXTS BY A PUBLIC EXAMINATION PROVIDER

N D Saville

This is a digitised version of a dissertation submitted to the University of Bedfordshire.

It is available to view only.

This item is subject to copyright.
DEVELOPING A MODEL FOR INVESTIGATING THE IMPACT OF
LANGUAGE ASSESSMENT WITHIN EDUCATIONAL CONTEXTS
BY A PUBLIC EXAMINATION PROVIDER

by

N. D. SAVILLE

A thesis submitted for the degree of Doctor of Philosophy
of the University of Bedfordshire

January 2009
DEVELOPING A MODEL FOR INVESTIGATING THE IMPACT OF LANGUAGE ASSESSMENT WITHIN EDUCATIONAL CONTEXTS BY A PUBLIC EXAMINATION PROVIDER

N.D. SAVILLE

ABSTRACT

There is no comprehensive model of language test or examination impact and how it might be investigated within educational contexts by a provider of high-stakes examinations, such as an international examinations board. This thesis addresses the development of such a model from the perspective of Cambridge ESOL, a provider of English language tests and examinations in over 100 countries. The starting point for the thesis is a discussion of examinations within educational processes generally and the role that examinations board, such as Cambridge ESOL play within educational systems. The historical context and assessment tradition is an important part of this discussion.

In the literature review, the effects and consequences of language tests and examinations are discussed with reference to the better known concept of washback and how impact can be defined as a broader notion operating at both micro and macro levels. This is contextualised within the assessment literature on validity theory and the application of innovation theories within educational systems.

Methodologically, the research is based on a meta-analysis which is employed in order to describe and review three impact projects. These three projects were carried out by researchers based in Cambridge to implement an approach to test impact which had emerged during the 1990s as part of the test development and validation procedures adopted by Cambridge ESOL. Based on the analysis, the main outcome and contribution to knowledge is an expanded model of impact designed to provide examination providers with a more effective “theory of action”.

When applied within Cambridge ESOL, this model will allow anticipated impacts of the English language examinations to be monitored more effectively and will inform on-going processes of innovation; this will lead to well-motivated improvements in the examinations and the related systems. Wider applications of the model in other assessment contexts are also suggested.
DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the degree of PhD at the University of Bedfordshire.

It has not been submitted before for any degree or examination in any other University.

Name of candidate: Nicholas D. Saville
Signature: ...........................................

Date: 20th January 2009
ACKNOWLEDGMENTS

I would like to acknowledge the support of Cambridge ESOL in the preparation of this thesis, and in particular the encouragement given to me over many years by Dr Michael Milanovic, (Chief Executive, Cambridge ESOL). Many other colleagues in Cambridge have also contributed to my thinking on the issues covered in this work, but in particular I am grateful for thought provoking discussions and collaboration with Dr Neil Jones and Dr Lynda Taylor. I am also grateful to colleagues who assisted in the preparation of the manuscript, especially Kirsty Sylvester.

Several eminent academics provided me with invaluable advice and guidance over an extended period. Professor Peter Skehan initially provided supervision and supported my candidacy at the University of Bedfordshire, and in taking over this role Professor Cyril Weir has shown great patience and insight in overseeing the completion of the thesis. Dr Roger Hawkey has probably contributed more than anyone else to this work through his longstanding collaboration with Cambridge ESOL and especially in his role as coordinator of impact-related research between 1999 and 2004. He read the text closely on several occasions and I am grateful for his many suggestions for improvements and clarifications. I am of course responsible for the weaknesses which undoubtedly remain.

I received helpful comments and suggestions from colleagues following presentations of the developing thesis at language testing conferences in Australia and Poland in 2006. Previously, members of the language testing group in Lancaster University were influential in the early days of the IELTS studies, and mentors and colleagues from the University of Reading, especially Don Porter, Arthur Hughes and Barry O’Sullivan, have all influenced my thinking on the ideas presented here. Antony Kunnan and Jim Purpura have helped me through their friendship and professional collaboration over many years, since their days as graduate students of Professor Lyle Bachman whose influence on this work has been particularly important.

Finally, I would like to acknowledge the support of my wife Cinzia and children Danny and Julia who could have been forgiven for thinking that this would never be finished, despite the many weekends and holidays sacrificed to “work on the PhD”.

## LIST OF CONTENTS

Abstract ............................................................................................................................. ii
Declaration ....................................................................................................................... iii
Acknowledgements ......................................................................................................... iv
List of contents .................................................................................................................. v
List of Tables ..................................................................................................................... xi
List of Figures .................................................................................................................. xii
List of graphs ................................................................................................................... xiii
Abbreviations .................................................................................................................... xiv

### Chapter 1: Introduction ............................................................................................. 1
  1.1 Overview .................................................................................................................. 1
  1.2 The impact of language examinations in educational systems ....................... 3
  1.3 Examinations within society – the British Tradition ....................................... 4
  1.4 Locating impact within educational systems .................................................. 5
  1.5 Locating impact research within Cambridge ESOL ....................................... 6
  1.6 Locating impact within a socio-cognitive model of test validity ..................... 9
  1.7 Locating impact in the test development and validation cycle .................... 11
  1.8 Methodology – meta-analysis and case studies as data ................................ 12
  1.9 Outcomes of the thesis and contribution to knowledge ............................... 13

### Chapter 2: Literature Review ................................................................................. 15
  2.1 Definition of the term impact ........................................................................... 15
     2.1.1 Influences, effects and consequences in society .................................... 15
     2.1.2 Impact – educational policy and the role of examinations .................... 18
     2.1.3 Concerns for the socio-political dimensions of assessment ................ 19
  2.2 Impact and washback in the literature .............................................................. 21
     2.2.1 Some definitions .................................................................................. 21
     2.2.2 The washback continuum – positive, neutral, negative? ....................... 21
     2.2.3 Impact in language assessment .......................................................... 24
  2.3 Washback – hypotheses and models ............................................................... 26
     2.3.1 Does washback exist? .......................................................................... 26
     2.3.2 Washback models .............................................................................. 28
     2.3.3 Researching washback – washback studies in the literature ............... 34
  2.4 Beyond washback ............................................................................................ 39
     2.4.1 Impact research in language assessment ........................................... 39
4.2.3 Period 3 ............................................................................................... 91
4.3 Delimiting the IELTS case ................................................................. 92
4.4 Phases 1 and 2 - development and validation of the instruments .......... 95
  4.4.1 Project 1 .............................................................................................. 96
  4.4.2 Project 2 .............................................................................................. 96
  4.4.3 Project 3 ............................................................................................. 101
  4.4.4 Project Four ....................................................................................... 102
4.5 Phase 3 – the impact survey ............................................................. 103
4.6 Analysis of the IELTS case and lessons to be learned ...................... 106
  4.6.1 Research design and procedures..................................................... 109
  4.6.2 Time-related factors and the iterative cycles ................................. 110
  4.6.3 Better understanding of contexts and the roles of stakeholders .... 112
4.7 Conclusions and implications for the revised impact model ............. 118

Chapter 5: Case Study Two ................................................................. 122

  5.1 Introduction to Case Study Two ........................................................... 122
    5.1.1 The PL 2000 Impact Study .............................................................. 123
  5.2 Delimiting the PL2000 Case ............................................................... 124
  5.3 Main Objectives of the Progetto Lingue 2000 .................................... 125
  5.4 The PL2000 Impact Study ................................................................. 128
    5.4.1 The Research Questions ............................................................... 129
    5.4.2 Design and Data Collection .......................................................... 131
    5.4.3 Summary of the outcomes: .......................................................... 135
      a) Impacts on the classroom ............................................................. 136
      b) Impacts on perceptions ............................................................... 139
      c) Impacts on issues related to assessment ..................................... 147
      d) Impacts on school management and teacher support ............... 150
  5.5 The Cambridge exams within the PL2000 ........................................ 153
  5.6 Follow up to the Impact Study .......................................................... 154
  5.7 Insights and lessons learnt from this case for the Impact Model ....... 155
    5.7.1 General points .............................................................................. 156
    5.7.2 The Context ................................................................................. 157
    5.7.3 Processes at the Macro Level ....................................................... 158
    5.7.4 Processes at the Micro Level ....................................................... 162
    5.7.5 The Participants (stakeholders) .................................................. 164
    5.7.6 Conclusions from the PL2000 case ............................................ 168
  5.8 Summary of implications for the revised model of Impact ............... 171
    5.8.1 General comments .................................................................... 171
Chapter 6: Case Study Three

6.1 Introduction to Case Study Three

6.2 Background to the FLLGP

6.3 Delimiting the case – unit of analysis
   6.3.1 Context and timeline
   6.3.2 Participants
   6.3.3 Scope and research design
   6.3.4 The action plan and data collection schedule
   6.3.5 Data collection
      a) Learning gain and level of proficiency
      b) The characteristics of the participants – learners and teachers
      c) The classroom context: classroom observations and data
      d) Data summary

6.4 Outcomes of the FLLGP
   6.4.1 Project reports and views of the coordinator
   6.4.2 Summary and conclusions

6.5 Analysis of the FLLGP case
   6.5.1 Modelling the classroom context
      a) The milieu – the wider educational context
      b) The school context
      c) The language classroom context
      d) Conditions for language learning in the classroom context
      e) Language proficiency in the classroom context
      f) The role of assessment within the classroom
      g) Individualising assessment – learner’s needs in context
      h) Integration of assessment within the classroom context
   6.5.2 Modelling individual differences in the language classroom
      a) Learners/test takers as individuals
      b) Teachers as individuals

6.6 Conclusions and implications for the expanded model of impact
6.6.1 Context and the impact of assessment ........................................... 226
   a) The communicative language construct ........................................... 226
   b) Approaches to language learning .................................................... 226
   c) Linking formative and summative assessment ................................ 227
   d) New models of assessment to support learning .............................. 228
6.6.2 Implications for the revised model of impact ................................... 230
   a) Local contexts and the hierarchical nature of educational systems. 230
   b) Individual differences of the participants and contextual variables .. 232
   c) Language learning/teaching needs to be understood as change ... 232
   d) How the language construct is communicated .............................. 234
   e) Impact research - case studies and mixed methods ....................... 235

Chapter 7: The revised model of impact and future applications .......... 237

7.1 Introduction .......................................................................................... 237
7.2 Overview of the three case studies ....................................................... 237
7.3 Lessons learnt from case studies .......................................................... 239
   7.3.1 Test features ................................................................................ 239
   7.3.2 Context .......................................................................................... 240
   7.3.3 Participants .................................................................................. 243
   7.3.4 Outcomes .................................................................................... 244
   7.3.5 The Researcher ............................................................................ 245
   7.3.6 Research methods ....................................................................... 246
   7.3.7 The timeline ................................................................................. 247
   7.3.8 Summary ..................................................................................... 249
7.4 The revised and expanded model – a meta-framework ...................... 250
   7.4.1 Dimension 1: re-conceptualisation of impact ............................. 252
   7.4.2 Dimension 2: impact by design .................................................. 255
   7.4.3 Dimension 3: finding out and understanding what is going on .... 257
   7.4.4 Dimension 4: a theory of action ................................................ 259
7.5 Implications for Cambridge ESOL ......................................................... 260
   7.5.1 Revising the maxims ................................................................... 260
   7.5.2 Implementing the revised model within Cambridge ESOL ......... 264
7.6 Applying the model in other contexts .................................................. 265
7.7 Limitations of the research ................................................................. 267
7.8 Concluding remarks ........................................................................... 268
Appendices: ................................................................................................................................................................................. 270

Appendix 4.1 The IELTS Format ................................................................................................................................. 270
Appendix 4.2 The development of ELTS/IELTS ........................................................................................................... 271
Appendix 4.3 The Cambridge-UCLA Language Learner Questionnaire ............................................................... 281
Appendix 5.1 The Italian educational system ............................................................................................................. 287
Appendix 5.2 Educational Reform in the 1990s ............................................................................................................. 291
Appendix 5.3 Foreign Language Study in Italy in the 1990s ....................................................................................... 295
Appendix 5.4 Ministerial circular – 1998 ...................................................................................................................... 298
Appendix 5.5 Details of Progetto Lingue 2000 (original version) ........................................................................... 305
Appendix 5.6 Official Announcement of PL2000 ........................................................................................................ 306
Appendix 5.7 CEFR as used in the planning for PL2000 by MPI staff ........................................................................ 313
Appendix 5.8 Profiles of the Case Study schools ....................................................................................................... 314
Appendix 5.9 Memorandum of Understanding (renewal 2002) ................................................................................ 317
Appendix 5.10 EU White Paper on Education and Training ..................................................................................... 320
Appendix 5.11 Language certification in Lombardy .................................................................................................. 323
Appendix 6.1 The British Institute of Florence ........................................................................................................ 332
Appendix 6.2 Summary of BIF class groups and teachers in FLLGP ......................................................................... 333
Appendix 6.3 Data Summaries and Reports .................................................................................................................. 336
Appendix 6.4 Summary of the 11 FLLGP/BIF class groups ....................................................................................... 338
Appendix 6.5 Summary of responses to Teacher Questionnaire .............................................................................. 342
Appendix 6.6 Participant profile grid - the learners .................................................................................................... 346
Appendix 6.7 Learner Profiles based on teacher logs ................................................................................................. 349
Appendix 6.8 Responses to the LLQ .......................................................................................................................... 371
Appendix 6.9 Note on latent growth and multilevel modelling ................................................................................... 373

Bibliography ............................................................................................................................................................................ 374
LIST OF TABLES

Table:

2.1 Dimensions of purposive action ............................................................... 16
2.2 Test takers ................................................................................................. 35
2.3 Language teachers .................................................................................... 35
2.4 Test developers ......................................................................................... 36
2.5 Teacher trainers and curriculum planners ................................................ 36
2.6 Teacher advisors ....................................................................................... 36
2.7 Head teachers and other school administrators ....................................... 36
2.8 Inspectors ................................................................................................... 36
2.9 Score users ................................................................................................ 36
2.10 Parents ....................................................................................................... 37
2.11 Materials developers ................................................................................. 37
2.12 Teaching processes .................................................................................. 37
2.13 10 year trend in washback/impact model development from 1993 .......... 57
2.14 Spatial/temporal dimensions of washback ............................................... 59
3.1 Summary of flexible designs ..................................................................... 66
3.2 Purposes of an evaluation and possible questions posed by sponsor .... 69
3.3 Characteristics of quantitative and qualitative research ......................... 74
3.4 Mixed method designs (based in Cresswell and Plano Clark 2007) ......... 78
3.5 Case study designs ................................................................................... 83
3.6 Spatial and Temporal variation - reproduced from Gerring ...................... 84
4.1 Phases of IELTS Impact studies 1995 to 2003 ........................................ 92
4.2 Impact Projects 1 to 4 ................................................................................ 93
4.3 Summary of lessons to be learnt from the IELTS case .......................... 106
4.4 People, events, places and actions ............................................................ 115
4.5 Broad categories and groups of stakeholders ......................................... 116
5.1 Overall participations in case studies ...................................................... 133
5.2 Reproduced from Hawkey, PL2000 Impact study, Final report, 2003... 133
5.3 Participation in the pilot visits (March 2002) ......................................... 134
5.4 Reasons for learning English ................................................................. 139
5.5 Activities out of class .............................................................................. 140
5.6 Time spent learning English out of class .............................................. 141
5.7 Candidates for Cambridge Exams at all Levels ................................... 153
5.8 Influences from the milieu – international, national, regional, local ...... 157
5.9 Broad categories and specific individuals .............................................. 164
5.10 Political changes in Italy 1997 to 2007 .................................................... 172
6.1 Summary of data collected: before, during, after .................................. 187
6.2 Degrees of integration of learning and assessment ................................. 217
LIST OF GRAPHS

Graph:

5.1 Number Cambridge exams taken (PL2000)............................................. 152
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALTE</td>
<td>Association of Language Testers in Europe</td>
</tr>
<tr>
<td>ANILS</td>
<td>Associazione Nazionali Insegnanti di Lingue Straniere</td>
</tr>
<tr>
<td>AFL</td>
<td>Assessment for Learning</td>
</tr>
<tr>
<td>BAAL</td>
<td>British Association of Applied Linguistics</td>
</tr>
<tr>
<td>BC</td>
<td>British Council</td>
</tr>
<tr>
<td>BEC</td>
<td>Business English Certificates</td>
</tr>
<tr>
<td>BIF</td>
<td>British Institute of Florence</td>
</tr>
<tr>
<td>BULATS</td>
<td>Business Language Testing Service</td>
</tr>
<tr>
<td>CAE</td>
<td>Certificate in Advanced English</td>
</tr>
<tr>
<td>CBT</td>
<td>Computer-based testing</td>
</tr>
<tr>
<td>CCSE</td>
<td>Certificates in Communicative Skills in English</td>
</tr>
<tr>
<td>CEFR</td>
<td>Common European Framework of Reference</td>
</tr>
<tr>
<td>CELS</td>
<td>Certificates in English Language Skills</td>
</tr>
<tr>
<td>CLA</td>
<td>Communicative Language Ability</td>
</tr>
<tr>
<td>CPE</td>
<td>Certificate of Proficiency in English</td>
</tr>
<tr>
<td>CRT</td>
<td>Centro Risorsa Territoriale</td>
</tr>
<tr>
<td>CRUI</td>
<td>Conference of Rectors of Italian Universities</td>
</tr>
<tr>
<td>CUEFL</td>
<td>Communicative Use of English as a Foreign Language</td>
</tr>
<tr>
<td>EAP</td>
<td>English for Academic Purposes</td>
</tr>
<tr>
<td>ECAD</td>
<td>Evidence-Centred Assessment Design</td>
</tr>
<tr>
<td>EFL</td>
<td>English as a Foreign Language</td>
</tr>
<tr>
<td>ELT</td>
<td>English Language Teaching</td>
</tr>
<tr>
<td>ELTS</td>
<td>English Language Testing Service</td>
</tr>
<tr>
<td>ESOL</td>
<td>English for Speakers of Other Languages</td>
</tr>
<tr>
<td>ESP</td>
<td>English for Specific Purposes</td>
</tr>
<tr>
<td>FCE</td>
<td>First Certificate in English</td>
</tr>
<tr>
<td>GCSE</td>
<td>General Certificate in Education</td>
</tr>
<tr>
<td>GCE</td>
<td>General Certificate of Secondary Education</td>
</tr>
<tr>
<td>IATM</td>
<td>Instrument for the Analysis of Textbook Materials</td>
</tr>
<tr>
<td>IDP</td>
<td>International Development Programme (Australia)</td>
</tr>
<tr>
<td>IELTS</td>
<td>International English Language Testing System</td>
</tr>
<tr>
<td>ILTA</td>
<td>International Language Testing Association</td>
</tr>
<tr>
<td>INVALSI</td>
<td>Istituto Nazionale per la Valutazione del Sistema Educativo di Istruzione e di Formazione</td>
</tr>
<tr>
<td>IRRE</td>
<td>Istituto Regionale di Ricerca Educativa</td>
</tr>
<tr>
<td>IRT</td>
<td>Item Response Theory</td>
</tr>
<tr>
<td>KET</td>
<td>Key English Test</td>
</tr>
<tr>
<td>Acronym</td>
<td>Full Form</td>
</tr>
<tr>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>LEND</td>
<td>Lingue e Nuova Didattica</td>
</tr>
<tr>
<td>LIBS</td>
<td>Local Item Banking System</td>
</tr>
<tr>
<td>LLQs</td>
<td>Language Learning Questionnaires</td>
</tr>
<tr>
<td>LSP</td>
<td>Language for Specific Purposes</td>
</tr>
<tr>
<td>LTRC</td>
<td>Language Testing Research Colloquium</td>
</tr>
<tr>
<td>MPI</td>
<td>Ministero della Pubblica Istruzione</td>
</tr>
<tr>
<td>MS</td>
<td>Main Suite</td>
</tr>
<tr>
<td>NFER</td>
<td>National Foundation for Educational Research</td>
</tr>
<tr>
<td>OCR</td>
<td>Oxford, Cambridge and Royal Society of Arts</td>
</tr>
<tr>
<td>PET</td>
<td>Preliminary English Test</td>
</tr>
<tr>
<td>QCA</td>
<td>Qualifications and Curriculum Authority</td>
</tr>
<tr>
<td>QPT</td>
<td>Quick Placement Test</td>
</tr>
<tr>
<td>RITCME</td>
<td>Recruitment, Induction, Training, Co-ordination, Monitoring, Evaluation</td>
</tr>
<tr>
<td>RSA</td>
<td>Royal Society for Arts</td>
</tr>
<tr>
<td>SEM</td>
<td>Structure Equation Modelling</td>
</tr>
<tr>
<td>SIA</td>
<td>Social Impact Assessment (or Analysis)</td>
</tr>
<tr>
<td>SILT</td>
<td>Studies in Language Testing</td>
</tr>
<tr>
<td>SLA</td>
<td>Second Language Acquisition</td>
</tr>
<tr>
<td>TEFL</td>
<td>Teaching English as a Foreign Language</td>
</tr>
<tr>
<td>TESOL</td>
<td>Teachers of English to Speakers of Other Languages</td>
</tr>
<tr>
<td>TOEFL</td>
<td>Test of English as a Foreign Language</td>
</tr>
<tr>
<td>UCLA</td>
<td>University of California, Los Angeles</td>
</tr>
<tr>
<td>UCLES</td>
<td>University of Cambridge Local Examinations Syndicate</td>
</tr>
<tr>
<td>UODLE</td>
<td>University of Oxford Delegacy of Local Examinations</td>
</tr>
<tr>
<td>USR</td>
<td>Ufficio Scolastico Regionale</td>
</tr>
<tr>
<td>VRIP</td>
<td>Validity, Reliability, Impact, Practicality</td>
</tr>
<tr>
<td>YLE</td>
<td>Young Learners English tests</td>
</tr>
</tbody>
</table>
Chapter 1 Introduction

1.1 Overview

Providers of high-stakes language examinations such as international examinations board have an obligation to investigate the effects and consequence of their examinations. These effects and consequences will be considered in this thesis under the concept of test and examination "impact" (as in the title). To date no comprehensive model of impact has been developed and implemented for language assessment. The aim of this thesis, therefore, is to propose such a model from the perspective of a provider of international language examinations.

More specifically, the data discussed in this thesis and which inform the model were collected in relation to the University of Cambridge ESOL examinations board, which as part of Cambridge Assessment provides public examinations of English in the UK and in about 150 countries around the world.

The stance in this work reflects the author's own interests and responsibilities in developing a model of impact to guide practice within the organisation. The author, as Director of Research and Validation at Cambridge ESOL since the late 1990s, has had responsibility for developing the validation policy for the ESOL examinations and procedures to investigate impact are seen as an integral part of this. His voice as participant, reviewer and developer of the impact model, as well as his relationships with other participants and researchers, is an important feature of this work and its methodological framework.

Starting in the early 1990s a series of projects were carried out to implement an approach to impact which had begun to emerge in Cambridge ESOL at that time. A differentiating feature compared with research being conducted elsewhere was the emphasis on actions and activities which would allow an examination provider to "work for positive impact" and to avoid negative consequences for test users. This thesis employs a meta-analysis to describe and review three of these projects. Based on the analysis, the main outcome and contribution to knowledge is an expanded model of impact designed to provide examination providers with a more effective "theory of action". When applied within Cambridge ESOL it will allow anticipated impacts of the examinations to be monitored and will inform processes of change leading to well-motivated, continuous and iterative improvements in the examinations and the related systems. The model emerging should also be applicable within other institutions which provide public examinations.
The starting point for the thesis is a discussion of examinations within educational processes in general, and in particular the role that examinations boards such as Cambridge ESOL play within educational systems. The historical context of examinations within the British tradition is an important aspect of this discussion and forms part of the review in the first chapter.

The literature review in Chapter 2 covers relevant work in applied linguistics, assessment and education, mainly focusing on a 15 year period up to 2004. Despite the increased discussion and growing use of the term impact when applied to tests and other kinds of educational assessment, the term needs to be carefully defined and distinguished from related terms and concepts. The notion of washback which was developed in the 1990s to take account of changing views of validity in language testing provides a useful basis for building an expanded model of impact. Much of the research in the language testing literature has been small-scale projects and to date no systematic programme has been initiated and carried out by staff within a major examination provider.

Since this thesis aims to focus more broadly on how impact operates within educational systems and in relation to society as a whole, the literature on educational reform and management of change is particularly relevant. The complex interactions between wider contexts and specific learning environments form part of the broader approach to impact and so an understanding of how socio-political change processes work within education is crucial.

With specific reference to educational systems, there are several concepts which emerge from the literature and will be explored in developing the expanded model of impact:

- a definition of stakeholders and the roles they play in many varied contexts or settings where the language learning and assessment operate;
- a view of educational systems as complex and dynamic in which planned innovations are difficult to implement successfully;
- an understanding of how change can be anticipated (if not fully predicted) and how change processes related to assessment systems can be successfully managed through the agency of an examination provider.
- the critical importance of the evidence collected by an examination provider as part of the validation system and as the basis for claims about validity.

These points are developed and interpreted from the perspective of the examination provider and the requirement to establish on-going and iterative processes designed to lead to improvements to examination systems.
The methodology chapter discusses:

a). research paradigms which are appropriate for impact research within educational contexts;
b). the use of meta-data for gaining insights and in-depth understanding and the basis for future impact research.

With regard to a), recent approaches to “real world” research are discussed; the evidence collected within educational contexts needs to include both quantitative and qualitative data and related analyses. The development of suitable instruments, procedures and deployment of resources are all important aspects of this research and relevant to the action-oriented approach to the impact model which is required within an examinations board. The use of case studies as meta-data is discussed with reference to the literature on case studies in social research. Three case studies provide the basis for the meta-analyses conducted within the context of Cambridge ESOL examinations over the period of 15 years under review. References to relevant literature, especially work since 2004, are added as the thesis develops and extensive appendices illustrate the data used in the meta-analysis of each of the three cases.

1.2 The impact of language examinations in educational systems

Impact is relatively new in the field of language assessment and has only fairly recently appeared in the literature as an extension on the notion of washback. Both terms are discussed at greater length in Chapter 2 as part of the literature review. Broadly speaking, impact is the superordinate concept covering the effects and consequences of tests and examinations throughout society, whereas washback is more limited and refers to the influence of tests and examinations in teaching and learning contexts (e.g. classrooms and schools). We shall adhere to this distinction throughout this thesis.

Significant contributions have been made to the language testing literature by PhD studies conducted in the 1990s. For example, the theses and subsequent books by Wall (2005), Cheng (2005), and Green (2007) looked at different aspects of washback and began to focus more broadly on impact issues. But neither in these studies nor elsewhere, has there been a serious attempt to bring all the features of impact together within a comprehensive model which allows for the complex relationships to be examined across broader educational and societal contexts. It is thus important to clarify the concept of a model of impact and to consider its implications when applied to language assessment policy and language testing practices in society. It is also necessary to place this discussion within the educational traditions which have led to current approaches to the teaching and
assessment of languages and to distinguish it from a number of related concepts which appear in the literature with reference to both education and language testing.

Given that the stance in this thesis is that of an examination board located within the British tradition of educational assessment, it is appropriate in this introduction to review the historical influences which have determined the contemporary systems and which continue to exert an influence on approaches to language assessment.

1.3 Examinations within society – the British Tradition

Public examinations have played a major role in educational processes and in British society for over 150 years since the middle of the Victorian era. The interplay between examinations and these processes has been an important consideration since the outset and is a main focus of this thesis.

Under the influence of the utilitarian philosophers Bentham (Constitutional Code, 1827) and Mill (On Liberty, 1859) the idea that it is the responsibility of government to stimulate society to improve itself took root, and against this backdrop examination systems were developed. Bentham worked out an examination system for applicants to the Civil Service and Mill proposed a system of compulsory education based on examinations. Another influential figure was Thomas Babington Macaulay who entered parliament in 1830; he developed the argument that promise in youth is a good predictor of a person’s potential for future tasks, and set out a framework of examinations to recruit Civil Servants (1853). The main driver behind these developments was the view that examinations could open up opportunities and provide fairer access to employment.

At about the same time, in the field of education, exams for teachers were seen as the means of improving the quality of teaching and from 1846 common exams were set simultaneously in all training colleges. Roach (1971) has pointed out that this examination for teachers was the first common test in England based on a general syllabus and taken in a number of separate places. This observation makes two important points about the way examinations have been viewed in the British tradition. First, examinations have been explicitly used to bring about educational change with the intention of improving aspects of society. Secondly, the relationship between the learning objectives (syllabus) and the assessment of achievement (the examination), has been an intrinsic part of this approach since its origins. The stance of Cambridge ESOL described in this thesis reflects this long tradition and still underpins the assessment systems which are in place today.

The setting up of the examinations boards in Oxford and Cambridge happened at around the same time. In June 1857, the University of Oxford Delegacy of Local Examinations
(UODLE) was established following suggestions by school representatives from Devon. They were seeking external assessments for schools in Exeter, with Oxford University conducting exams for pupils in their “local” school contexts. This suited the University’s aims at that time and was part of a movement to reform the universities and make them more socially involved. Cambridge University received a similar request from schools in Birmingham, Cheltenham, Leeds and Liverpool to offer “local exams” and the Council of Senate recommended that a “Syndicate” be set up. The University of Cambridge Local Examinations Syndicate (UCLES) was established in February 1858 and its first examinations were held in December that year. Both examinations boards continued to provide external examinations for schools until the mid-1990s when government policy required a reduction in the number of Boards offering school examinations; this saw the merging of the Oxford and Cambridge boards with the Royal Society for encouragement of Arts (RSA) board to form OCR. OCR is now one of the three unitary awarding bodies which offer examinations in English schools under regulation by the Qualifications and Curriculum Authority (QCA) (see Raban, 2008). The trend towards centralisation and increased regulation is taken up in Chapter 2 and in subsequent discussion as it is a major factor in determining the influence of assessment on schools and teaching in the classroom.

Given these origins and influences, it is not surprising that school examinations in England have continued to involve teachers and other “local stakeholders” in the assessment processes; there has been an accepted view that assessments should be relevant to the teaching/learning contexts and should also meet societal demands for accountability, including maintenance of standards of achievement and fairness in terms of impartiality. Against this backdrop the thesis will focus on the concept of test impact within contemporary educational systems taking into account the role that examinations boards and assessment systems now play.

1.4 Locating impact within educational systems

It has been suggested that educational processes take place within complex dynamic systems (see Fullan, 1993, 1999; Thelen & Smith, 1994, Van Geert, 2007) with interplay between many sub-systems and “cultures” and where understanding the roles of stakeholders as participants is a critical factor. This thesis, therefore, situates the current discussion of impact within the work of researchers who focus on how change can be managed successfully within educational systems. The importance of evidence on which to base analyses and to support claims about the nature of impact is seen as central to this approach.

In Figure 1.1 Saville sets out a representation of macro and micro contexts within society; it shows how diversity and variation between contexts tend to increase as the focus moves
from the general, macro context at national level to the multiple micro contexts within the educational system at the local level (i.e. schools, classes, groups, individual teachers and learners).

Understanding the nature of context within educational systems and the roles of stakeholders in those contexts are clearly important considerations for an examinations board. This is taken up in Chapter 2 and Figure 1.1 is referred to again in subsequent chapters.

Saville (2003:60) suggests that the extensive “taxonomy of stakeholders” working with Cambridge ESOL places demands on the board to ensure that it can “review and change what it does in the light of findings on how the stakeholders use the exams and what they think about them”. This requirement provides a primary justification for impact studies being part of the test validation process.

In education, there is a growing literature focusing on the effects of planned interventions, including teaching and assessment programmes, on the people participating in them (the stakeholders referred to above). These studies generally seek to measure and analyse the processes of educational change, and in considering the impact of language assessment the nature of these processes needs to be taken into account.

1.5 Locating impact research within Cambridge ESOL
The study of impact has featured on Cambridge ESOL’s research agenda since the early 1990s and findings from impact research now form part of the body of evidence to support
the claims to the validity of its tests and examinations. However, a major consideration for Cambridge ESOL and a fundamental concern in this thesis is how the study of impact can be integrated into operational processes in a more systematic way. This issue is resolved in Chapter 7, the concluding chapter, where there is a discussion of what has been learnt from the three case studies and where their outcomes are used in redefining and refining an expanded and more comprehensive model of impact.

In order to be integrated into Cambridge ESOL’s assessment procedures, impact research needs to be conceptualised in such a way that it combines theoretical substance with practical applications. In other words, impact and impact research need to be seen as an integral part of the operational test development and validation processes alongside other essential considerations. To date, such an approach has not been adopted by test developers and, as we shall see in Chapter 2, most washback studies in the literature have been small-scale and conducted post-hoc.

In placing impact within a validation framework, we acknowledge the work of Bachman who was one of the first to discuss impact as a “quality” of a test or examination. Bachman’s seminars delivered in Cambridge in the early 1990s had a significant influence on Cambridge ESOL at that time (two summer schools in 1991 and 1992). He suggested that impact should be considered within the overarching concept of test usefulness and that the development of “useful tests” should involve the balancing of four test qualities: validity, reliability, impact and practicality - the VRIP features as they became known in Cambridge. This thesis builds on this approach, as Saville (2003: 61) points out: “Individual examination qualities cannot be evaluated independently. Rather the relative importance of the qualities must be determined in order to maximise the overall usefulness of the examination”. Bachman himself was influenced by Messick’s work on validity theory in the 1980’s and the view that validity is a unitary concept. This is also a fundamental concept which underpins the current work and is discussed in Chapter 2.

In an internal working paper in 1996, Milanovic and Saville first set out their ideas on an expanded concept of test impact to meet the needs of Cambridge ESOL. Their paper set out the question of how examinations can be developed with appropriate systems in place to monitor and evaluate their impact. Practical guidance on how to put in place operational working practices was required. Aware of the work of Hughes (1989) and others (e.g. Bailey, 1996) who used checklists of behaviours to encourage positive washback, Milanovic and Saville proposed four maxims – outlined below.

Saville had worked on the concept of practicality in language testing and how practicality relates to validity and reliability (see Saville, 1990; unpublished dissertation, University of
Reading. This work represented early stage thinking about ways in which impact could be addressed by an examination board and how the effects of a test, once “installed” within its intended contexts of use, could be monitored. Saville used the metaphor of the test creating “ripples” which flow out into the educational system and society at large. These ripples, he suggested, produce both intended and unintended effects.

Figure 1.2 Impact “ripples”

As educational systems are complex with many inter-related features, the ripples can be far reaching as Milanovic and Saville (1996) point out, recognising impact as well as washback: “… because of the social and cultural implications, it is clear that impact relates to more than classroom practices. It concerns attitudes, motivations, standards, and qualifications, as well as issues such as fairness, discrimination and equal opportunity for all”.

The four maxims proposed by Milanovic and Saville were used to support working practices and were set out as follows:

Maxim 1 PLAN
Use a rational and explicit approach to test development
Maxim 2 SUPPORT
Support stakeholders in the testing process
Maxim 3 COMMUNICATE
Provide comprehensive, useful and transparent information
Maxim 4 MONITOR and EVALUATE
Collect all relevant data and analyse as required
The statements were deliberately designed to be short and memorable, to capture the key principles and what is most relevant, and in so doing to provide a basis for decision-making and action planning. Under Maxim 1 there was a requirement to plan effectively and for the organisation to adopt a rational and explicit model for managing the test development processes in a cyclical and iterative way. Maxim 2 focused on the requirement to provide adequate support for the stakeholders involved in the many processes associated with international examinations, i.e. all those who, together with the examination board itself, have a stake in the examination, including the test takers and the users of the test results such as parents, schools, employers etc. Maxim 3 focused on the importance of communication and of providing useful and transparent information to the stakeholders and Maxim 4 on the requirement to collect relevant data and to carry out analyses as part of the iterative process model.

This approach represented a new development in working practices at that time. By conceptualising impact within VRIP-based validation processes, there was an explicit attempt to integrate impact research into on-going procedures for accumulating validity evidence. The Cambridge perspective on impact is framed by these considerations; they provide the starting point for the model developed in this thesis and a context for the three case studies which are described and analysed in detail in Chapters 4, 5 and 6 below.

1.6 Locating impact within a socio-cognitive model of test validity

While advances in linguistic theory and second language acquisition (SLA) have provided useful insights for language test developers, there is still no comprehensive theory of language ability and how it is acquired. This means that the definition of the constructs in language assessment is particularly challenging. Yet the nature of these constructs and how they relate to the learning/teaching context has been a long-standing consideration in washback research, (discussed in Chapter 2: see Green, 2003).

Cambridge ESOL’s approach to validation has attempted to introduce explicit reference to construct definition covering both social and cognitive dimensions. This approach draws on the work of Weir (2005) in that it outlines a socio-cognitive framework to produce theory-based evidence of test validity. A key feature is the recognition that the validity of an assessment procedure will depend crucially on the interaction between the characteristics of intended test takers and the characteristics of the test, i.e. the specific tasks and events which elicit a response from the test taker within a testing system.

The work of Messick (op cit) has also been influential in developing this approach. His ideas underlie recent literature in educational assessment, including important work by Kane from
1992 onward. Bachman has also extended Kane’s work to support a “case for test utilization” (e.g. paper presented at LTRC 2004) and Mislevy et al (2003) have recently described the structure of educational assessments based on the concept of evidence-centred assessment design, sometimes abbreviated to ECD or ECAD. (See also Pellegrino and Chudowsky, 2003).

In summary, a socio-cognitive approach depends on the following two points:

a). The abilities to be tested are mental constructs within the brain of the test taker (the cognitive dimension). In the case of language ability, the effective use of language for communication also depends on other mental constructs, such as declarative knowledge, and other cognitive features (strategic, meta-cognitive, etc.).

b). The use of language is a social phenomenon. This suggests a need to design the test tasks and other aspects of the test taking context so that inferences about the test taker’s ability to use language for communication may be adequately evidenced.

The interpretation and use of test results are also social phenomena so that in addition to the argument for the validity of a testing instrument or system itself (its technical qualities), the socio-cognitive approach also builds in a component for test use. This is related to consequential aspects of validity and test impact and is consistent with Messick’s “progressive matrix” (1989). Figure 1.3 summarises the components of the socio-cognitive approach in use by Cambridge ESOL (Saville, 2004). In this framework the central “triangle” represents the three components which are essential for construct validity – i.e. the specification of language abilities being assessed, the test tasks and the scale used as the basis for score interpretation.

Figure 1.3 Socio-cognitive approach based on Saville, 2004
The test taker is represented in the framework as an individual sharing features in common with other people in the same test taking context (e.g. age range, gender, profile of prior learning, etc.) and sharing similar goals in using the test results, (e.g. for future learning or other societal goals). Each test taker will also be differentiated by individual characteristics, including knowledge and skills profiles and social, cultural, cognitive and meta-cognitive features (e.g. attitude, motivation, strategy use etc).

A theory of cognitive processes and abilities which constitute the principal object of assessment is central to this framework, i.e. a definition of the focal constructs and the processes of the brain which are engaged when carrying out language-related skills such as reading, writing, listening, speaking (cf. Bachman’s 1990 model of communicative language ability - CLA). The tasks in the test should engage the candidate's abilities and skills so that in their externalised form, the test taker’s responses provide the basis for the scoring of those abilities. The social dimension of the framework is represented by the test taking context and the contexts of test use. “Situational authenticity” (Widdowson, 1989) is achieved by ensuring that the test-taking context recreates conditions similar to those in the target language use domain (see discussion in Chapter 6).

The candidates’ responses have to be reliably scored and transformed into a scale of some kind to be interpretable and to have meaning for test users. Apart from the accuracy of scoring, consistency and stability over time also need to be demonstrated (e.g. through the application of a measurement model such as the Rasch model). The test result (score/grade) attributed to the test taker can be used in various societal contexts and the overall validity (or usefulness) is determined by the validity of the inferences and the fairness of the decisions which are made using the test results. This is consistent with Messick (1989: 13) and is discussed at greater length in Chapter 2.

1.7 Locating impact in the test development and validation cycle

In Messick’s approach cited above, the consequential aspect of validity is a key element of his “progressive matrix” in which “the social values and the social consequences [of a test] cannot be ignored in considerations of validity”. Bachman and Palmer (1996: 22) also emphasize this when they define validation as “the on-going process of demonstrating that a particular interpretation of test scores is justified, and [validation] involves, essentially, building a logical case in support of a particular interpretation and providing evidence justifying that interpretation”.

With this in mind, test impact research should entail the collection of adequate data to investigate how and to what extent the test affects the stakeholders, the processes and systems in the contexts in which the test has been designed to play a useful role. The
nature of the test construct itself and the ways in which the scores are interpreted in society are clearly important aspects of the validation process; in order to address these and other features effectively the study of impact needs to be embedded within operational processes (as noted above). From a practical point of view, those involved in test development and validation within an examinations board need to think carefully about the nature of the data which will be necessary to substantiate claims of validity and at what points in the process adequate data can be collected. Impact research therefore is akin to other kinds of routine validation work.

1.8 Methodology – using case studies as meta-data
A range of data collection and analysis techniques needs to be employed in impact-related research depending on the aspects of impact which are focused on. These are discussed in Chapter 3 with reference to the literature on social research, and particular reference is made to the use of case studies as meta-data. Ways in which both quantitative and qualitative approaches can be effectively combined are noted and the design, validation and application of instruments are illustrated in Chapters 4, 5, and 6. The three case studies which form the central part of the thesis are as follows:

Case 1 is a survey of the impact of IELTS (the International language Testing System). It can be considered the starting point for this work, and the original model for what has followed. It sets out the conceptualisation of impact and describes the design and validation of suitable instruments to investigate it, as applied within four Impact Projects as part of an ongoing programme of validation following a major revision. This case includes a description of the IELTS development and the underlying constructs, the nature of the impact data which were targeted and the necessary instrumentation to collect those data. The lessons learned are summarised in relation to the developing model and how they informed the next phase of development summarised in Case 2.

Case 2 is the Italian Progetto Lingue 2000 (PL2000) Impact Study. The impact study is an application of the original model within a macro educational context and describes an initial attempt at applying the approach within a state educational context. This context is the Italian state system of education and a government reform project intended to improve standards of language education at the turn of the 21st century, i.e. Progetto Lingue 2000. The impact of the reforms generally and the specific role of external examinations provided by Cambridge ESOL form the basis of this case. The PL2000 Impact Study provides a greater focus on the contextual variables and the roles and responsibilities of particular stakeholder groups and individuals within the educational system.
Case 3 is the Florence Learning Gains Project (FLLGP). Still within Italy, this project built directly on the PL2000 case and is an extension and re-application of the model within a single school context (i.e. at the micro level). It focuses on individual stakeholders in one language teaching institution, namely teachers and learners preparing for a range of English language examinations at a prestigious language school in Florence. The complex relationships between assessment and learning/teaching in a number of language classrooms, including the influence of the Cambridge examinations, are examined against the wider educational and societal milieu in Italy. The micro level of detail, as well as the longitudinal nature of the project conducted over an academic year, are particularly relevant in this case.

Given the focus on human participants and their behaviours in real world contexts, impact research needs to take into account ethical and legal considerations which pertain to the collection, storage and uses of such data (Fetterman, 1989; Miles & Huberman, 1994; Punch, 1986, 1994). The collection and storage of data as part of the research reported in this thesis conform to best practice ethical guidelines (e.g. issued by BAAL) and to data protection legislation in the UK.

1.9 Outcomes of the thesis and contribution to knowledge

The final chapter, Chapter 7, asks what can be learned from the three case studies and how the outcomes of the analysis can contribute to an expanded model of impact. Four main features of a meta-framework are proposed in answer to this question.

The outcome and the main contribution to knowledge in the thesis, therefore, will be the expanded model designed specifically to help Cambridge ESOL and other examination providers to address the challenge of finding out and understanding how their examinations impact on society. Concrete and relevant applications for investigating the impact of language assessment at micro and macro levels within the routine work of the examinations board are suggested.

The integrated nature of the approach, both in terms of the relationship between the macro and micro elements and the situating of the research within test validation more generally, is new. Unlike much research reported in the literature to date which has been conducted post hoc by external researchers, the expanded model “builds in” impact considerations from the start, and seeks to anticipate potential effects and consequences with a commitment to monitoring and changing things as required. The suggested approach, referred to as “impact by design”, is thought to be particularly relevant for future impact research conducted by examination providers as part of their ongoing test development and validation processes.
Finally, a proposal is made suggesting ways in which the expanded model might be applied within other context. Limitations of the research are summarised and final conclusions are drawn.

In the next chapter the relevant literature is reviewed and the research questions are posed.
Chapter 2  Literature Review

2.1  Definition of the term impact

The word “impact” in modern English carries a number of literal meanings and connotations which are important in understanding why this word is used in relation to assessment.

The noun impact is typically defined as an effect on something, often with the implication that the effect is strong and with force behind it. Other common words associated with impact in a thesaurus include, influence, consequence, impression, repercussion and ramification. The use of impact as a verb has also become common in recent years, as in to impact (on) something and it is used with the meaning to have an effect on someone or something. Again there is a connotation of force, as in to strike or to hit (see for example the definitions and examples in Collins Essential English Dictionary, 2nd Edition 2006).

These meanings explain why the term impact has come to be widely used in social contexts where it refers to the influences, effects and consequences of social policies and the associated practices. This general usage has been adopted in this thesis and the more precise meanings when applied to language assessment are taken up below and discussed in subsequent chapters.

2.1.1  Influences, effects, and consequences of policies in society

Sociologists have discussed the effects and consequences of social action within society for many years, with a particular focus on whether desirable and planned effects are achieved or whether unplanned or undesirable effects occur in practice as a result of social policies. The seminal paper by the sociologist Robert K. Merton in 1936 has been particularly influential and is relevant to the discussion of effects and consequences of assessment.

In his paper Merton applied a systematic analysis to the problem of "unanticipated consequences of purposive social action" (Merton, 1936: 897-904). He emphasised that planned ("purposive") action is concerned with human "conduct" as distinct from "behaviour", that is, "with action which involves motives and consequently a choice between various alternatives". (1936: 895).

Actions, he claimed, have more than one effect, and these effects invariably include at least one unforeseen side effect. In some cases, the unintended side effect can be
more significant than any of the intended effects. His ideas have been widely applied to many contexts where planned activity takes place and Merton’s analysis has become known as the law of unintended consequences. Merton noted possible causes of unanticipated consequences - what he referred to as “limiting factors”. These include “human ignorance” and “limitations of knowledge”, “error”, and the “immediacy of interests”. The last of these he refers to as “… instances where the actor’s paramount concern with the foreseen immediate consequences excludes the consideration of further or other consequences of the same act. The most prominent elements in such immediacy of interest may range from physiological needs to basic cultural values”. (1936: 900).

While not actually a scientific law, Merton’s maxim reminds us of the unpredictability of social action and of the problem of causal imputation, i.e. of ascertaining “the extent to which ‘consequences’ may justifiably be attributed to certain actions” (Merton, 1936: 897). Merton pointed out that the attribution of effects and consequences to specific causes can be compounded by problems of “ascertaining the actual purposes of a given action”. Where uncertainty of purpose exists, it can lead to post facto rationalisations and justifications for what has occurred, as Merton states “there is the difficulty … of discriminating between rationalization and truth in those cases where apparently unintended consequences are post facto declared to have been intended”. (1936: 897).

In discussing effects and consequences of tests and examinations it is important to bear in mind Merton’s analysis; his work reminds us that our understanding of assessment policies needs to be developed within a broader understanding of societal processes and that we need to consider carefully how causes of particular effects, planned or otherwise, might be attributed.

Table 2.1 shows the interfaces between two dimensions of purposive action (e.g. plans, policies, laws etc.): planned versus unplanned outcomes, and positive versus negative effects and consequences.

<table>
<thead>
<tr>
<th>Effects and consequences</th>
<th>Planned, Intended outcome</th>
<th>Unplanned and unintended outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive effect and consequences</td>
<td>The policy goal</td>
<td>An unexpected beneficial effect <em>(windfall)</em></td>
</tr>
<tr>
<td>Neutral</td>
<td></td>
<td>A problem to be resolved</td>
</tr>
<tr>
<td>Negative effect and consequences</td>
<td>Not applicable?</td>
<td>An unexpected adverse or perverse effect</td>
</tr>
</tbody>
</table>
The unintended consequences which occur as a result of purposive actions typically fall into three categories:

- **positive**, unplanned benefits;
- **negative or perverse** effects;
- **problems** which arise as a result of error or lack of prior knowledge.

The second point is particularly relevant to the discussion of impact as it concerns the circumstances in which negative consequences result from a specific policy which encourages people to take actions contrary to what was intended.

An additional dimension is that of unpredictability. Social planners may be aware of possible unintended effects which pose a potential threat or risk to the policy goal. On the other hand, some effects may be unknown to the planner in advance or may not be predictable based on the current state of knowledge. Merton comments that: “the most obvious limitation to a correct anticipation of consequences of action is provided by the existing state of knowledge” and that the importance of ignorance as a factor “is enhanced by the fact that the exigencies of practical life frequently compel us to act with some confidence, even though it is manifest that the information on which we base our action is not complete.” (1936: 898-900).

These points highlight the importance of human agency in determining the effects and consequences of social policies generally, and this also applies in assessment policy. In addition to human conduct in developing and carrying out policies, the unpredictability of human behaviour plays an important role in societal systems. The actions of individuals in responding to policies are key factors in determining whether the planned activities turn out as intended. It has been noted that educational systems are particularly prone to unintended effects of government policies; so, for example, Fullan (1993) reminds us that change in education “roams somewhere between over control and chaos” (quoting Pascal, 1990), and notes that “controlling strategies” don’t work because the change process is “uncontrollably complex and in many cases unknowable”.

The unintended and/or unexpected effects and consequences of educational policies are particularly problematic if they turn out to be “perverse”, that is contrary to the stated objectives. This can occur when a policy creates perverse incentives which lead to the perverse outcomes; this notion is discussed below in relation to high-stakes testing and Goodhart’s Law (Goodhart, 1975).
2.1.2 Impact: educational policy and the role of examinations

Social policies are always developed within the social contexts in which they are intended to operate, and all policies are influenced by the values and beliefs of those who have the responsibility to draw them up and to implement them. This is particularly true of educational policies and assessment systems (see Broadfoot, 1996).

In this section the development of examinations within formal educational systems is briefly reviewed and recent literature focusing on the social dimension of language assessment is covered. The interaction between policy decisions and the use of language tests is also discussed (see McNamara and Roever, 2006).

Education has been a tool of governments in bringing about planned reforms since the Victorian era. The influence of the 19th century philosophers referred to in Chapter 1 and of social thinkers such as Marx, Durkheim, Weber and Bernstein has been particularly significant over the past 100 years. In England in the 1800s the first industrialised society and its associated bureaucracies were emerging and increased social control and accountability mechanisms became important social phenomena (Durkheim, 1947; 1961). Supporting the reform of society through education became particularly important and public examinations played a role for the first time.

In discussing the social role of assessment, Broadfoot (1996: 25) points to the “sociological rationale for the emergence of assessment procedures as part of the development of formal educational provision”. She highlights the role of examinations in regulating and controlling competition, individuals and systems dating back to the origins of formal education. So, for example, examinations were seen by the Victorians as a way of providing access to opportunity for talented individuals who were prepared to put in effort to better themselves i.e. the concept of ‘merit’. However, as we saw in Chapter 1, examinations also helped to regulate competition and sought to introduce greater fairness into systems, such as the selection processes for the universities, the Civil Service, and the professions. Examinations in this context put a check on patronage and corruption, although they could not completely “eliminate the advantages that derived from social status and wealth” at that time (Eckstein and Noah, 1993: 11). The concern for social justice and the raising of educational standards through the external monitoring of pupils and schools saw the setting up of the first “local examinations” boards (Broadfoot, 1996: 168-216; Raban, 2008; Roach, 1971).
From the start, the university-based examinations boards in England, not only conducted the assessments, but also exerted a powerful influence on the content of the curriculum which was taught within the schools. This is still the case today (e.g. with regard to choice of syllabus for GCSE and A level examinations), although the introduction of the National Curriculum in 1988 has led to greater centralisation and has put more control over the content of teaching in the hands of central government. The same education act also led to national testing at the Key Stages and the use of league tables to evaluate the performance of schools which is discussed below (see Mansell, 2007).

The close relationship between assessment and the content of learning was an important consideration when UCLES (now known as Cambridge Assessment) started to administer English language examinations in 1913. Weir (in Milanovic and Weir, 2003) refers to the influences of Jack Roach, the Assistant Secretary in Cambridge who was largely responsible for the development of the Cambridge ESOL examinations until the 1940s. Roach noted that “examinations act as a stimulus and a focusing point for both teachers and the taught, and thereby promote the expansion of the studies they are designed to test.” (Roach, J O, 1944: 36).

The assumption that public examinations can define a teaching syllabus and help to determine learning objectives (learning-led assessment) has remained a critical factor in the development and revision the ESOL examinations and is relevant to the model of impact being developed in this thesis. (See Hawkey, 2009).

2.1.3 Concerns for the social dimensions of language assessment
Concerns for social consequences, ethicality and accountability have re-emerged as important themes in educational assessment in the early 21st century, and these concerns are now seen as particularly relevant to language assessment (Hamp-Lyons, 1997).

Spolsky was one of the first scholars to address the use of language tests for high-stakes purposes where people’s lives are significantly impacted by the decisions made using the results of the test (Spolsky 1981, 1997). In his 1981 paper, he discussed a number of ethical concerns in contexts of test use related to “gatekeeping” for migration and citizenship. He urged that test results should be used with great care, and from an ethical point of view, he pointed out that it is important to understand the effects of a test on the test taker, on the instructional processes, and on other facets of the social context in which the test operates. Spolsky (1995: 354-357) has more recently referred to “competing sets of forces” including: the
institutional context of a test, the need for general acceptance, and the “unavoidable uncertainty” or “probable error” (see also Hamp-Lyons, 2000: 587).

More generally, the potential for assessment to function as a mechanism of control in society has been highlighted by post-modern philosophers and academics. The deliberate attempt by governments to influence aspects of education and/or society through assessment policies and the use of particular tests or examinations has been seen by some to be a necessarily negative impact of assessment. The work of Foucault (1975/1979), has been referred to in support of this position; for example, Foucault sees the use of educational examinations as a mechanism in the operation of power in society (both overt and covert) with a significant disciplinary function (e.g. to judge, to normalise, to quantify, to classify and to punish). In Foucault's critique, tests and examinations help “construct the social identities” of the individuals who are also “policied” through the same assessment mechanisms. Foucault is quoted by McNamara and Roever (2006: 191) to illustrate the relationship between language examinations, power and identity.

This is a “social critique” which many educationalist do not subscribe to, but nevertheless this view has been influential and has led to a questioning of the “power of tests” and of the influence which they can exert on educational processes and within society generally (Bourdieu, 1991; Shohamy, 1993, 2001, 2006, 2007).

The influence of post-modern thinking also saw the development of a “critical language testing” movement in the 1990s. McNamara (1997), taking a lead from Pennycook (1994) and Kramsch (1993), suggests ways in which critical thinking might influence future directions in language testing, including making language testers more self-reflective and socially aware and by introducing Foucauldian thinking into their critiques of language testing practices. He also suggests that this might lead to alternative types of testing being developed. Davies (1997: 328), in focusing of professional practices, agrees with this when he notes that critical approaches to language testing can "expose the importance of carefully examining alternative assessment proposals and making clear the validity of assessment methods used by the profession." (See discussion of formative assessment below and in following chapters).

The effect of government policy and the use of high-stakes standardised tests (e.g. as researched by Kellaghan et al, 1982) is discussed in more detail later in this chapter. We will also return to this question in Chapters 5 and 6 in discussing the interface between micro and macro levels of impact.
2.2 Impact and washback in the literature

2.2.1 Some definitions

The terms impact and washback were introduced in Chapter 1. The concept of impact has also been defined above as the influences, effects and consequences which tests and other assessment procedures have on education and various aspects of society at large. What these effects are called, who or what is affected, and whether or not they are intentional or accidental, positive or negative are key considerations which need to be addressed in this thesis.

The related term washback, often used in free variation with backwash, can be found in literature on educational assessment, applied linguistics and language testing dating back to the middle of the 20th century (see Alderson & Wall, 1993; Bachman, 1990; Broadfoot, 1996; Davies et al, 1999; Gipps 1994; Hughes 2003; Pearson 1988; Vernon, 1956). The meaning derives from the literal meaning of the word backwash; the backward flow or movement of something on to something else, e.g. sea water on to the beach. In the assessment literature and in language testing it is used figuratively and is understood as the “backward flow” of effects from testing on to teaching and learning (and as such is one aspect of the broader phenomenon of impact). Andrews (1994) notes that in the general educational literature backwash has been the preferred terms, whereas in language education there is a preference for washback (see also Alderson, 2004). In this thesis washback will be used.

Alderson and Wall (1993: 121) in their influential paper on washback, discussed at length below, ask whether washback should be taken seriously or if it simply provides a “metaphor which is useful in that it encourages us to explore the role of tests in learning and the relationship between teaching and testing”. The attention which has been given to the concept since they posed this question has confirmed that washback and washback-related research certainly need to be taken seriously. It is now accepted that washback studies help us to understand the role that language tests play in relation to factors such as the individual learner, the teacher's attitudes and behaviour, the classroom environment, the choice and use of teaching/learning materials, etc.

2.2.2 The washback continuum - positive, neutral, negative?

The term washback is neutral, being related to the influence and effects of the test or examination which might be beneficial or harmful. Alderson and Wall (1993) reflect this view – i.e. that language tests can have both positive and negative effects. In other words, the effects of tests on teaching and learning lie somewhere on a
continuum, stretching from negative (harmful) at one end, through neutral to positive (beneficial) at the other end. Harmful effects might occur when test content or format is based on a narrow definition of language ability and so tends to constrain teaching and learning too narrowly (see below: “teaching to the test”). Beneficial washback might occur when the assessment procedures encourage ‘good practice’ (however that is defined) and lead to positive learning outcomes. Whether only “good tests” can have positive washback, or “bad tests” have only negative washback has remained an important part of the washback debate (see Messick, 1996 for example).

The use of washback as a neutral term is reflected in the work of other academics, for example, Bachman & Palmer, 1996; Buck 1988; Davies et al., 1999; Hughes, 1989; 2003. They acknowledge that washback can be planned (intentional) and have positive effects, or unintentional with effects that might be positive or negative.

The idea that language tests can be used explicitly within a policy to bring about positive effects in classroom contexts began to be discussed in the literature during the 1980s. Hughes (1989) was one of the first language testers to focus on this in his introductory text on testing for language teachers. In this volume, he saw the relationship between teaching and testing as a key element which he discussed under the heading of “backwash” (his preferred term). An entire chapter (1989; 44-46) is devoted to achieving positive backwash and Hughes suggests that test developers should follow seven wide-ranging “imperatives” intended to “promote this beneficial effect”:

- Test the abilities whose development you want to encourage.
- Sample widely and unpredictably.
- Use direct testing.
- Make testing criterion-referenced.
- Base achievement tests on objectives.
- Ensure test is known and understood by students and teachers.
- Where necessary provide assistance to teachers.

This list was perhaps the first attempt “to model” the washback concept and to provide language testers with a guide for action. His bullet points cover various aspects of the test design and construction, as well as support for the test users. It seems likely that Hughes based these recommendations on his experience in test
development projects in various parts of the world (e.g. Turkey), but there was no explicitly stated theory to support these points.

In the 1980s several other authors suggested similar approaches, largely based on practical experience and what seemed to work well for practitioners. For example, Heyneman and Ransom (1990) recommended three strategies for specifically improving test content as a way of generating positive washback effects: use more open-ended items; test higher-level cognitive skills; provide feedback to teachers and others so meaningful change can be effected. See also Swain, (1985: 42-4) who recommends that test developers should "work for washback" and "bias for best".

While Hughes was influential in raising awareness of positive washback in language testing, some writers warned about negative effects of tests, sometimes as the direct result of assessment policies (e.g. Davies, 1968; Kellaghan et al 1982; Alderson, 1986). As Spolsky (1981) pointed out, negative effects are a particular danger with certain uses of tests (e.g. for migration purposes) or where certain test formats are prevalent. For example, tests using discrete-point, multiple choice items may narrow the curriculum and encourage too much practicing for the test (Buck, 1988).

Some researchers question whether washback might be independent of the quality of the test and more to do with the uses and stakes associated with it. In other words, the washback effect might be caused by the testing regime rather than the particular features of a test. This has been discussed in the general educational literature and a number of related concepts emerged highlighting potentially negative effects of assessment; nowadays these are also frequently referred to in the language testing literature and include:

- teaching to the test (Madaus, 1988; Wiliam, 1996; 2008);
- measurement-driven instruction (Airasian, 1988; Bracey, 1987; Popham, 1987; Ramirez, 1999);
- curriculum alignment or "distortion" (Resnick & Resnick 1992; Shepard ,1990,1993; Smith, 1991; Vernon, 1956).

For these writers, it is the effects of the assessment policy which lead to undesirable consequences in schools (although this might be compounded by "bad tests" which themselves derive from the policy decisions or may be created by invalid practices on the part of the examination provider). Curriculum alignment, for example, refers to the matching of the curriculum to test content by teachers who are overly influenced by the test, usually because of the high-stakes associated
with passing or failing the test. This convergence is claimed to be negative as there tends to be a narrowing of the curriculum to match the narrower coverage of the test. This is informally referred to as “teaching to the test” or “the tail wagging the dog” (Green, 2003) where the powerful influence of the test constrains what goes on in classrooms in unhelpful and undesirable ways. This view, however, assumes that the curriculum and the test are not based on the same knowledge constructs and educational objectives. As we will see below, this is a key consideration in determining the effect of this kind of alignment and the phenomenon need not necessarily be undesirable or have negative impacts (see discussion below of test validity and construct-irrelevant variance and construct under-representation).

A test may also affect individual people in various ways – as noted above in our discussion of Merton; the attitudes, perceptions and behaviours of teachers and learners may be affected in ways which are thought to be negative, e.g. by reinforcing unproductive learning practices or attitudes to learning, increasing anxiety, etc.. This point is taken up below in a discussion of stakeholders and participants.

### 2.2.3 Impact in language assessment

It is now widely recognised in the language testing literature that language tests have consequences beyond the teaching/learning context. In 2000, Hamp-Lyons (2000: 586) claimed that there is “a growing awareness by language testers that the societies in which and for which we work are, whether we approve or not, using tests as their levers for social and educational reform”. (See also the work of Shohamy referred to above).

Language tests often carry high-stakes and significantly affect the life chances of individual test takers through their use in controlling access to international education and employment opportunities. Language tests also affect the educational and social systems in which they play a part, e.g. when language test results are used to make important decisions about curriculum planning or funding allocation for schools, immigration policy or the licensing of health professionals such as doctors and nurses. There may also be economic and commercial interests related to language assessment especially when they carry high-stakes. If there is a viable market, publishers produce preparation materials and private institutions set up preparation courses and other businesses linked to assessment.

It is these wider effects and consequences within society which we refer to as language test impact. This impact operates broadly on at least two levels: on a
socio-cultural level, within educational systems and society in general; and on a local and personal level in terms of the people who are directly affected by tests and their results, especially in school contexts. The division of impact along these lines is consistent with the distinction made by Bachman and Palmer (1996) who refer to "macro" and "micro" levels in their discussion of impact and test usefulness (noted in Chapter 1; see also Wall, 1997, 2005, and discussion of Bachman’s work below).

Wall (1997: 29) agrees with this distinction when she refers to impact as “any of the effects that a test may have on individuals, policies or practices, within the classroom, the school, the educational system or society as a whole”. Similarly, Hamp-Lyons (1997) defines impact as that which pertains to high-stakes tests whose influence extends to the school, educational systems, and society.

This distinction is useful in conceptualising the notion of impact, but it does not, of course, imply that the levels are unconnected. On the contrary, the potentially complex relationships between individuals, the institutions to which they belong, and the broader systems in society are clearly of crucial importance in reaching an understanding of what impact is and how it works. The dynamics between the micro and macro contexts mean that this is likely to be a complex rather than a simple relationship. In other words, these phenomena operate within complex systems which have “dynamic” properties, a point which will be discussed below. Larsen-Freeman, (1997) discusses the relevance of chaos theory (Gleick, 1987) to linguistic research, and Larsen-Freeman and Cameron (2008) discuss the notion of complex systems and the dynamics of change processes in applied linguistics.

The complex interactions between the factors which make up the teaching/learning context (including the individual learner, the teacher, the classroom environment, the choice and use of materials, etc.) make it difficult to attribute outcomes to any one factor alone such as a test. The extent to which a test interacts with or is conditioned by factors beyond the classroom is a feature of the test’s wider impact on society and educational systems.

In summary, language test impact concerns the effects a test or assessment procedure on various systems in a range of contexts. Because of the social and cultural implications, the scope of impact research includes the effects of a test on people’s attitudes and motivations, their views of educational standards and choice of suitable qualifications, as well as ethical issues such as fairness, discrimination and equal opportunity.
2.3 Washback - hypotheses and models

2.3.1. Does Washback Exist?

Alderson and Wall's paper, Does Washback Exist (1993) is now seen as a seminal work on washback; they suggest that while washback provides a useful metaphor, one of the problems when looking at real situations is that it becomes difficult to establish whether it is the test itself which has the influence or other factors from the educational and socio-cultural contexts (i.e. the problem of attribution of effects to specific causes).

In their paper, they reviewed their own research in Sri Lanka (Wall and Alderson, 1993), as well as projects in Nepal, Turkey and the Netherlands by other researchers (Hughes, 1988; Khaniya, 1990; Wesdorp, 1982). Wall later went on to complete her doctoral thesis based on the Sri Lanka project which is dealt with in more detail below (Wall, 1999; Wall, 2005). They pointed out that, while many claims had been made about "washback effects", these were rarely backed up by evidence to allow us to understand the nature of the phenomenon. In other words, there was no theoretical basis underpinning the concept, nor systematic ways of relating it to the wider concept of test validity. In posing their question Does Washback exist? they attempted to problematise the concept, and suggested that appropriate research should be carried out to investigate it. To do so, they put forward 15 propositions or washback hypotheses (Alderson and Wall, 1993: 120). The first and most general hypothesis states that "a test will influence teaching" and on the assumption that teaching and learning are related, the next states that "a test will influence learning". The other hypotheses state that a test will influence some, if not all, classrooms in terms of the following:

- what and how teachers teach;
- what and how learners learn;
- the rate and sequence of teaching;
- the rate and sequence of learning;
- the degree and depth of teaching;
- the degree and depth of learning;
- attitudes to the content, method of teaching and learning etc. of teaching and learning.

In addition, they hypothesize that tests which have important consequences (i.e. are high-stakes) will have washback (but not necessarily on all teachers and learners), whereas tests which do not have important consequences will have little or no washback.
The significance and major influence of Alderson and Wall’s paper was that it brought to attention the fact that washback was a more complex concept than many had imagined, and their call for an increase in empirical studies was taken up by other researchers. Bailey (1996, 1999), Brown (1997) and Wall herself (1997) all lamented the scarcity of empirical studies into the complexities of the washback phenomenon and began addressing the issues in their own work.

Alderson and Wall guest edited an influential issue of *Language Testing* (1996: 13,3) which included Messick’s paper on validity and washback (the most referenced paper in the history of the journal) and five other papers by Alderson and Hamp-Lyons, Bailey, Shohamy et al, Wall and Watanabe.


Shohamy et al, in their paper (1996: 298-317) were among the first researchers to consider the “washback effect over time” and how this varies owing to societal factors such as the status of the language being tested and the uses of the test. The timeline and the timescale as a factor in monitoring impact is taken up again in our discussion below.

Alderson and Hamp-Lyons’ paper (1996: 280-297) is of particular relevance as it was the first attempt to explore the washback from the TOEFL which by the mid 1990s had nearly one million test takers a year. Taking the washback hypothesis as a starting point, their work was innovative in that they collected data based on interviews with teachers and learners and conducted classroom observations. They set out to challenge and investigate the commonly heard claim at that time that TOEFL tended to exert an undesirable influence on language teaching because of the multiple-choice item formats. Their results led them to conclude that simple forms of washback hypotheses are “too naïve” and that influences on what happens in classrooms are more complex. For example, they suggest that the *amount* and *type* of washback will varying according to a range of other factors, such as: the level of the stakes associated with the test; the divergence of the test from current teaching practices; the extent to which teachers think about test preparation; and the extent to which teachers and textbook writers are prepared to innovate. For them “the existence of the test by itself does not guarantee washback, either positive or
negative” and where it does exist it will vary from individual to individual thus providing an expansion of the Alderson and Wall washback hypothesis, as follows: “Tests will have different amounts and types of washback on some teachers and learners than on other teachers and learners.” (1996; 296).

It had become clear by the mid-1990s that many variables affect the type and extent of washback that might occur in educational contexts and that cause and effect explanations are rarely adequate. Alderson, in his forward to Cheng and Watanabe (2004), no longer asks the question “does washback exist?” on the contrary, he claims that there is no doubt that “washback does indeed exist” and in light of the growing body of research (e.g. as represented by the papers in the volume), that “the phenomenon is a hugely complex matter” and “very far from being a simple case of tests having negative impact on teaching”. The question is no longer “does washback exist?” but rather, “what does washback look like?”, “what brings washback about?” and “why does washback exist?”

2.3.2 Washback models

Taking a lead from Alderson and Wall, several researchers developed more complex “washback models” and suggested ways in which the kind of questions raised by Alderson might be addressed. The various models have sought to conceptualize washback more precisely and to guide a research agenda for investigating washback effects systematically. The publications of Bailey (1996, 1999), Cheng (1998, 2004, 2005), Watanabe (1997, 2004) and Green (2004, 2007) are particularly noteworthy and are reviewed in the next section.

Bailey (1996)

Bailey (1996) in her contribution to the special issue of Language Testing proposed a three-pronged approach to washback focusing on the three Ps - participants, processes and products. In this overview article entitled Working for Washback; a review of the washback concept in language testing she assembled these three facets into a basic model (acknowledging the influence of Hughes from an earlier, unpublished paper of 1993 in which he discussed the mechanisms of backwash).

Bailey started by posing the following questions:

- What is washback?
- How does washback work?
- How can we promote beneficial washback?
- How can we investigate washback?
These four questions served to structure her article and the washback model which she proposes was probably the first of its kind. The participants, processes and products, referred to by Hughes as a trichotomy, are set out in the following diagrammatic way (from Bailey 1996: 264 – figure 1) illustrating the presumed relationships between them. See Figure 2.1.

Figure 2.1 Bailey’s washback model

![Diagram of Bailey's washback model]

This analysis and its presentation were helpful in providing a systematic way of looking at the washback phenomenon. She listed four categories of participant (or stakeholders as we have also referred to them): students, teachers, materials writers/curriculum designers and researchers. She also listed four products: learning, teaching, new materials and curricula and research results.

The arrows in her figure show interactions between all the participants, and there are also arrows showing the influence of materials on the teaching and learning. The test, represented on the left of the Figure, influences the students, the teachers, the materials writers and the researchers who produce the outcomes in terms of teaching/learning, materials and curricula and research results. The curricula and materials influence the teaching and learning and all the products loop back onto the test.

As far as the direct influence of the test is concerned, Bailey distinguished between learner washback and programme washback. For her, learner washback is “the
effects of a test on the language learners themselves," while programme washback refers to test effects "on personnel involved in language teaching" (1996: 264-268). From the learner's point of view, she listed "test taker buy-in" as one of the important characteristics of ensuring beneficial (positive) washback. In other words, she believes that the learners need to accept that the test is fairly assessing the abilities that they need to learn.

In terms of programme washback, she listed teachers, administrators, course designers, and material developers as those participants who are influenced by the test and who ultimately influence the courses, programs and materials which the students receive. While this list is useful, it is a rather simplistic analysis as the types of role and potential for influence of these various participants can vary greatly in different contexts, and there are often more complex interactions between them as participants in the educational processes.

The process aspect of the model remained relatively underspecified (arrows with no commentary in the figure); in essence, the processes described boiled down to lists of behaviours which the main participants (teachers and learners) engage in when faced with a test.

In the fourth section of her paper where she discussed how beneficial washback might be promoted, she suggested that (positive) washback can be achieved by including the following in the testing process:

- promoting of language learning goals;
- building in authenticity;
- introducing learner autonomy and self-assessment;
- including detailed score reporting.

She concluded with the view that a test will promote beneficial washback to the extent that it is based on sound theoretical principles, it uses authentic tasks and texts, and the test takers buy into the assessment process (1996: 275-7). In her 1999 reworking of this paper commissioned by ETS, Bailey re-proposed her model and reviewed the washback literature of the 1990s in the context of the TOEFL 2000 project (Bailey, 1999).
Watanabe (2004)

Watanabe discusses methodology in washback studies and illustrates his points with examples taken from his research in Japan (Watanabe, 1996, 1997, 2000). Watanabe’s work is also referred to below in Chapter 3 in discussion of impact research methods. In his 2004 paper, he conceptualized washback in terms of five dimensions: specificity, intensity, length, intentionality and value. He suggests that the aspects of learning/teaching that may be influenced by an examination and the factors facilitating the process of washback are:

- test-related;
- status-related;
- stakeholder-related.

He also sought to isolate the features of a test which might be directly responsible for observable behaviours in classrooms. So, for example, he has suggested (2004: 28) that washback exists: a) if teaching is different in exam preparation and non-exam-preparation classes taught by the same teacher; b) if teaching is similar in exam-preparation classes taught by different teachers.

His focus on comparative methodologies, including the use of qualitative techniques, is noteworthy and relevant to the impact research discussed below.

Cheng (2005)

Cheng reported on research into examination reforms in Hong Kong during the 1990s and the washback of these changes on teaching materials and methodology (1998, 2004). In her 2005 volume, she relates the classroom behaviour of the participants (teachers and students) to the wider reform objectives of the education authorities. Her research was conducted in three phases (1993-6) and she collected both attitudinal and observational data from school contexts. In concluding, she observed that the methodology employed by the teachers tended to change more slowly than the content of teaching. In other words, the what changed as a consequence of the new examinations, but changes in the how were much more limited and slower to take effect.

Two other conclusions are relevant for this thesis: a) she confirmed the “complex nature of washback effects” (2004: 162) which others researcher had begun to realise; b) she demonstrated that washback is an “educational phenomenon in which change is central.” (2004: 164). Both these points are taken up again below.
Green (2007)

Green has developed the most comprehensive model of washback to date based on his PhD studies and set out in his subsequent volume *Washback in Context* (2007). While building on Bailey’s work, Green adds a number of new features, particularly focusing on the process aspects which were underspecified by Bailey.

His model is summarised in Figure 2.2 (taken from Green, 2007: 24) and covers washback *direction, variability* and *intensity*. Unlike Bailey, Green chooses to focus in more detail on the test itself (in his case the International English Language Testing System - IELTS) and on the way in which the focal constructs influence the washback effects. This approach moves the model away from “a recipe” for achieving positive washback, and towards a descriptive and partially explanatory tool addressing what goes on in order to cause the various washback effects.

Green’s model starts from test design characteristics (the *focal construct*) and related validity issues of *construct representation* identified with washback by Messick (1996), and also encapsulated in Resnick and Resnick’s (1992) formulation of ‘overlap’ or the extent of congruence between test design and skills developed by a curriculum or required in a target language use domain. It is claimed that test design issues are most closely identified with the *direction* of washback – whether or not these effects are likely to be judged to have beneficial or damaging effects on teaching and learning.
Importantly, the model relates design issues to contexts of test use, including the extent to which participants (including materials writers, teachers, learners and course providers) are aware of and are equipped to address the demands of the test and are willing to embrace beliefs about learning embodied therein. These features are the ones most closely related to washback variability (differences between participants in how they are affected by a test) and washback intensity. Green claims that washback will be most intense – i.e. have the most powerful effects on teaching and learning behaviours – where participants see the test as challenging and the results as important, perhaps because they are associated with high-stakes decisions, such as university entrance as suggested by Alderson and Hamp Lyons (op cit). (See also Cheng, 2005; Gipps, 1994; Haertel, 1999; Shepard 1991, 1993; Wiggins, 1998).

A survey of IELTS preparation courses conducted in tandem with this study indicated that the IELTS test was regarded both as important and challenging by a majority of learners on the IELTS preparation courses involved (70% rated success on the test
as ‘very important’). It also indicated that almost all the learners were taking the test because they intended to enter higher education in the UK. Most demonstrated at least a basic knowledge of the test format and viewed the direct test of writing as a positive feature. In short, the "conditions for intense washback" is represented by the diagonal (blue) arrow in Figure 2.2.

2.3.3 Researching washback – washback studies in the literature

The edited volume by Cheng and Watanabe (2004) entitled *Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning* brought together a number of important papers based on work carried out in the 1990s in researching washback (including the introduction by Alderson referred to above).

In their own review in this volume, Cheng and Watanabe classify the research of the 1990s into two major types: studies related to traditional, multiple-choice, large-scale tests invoking negative impacts (as described above); studies related to improvements to examinations which had been modified to produce desirable, positive change.

The volume also includes:

- a paper by Andrews on washback and curriculum innovation (with reference to Hong Kong);
- two papers on IELTS, one by Saville and Hawkey dealing with washback on teaching materials and the other by Hayes and Read investigating IELTS test preparation classes in New Zealand (see chapter 4 below);
- a paper by Watanabe on teacher factors mediating washback (discussed above);
- five papers dealing with educational reforms involving high-stakes language assessment in different parts of the world - in USA (Stecher, et al), Australia (Burrows), Hong Kong (Cheng), China (Qi), and Israel (Ferman).

See also Spratt (2005) for a useful overview of washback and the classroom, including a discussion of the implications of washback studies up to 2003.

A number of other washback studies were carried out and reported in the 1990s focusing on language test washback, involving an increasingly wide range of participants and stakeholders, including: test takers, language teachers, test developers, teacher trainers, curriculum planners, teacher advisors, head teachers and administrators, inspectors, end users and parents, materials designers.
In Tables 2.2 to 2.12, these studies are summarised by the category of research focus showing the researchers involved, the research method employed and the geographical context in which the research was carried out.

Table 2.2  Test takers

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fullilove (1992)</td>
<td>Observation and interview</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Cohen (1984)</td>
<td>Questionnaire</td>
<td>United States</td>
</tr>
<tr>
<td>Wall and Alderson (1993)</td>
<td>Observation</td>
<td>Sri Lanka</td>
</tr>
<tr>
<td>Ingulsrud (1994)</td>
<td>Observation</td>
<td>Japan</td>
</tr>
<tr>
<td>Sturman (1996)</td>
<td>Questionnaire and ethnographic</td>
<td>Japan</td>
</tr>
<tr>
<td>Shohamy, Donitsa, &amp; Ferman (1996)</td>
<td>Questionnaire</td>
<td>Israel</td>
</tr>
<tr>
<td>Cheng (1997)</td>
<td>Questionnaire</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Alderson &amp; Hamp-Lyons (1996)</td>
<td>Group discussion</td>
<td>United States</td>
</tr>
<tr>
<td>Hughes (1988)</td>
<td>Criterion-related</td>
<td>Turkey</td>
</tr>
</tbody>
</table>

Table 2.3  Language teachers

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hughes (1988)</td>
<td>Observation</td>
<td>Turkey</td>
</tr>
<tr>
<td>Cheng (1997)</td>
<td>Interview and observation</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Lam (1994)</td>
<td>Questionnaire</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Andrews (1994a)</td>
<td>Questionnaire</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Wall &amp; Alderson (1993)</td>
<td>Observation</td>
<td>Sri Lanka</td>
</tr>
<tr>
<td>Shohamy et al. (1996)</td>
<td>Observation</td>
<td>Israel</td>
</tr>
<tr>
<td>Watanabe (1996)</td>
<td>Observation</td>
<td>Japan</td>
</tr>
<tr>
<td>Alderson &amp; Hamp-Lyons (1996)</td>
<td>Interview and observation</td>
<td>United States</td>
</tr>
<tr>
<td>Wall (1999)</td>
<td>Interview and observations</td>
<td>Sri Lanka</td>
</tr>
<tr>
<td>Researchers</td>
<td>Research Methods</td>
<td>Geographical Context</td>
</tr>
<tr>
<td>------------------------------</td>
<td>------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Andrews (1994)</td>
<td>Interview</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Andrews &amp; Fullilove (1994)</td>
<td>Interview</td>
<td>Hong Kong</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrews &amp; Fullilove (1994)</td>
<td>Interview</td>
<td>Hong Kong</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wall &amp; Alderson (1993)</td>
<td>Interview</td>
<td>Sri Lanka</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fullilove (1992)</td>
<td>Interview</td>
<td>Hong Kong</td>
</tr>
<tr>
<td>Hughes (1993)</td>
<td>Observation</td>
<td>Turkey</td>
</tr>
<tr>
<td>Shohamy (1993)</td>
<td>Interview and questionnaire</td>
<td>Israel</td>
</tr>
<tr>
<td>Shohamy et al. (1996)</td>
<td>Interview and questionnaire</td>
<td>Israel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shohamy et al. (1996)</td>
<td>Interview</td>
<td>Israel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Researchers</th>
<th>Research Methods</th>
<th>Geographical Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andrews &amp; Fullilove (1994)</td>
<td>Observation</td>
<td>Hong Kong</td>
</tr>
</tbody>
</table>
These tables, although not exhaustive, provide a useful way of looking at the research priorities and methods which characterised the period of the 1990s starting with the work of Hughes referred to above. The studies listed mainly focused on classroom-based participants, the teachers and students/test takers, and there was still relatively little research which examined the roles and activities of other stakeholders in the wider context. It is also interesting to note that there was a significant interest in washback in Asia during the 1990s (particularly Hong Kong and Japan) - largely under the influence of British academics (e.g. as PhD supervisors).

On balance, we see that there were relatively few researchers at that time who were involved in this kind of research despite Alderson and Wall’s call for more research dating back to 1993. Impact, although not an unknown concept, was not a major research focus.

Another point to note is the emergence of qualitative methods to investigate the washback phenomenon. Alderson and Wall (1993: 127) had advocated a more “ethnographic approach to the topic” and the studies listed above show that this plea had been heeded with an increase in the use of qualitative research methods to look
into classroom events (involving interviews, discussions, observations and various kinds of questionnaire). This point will be taken up in Chapter 3 when looking at the methodology for investigating washback and wider aspects of test impact within an expanded model.

With regard to the role of the researchers themselves, it is significant that most were academics working on relatively short-term projects (e.g. linked to a three year PhD research programme), and there were no cases of researchers working on washback studies within a major examinations board or testing agency (with the exception of Cambridge ESOL – as we shall see in Chapter 4).

Since 2004, there have been several notable publications on washback with greater emphasis on impact as the wider phenomenon; for example, a special issue of Assessment in Education (March 2007) edited by Rea-Dickens and Scott entitled Investigating Washback in Language Testing and Assessment. This volume provides a number of new papers looking at the washback from language tests on teaching, learning and policy, with evidence from different settings. Fox and Cheng examine a literacy test taken by first and second language test takers in Canada; Scott focuses on stakeholder perceptions of test impact in the UK; and Qi asks whether testing provides an efficient “agent for pedagogical change” by examining the intended washback of the writing task in the NMET, a high-stakes English test in China. In doing so she builds on her own PhD studies and her earlier work published in 2004 and 2005.

The volume includes a paper by Wall and Horak which reports on work undertaken on behalf of ETS from 2002 onwards looking into the impact of changes made to the TOEFL on teaching and learning in central and Eastern Europe (Wall and Horak, 2006). Their paper draws on Wall’s PhD research (see below) and develops an argument for the use of baseline studies in the investigation of test impact.

Shohamy’s paper (2007) entitled Language tests as language policy tools picks up the arguments she put forward in her book The Power of Tests (2001) and her earlier warnings about the impact of assessment (1993: 19): “Tests used for the purpose of improving learning can be effective only if they are connected to the educational system; they are not effective when used in isolation. But using tests to solve educational problems is a simplistic approach to a complex problem. It works on people’s fear of authority. It can even be said that testers are abused by educational leadership”.

Her more recent discussion (2007) is relevant to impact because of the relationship
she identifies between values and beliefs in society and policy making with leads to the use of tests in particular contexts (as noted above). Her analysis links “ideology” to “practice” through the operation of “mechanisms” in society. Tests, she argues, can be used as mechanisms “for creating language education policies to control and manipulate de facto language policy.” (2007: 122).

2.4 Beyond washback

2.4.1 Impact research in language assessment

A consensus was beginning to emerge by 2004 that washback, referring to specific effects of a test on teaching and learning was itself a complex phenomenon and that it should be subsumed within the superordinate concept of impact. One limitation of the early washback studies was a rather narrow focus and despite the fact that some washback researchers, such as Cheng and Green, had sought to relate the washback to factors in the wider community, they had tended not to look at broader contextual factors and to the relationship between society and education.

In moving beyond washback to impact research, we return to the potentially deterministic nature of washback (noted by Alderson and Wall 1993: 118) and to Cheng and Watanabe’s observation that examinations can help to bring about positive change. It was in relation to assessment policy as an instrument of change in education that impact research moved beyond the narrower views represented by the washback models summarised above. The work of Wall (1996, 1997, 2000, 2005) is particularly noteworthy in this respect. She began to question how assessment policies reflect educational ideologies and can operate overtly as a mechanism for change, leading to improved practices in the teaching and learning languages.

Starting in 1993 with Alderson, Wall was the first to examine the literature on the management of change and to recommend using insights from the diffusion of innovation to develop a possible framework for investigating impact in language testing. Her work based on her thesis and using data collected in Sri Lanka was published in her book, The impact of high-stakes examinations on classroom teaching: a case study using insights from testing and innovation theory, (2005).

At the end of the 1980s and early 1990s, Sri Lanka provided a backdrop for a range of educational reforms and the role of examinations in that context was widely discussed at the time. In its post-colonial era after independence (1948), Sri Lanka continued to follow many traditions in education and assessment passed on by the British during 150 years of colonial rule and the Sri Lankan O level examination
project was an example of this – the O level being an English-style examination taken at the end of compulsory education, usually at age 16.

This broad context provided the political and educational setting for the project which led to Wall’s thesis and was influential in the thinking of Alderson and Wall discussed above. The O level English project was part of the Ministry of Education’s ELT reform programme which had begun in the early 1980s and was supported by British government aid through the British Council and the Overseas Development Agency (now DfID – Department for International Development). Alderson and Wall were appointed as consultants to conduct an evaluation of the O level project (between 1988 and 1991).

The much quoted idea that tests can be used as “levers for change” was suggested by Pearson (1988), a British consultant attached to the Sri Lankan Ministry in the 1980s. His paper entitled, *Tests as Lever for Change* (or “Putting First Things First”) appeared in the British Council’s ELT Documents: 128 (1988), based on papers presented at an international conference held in Columbia in 1985.

While not influential in the testing literature in general, Pearson’s paper is often cited in relation to the “levers for change” notion in the title. Although not without merit, Pearson’s approach has been characterised as rather naïve, and on closer analysis the factors influencing curricular innovation are more complex than his idea might suggest (see in particular Fullan, 1991 and 1999). In fact the realisation that we need to take careful account of “the myriad forces that can both enhance and hinder the implementation of the intended change” (Alderson, 2004: xi) underpins the need to develop a more complex model of test impact.

Wall’s book (2005) is important in demonstrating how innovation theory and its application increases our understanding of the hows and the whys of washback and the broader aspects of test impact. In her chapter 2 the impact of high-stakes examinations on classroom teaching is reviewed. This is a case study which uses insights from innovation theory, and in chapter 4, Wall focuses specifically on innovation processes in education (Fullan, 1991; Rogers, 1995), including the meaning of change for those affected by it. In her analysis she addresses the composite question: “Who adopts what, where, when, why and how?” (Cooper 1989, cited by Markee, 1993, 1997) and concludes with discussion of several models of innovation, including the hybrid model of the diffusion/implementation process developed by Henrichsen (1989) – see below.
In terms of the *what*, Wall reconfirms the now widely-accepted view that an innovation can be any idea, concept or practice which is "perceived as new" and which is intended to bring about improvement in a deliberate and planned way (Nicholls, 1983). She also notes the importance of the socio-cultural context (the "where") in which change takes place. Although Wall does not focus on this in particular, it turns out to be a key point in our understanding of test impact.

In terms of the *how*, Wall reviews the traditional models of change found in the literature: the Social Interaction Model; the Centre-Periphery Model; the Research, Development and Diffusion (RD&D); the Problem-Solving model; and the Linkage models (see also Markee, 1997 for a review of these in the management of curricular change). She refers extensively to the comprehensive survey of innovation in education by Fullan (1991) and the emerging consensus that innovation is a process *not an event*. This process typically has three basic phases, which lead to the desired outcomes: initiation, implementation and continuation. The Henrichsen model of innovation which she used as the starting point for analysing the Sri Lankan data is a synthesis of this approach.

Henrichsen sees the diffusion of innovation process as having three key components within a sequence: antecedents (before the changes are made - existing circumstances or baseline); process (during the change process); consequences (outcomes after the changes are made). The purpose of the *antecedents* component is to examine the current conditions in the educational context before an innovation is implemented. The *process* component examines the source of the innovation, the message, the plans and strategies and the channels of communication, and the factors which seem to influence the change process (features within the innovation itself, the resource system, the intended-user system, and inter-elemental factors). The third component, *consequences*, refers to the complexity in the types of decisions and outcomes that an innovation might lead to (in other words, the impact). Wall used the hybrid model to describe processes and outcomes after the reforms were implemented, i.e. as part of the post-exam evaluation she was taking part in. She documented many of the factors related to Henrichsen’s model to explain why the reform did not have the type of impact that was expected. Henrichsen himself did a post-analysis of a project in Japan (1956-1968), but he and other innovation theorists recommend the use of an innovation model *throughout the process*, i.e. before the reform begins, while it is in progress, and after it has been completed (Markee, 1993; White, 1988, 1993).
Henrichsen’s model is compelling as a basis for washback and impact research, especially in cases where there are well-focused educational objectives involving assessment. The notions of an existing “baseline” and outcomes described in terms of “consequences” allow for research designs of a quasi-experimental kind (see Chapter 3), as well as being coherent with a range of flexible approaches, such as case studies employing triangulation of quantitative and qualitative data. In terms of consequences, impact can be characterised in this model according to a number of facets: immediate vs. delayed effects; direct vs. indirect effect; intended vs. hidden (latent) effects; and functional vs. dysfunctional outcomes.

Wall adapted this model in light of her own experience of the Sri Lankan O level examination project which she had been involved in since the late 1980s and for which she collected data in 1991. In her thesis she developed an argument which suggested that the hybrid model is appropriately comprehensive without being overly detailed and is able to give a clear picture of the factors which are most important at different stages in the innovation process. She concludes that: “the framework seemed to offer the right amount of descriptive power while retaining a degree of flexibility; it provides a clear user-friendly starting point for the first stage of my analysis.” (2005:88).

In terms of the who, it is important to consider the various stakeholders or participants in the innovation process. Different impacts may be experienced by different stakeholders, and so the role of stakeholders in impact research must be a major consideration.

Wall argues that understanding the impact of a new examination requires a fuller understanding of the context into which the examination is being introduced. She was one of the first researchers to point out that testing specialists who are interested in promoting beneficial examination impact, often fail to account to the characteristics of the educational system (i.e. context or contexts) into which a new test or examination will operate. She recognised the lack of an appropriate model within language testing which could account for a wide range of contextual variables involved in this type of research.

In summary, while Wall’s work moved the field forward especially focusing on the role of context and management of change, a gap remained in that her approach did not involve the policy makers and test designers from the start. It emerges from the discussion above that the role of the various participants or stakeholders in the educational context is an important and recurrent theme in the washback research.
but the list of stakeholders in this research has remained relatively limited and constrained.

For Wall, as with many other researchers noted in Bailey’s survey (1999), her impact research began *post hoc* and was designed to investigate existing contexts or to follow up innovations which had already been decided by policy makers without systematic consideration of the washback/impact implications during the planning phase. Wall’s more recent work with Horak on the new TOEFL (2006, 2007) has also been conducted from the perspective of *outsiders* without having had prior influence on or insights into the design and development of the revised test. The effect of this is to make it difficult to involve the relevant stakeholders in the research. What was missing is the active involvement of the examination provider (examination board, education department etc.) in the conceptualisation and investigation of test impact; in most cases where high-stakes tests or examinations are developed, the policy makers and examinations providers are likely to be the only stakeholders with sufficient authority and resources to conduct such work.

This thesis attempts to address this issue by developing an impact model from the perspective of an examinations board, i.e. one which can integrate considerations of impact into the design, development and validation of the examinations.

### 2.4.2 Impact and validity theory

In Chapter 1, the role of validity theory in the conceptualisation of washback and impact was introduced. The extent to which the effects and consequences of a test can be seen as an intrinsic aspect of the test’s validity has received particular attention in the literature and needs to be addressed in more detail.

Some writers have gone as far as to suggest that a test’s *validity* should be judged by the *degree of beneficial washback* that it has on teaching (i.e. the test should have pedagogical aims to be valid). For example, the concept of *systemic validity* was proposed by Frederiksen & Collins (1989), and *washback validity* was favoured by Morrow (1986). Frederiksen & Collins (1989: 27) suggest that an assessment will be “systemically valid” if it can induce in the education system “curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure”.

The rather simplistic notions of systemic and washback *validity* have been criticised by some language testers (Alderson and Wall, 1993: 116; Messick, 1996: 242) and generally there is a consensus in testing circles that notions such as these should not
be considered aspects of validity (see Bachman’s 1990 criticism of Morrow). Nevertheless, the relationship between the content of examinations and the teaching content cannot be ignored, as we have already noted and is reflected in much of the washback research reported above.

The notion of direct testing (as in the communicative approach to language testing associated with Morrow) and the presumed importance of the link between directness and authenticity have been important themes since the 1980s (e.g. in the work of Swain, Hughes, Bailey reported above and others such as Weir, 1990).

This view extends beyond language testing so, for example, Wiliam who has been associated with the “formative assessment movement” in the UK, has stated that “a test will be valid if one would be satisfied for teachers to teach to the test.” (1996). Goldstein, (1989) writing in a similar vein has referred to the need for educational assessment to be “connected” to teaching and learning processes rather than a separate measure which avoids any connection with specific learning environments (cf. Shohamy, 1993, quoted above). Messick himself in the paper quoted above (1996: 241) also noted that, as far as possible, “the move from learning exercises to test exercise should be seamless. As a consequence, for optimal positive washback there should be little if any difference between activities involved in learning the language and activities involved in preparing the test.”

It is important, therefore, to relate the qualities of a test or examination to curriculum policy and the associated teaching practices - as in the washback model of Green who notes that there is a “broad consensus that test content and its relation to criterion skills is central to washback” (Green 2007: 14). The concept of impact also needs to be considered within the development of validity theory and the measurement literature since Cronbach, the “father” of construct validity in psychometrics.

In 1955 Cronbach and Meehl set out the concept of the “nomological network” which became central to validity theory in the second half of the 20th century (Cronbach and Meehl, 1955). Their ideas dominated thinking in psychometrics in the USA until the late 1980s (although the influence on educational assessment and language testing in the UK emerged more slowly and did not impinge on educational practices until quite recently - see discussion in Bachman et al, 1995).

The nomological network was conceived as a “law-based network” enabling the central concern of construct validation to link the conceptual/theoretical with the
observable. Cronbach and Meehl suggested that in order to provide evidence that a measure has construct validity, a nomological network for that measure should be developed, including the theoretical framework for what you are trying to measure, an empirical framework for how you are going to measure it, and a specification of the linkages among and between the two.

This approach provided the basis for the “argument-based” approach to validation which has now become prevalent and for the notion that it is not the test but the “principles for making inferences” which are validated (1955: 297). The importance of the “trait” underlying the test (1955: 283) and thus the cognitive dimension of assessment was also established by Cronbach and Meehl.

However, while the philosophical foundation for construct validity derived from the prevalent logical and empirical philosophy of science of the 20th century, Cronbach’s early work did not provide a practical and usable methodology for actually evaluating construct validity in wider educational contexts. The next phase in the evolution of validity theory saw the development of methodological approaches for construct validation (e.g. multi-trait, multi-method – see Campbell and Fiske, 1959) and the incorporation of a social dimension, especially through the work of Messick, whose contributions have been particularly relevant to the notions of washback and impact since the 1980s.

In 1989 Messick proposed a “unified theory of validity” in the form of a progressive matrix referred to in Chapter 1, and he introduced “cells” on both value implications and social consequences (1989: 20). In 1995 he wrote: “…validity is broadly defined as nothing less than an evaluative summary of both the evidence for and the actual – as well as potential – consequences of score interpretation and use (i.e. construct validity conceived comprehensively). This comprehensive view of validity integrates considerations of content, criteria and consequences into a comprehensive framework for empirically testing rational hypotheses about score meaning and utility” (1995: 742).

The implications of his matrix and the points made in this quotation have been extensively discussed in the testing literature (e.g. Bachman; 1990; Kane 1992; Shepherd, 1993; Mislevy, 1996, 2003, etc.).

According to Messick, test outcomes should be considered in terms of their interpretation and use, and the justifications for the validity need to be related to: the test construct itself; its utility for the intended test interpretation and use of scores; the
implied values and social consequences of interpreting and using the test in context. Messick calls these the evidential and consequential bases of the validity argument.

2.4.3 Validity by design as a likely basis for washback

Messick’s article on Validity and Washback in Language Testing (1996: 248-9) returns to the discussion of construct validity with particular relevance to language assessment and the concept of washback. In this paper he makes a connection between validity and the delimitation of effects of tests “by attempting to minimise sources of invalidity in language test design, the test deficiencies and contaminants that stimulate negative washback are also minimised, thereby increasing the likelihood of positive washback.” Green (2007: 6) has referred to this as “washback by design”.

Messick focuses on six aspects of construct validity which, he suggests, function as general validity criteria for all educational and psychological measurement and provide the means to check that the validity argument “touches the important bases” and that any significant omissions can be justified. The scope of washback and impact can, therefore, be set against these six aspects of validity as follows:

- the content aspect - relevance, representativeness, technical quality;
- the substantive aspect i.e. theoretical rationales;
- the structural aspect – fidelity of scores to the construct domain;
- the generalisability aspect - score interpretations across population groups, settings and tasks;
- the external aspect - including criterion relevance and applied utility;
- the consequential aspect - value implications for score interpretation and consequences of test use.

The external and consequential aspects are particularly relevant to a discussion of washback and impact. The consequential aspect includes the value implications associated with the ways in which test scores and interpretations provide the basis for actions (e.g. decisions), as well as consequences for stakeholders relating to issues of bias, fairness “distributive justice” and the effect on teaching and learning. According to Messick, evidence and rationales for evaluating the intended and unintended consequences, both in the short- and long-term, should be sought (the timeframe being important from the perspective of an examinations board – as we shall see below). However, Messick cautions that the consequential aspect should not be seen as a separate type of validity but is a necessary part of construct validity. This is because the outcomes of test interpretation and use both derive from and
contribute to the meaning of test scores. [NB. Despite Messick's warning, "consequential validity" is now often used to refer to this concept.]

While intrinsically related to the focal construct, the value implications and effects of using test scores in a particular context for a particular purpose are social phenomena, and therefore it is important to consider the nature of these phenomena. For example, a policy decision to develop and administer a certain type of examination in schools embody certain values and goals, and will therefore have both intended and unintended consequences related to those values and goals. This is also true of language testing policies as Bachman (1990) points out: “language testing occurs in an educational and social setting, and the uses of language tests are determined largely by political needs that change over time and vary from one society to another. Therefore we need to move beyond applied linguistics and psychometric theories and consider the social and political functions served by our tests”.

Messick acknowledges that the test developer is not able to guarantee positive impacts and he speaks of probabilities and likelihoods only. He suggests, however, that it may be possible to delimit negative impacts and to improve the chances that the impacts will be benign or at best positive. This can be achieved by minimising sources of invalidity caused by construct under-representation and construct irrelevant variance and by ensuring that test validation is carried out to gather “sufficiently compelling evidence to counter these two major threats to validity”.

This means that if differences in the scores are caused, in part, by differences which are not relevant to the traits of interest, the test actually measures things it shouldn’t (it is too “wide”); this is a case of construct-irrelevant variance threatening validity. If, on the other hand, important differences in the traits of interest are not actually reflected in the scores, the test doesn’t measure things it should (it is too “narrow”); this is a case of construct under-representation. In both cases, the threats to validity can lead to undesirable washback effects, but both can be addressed through the design of the test which addresses the technical qualities of the test – hence washback by design.

For Messick, therefore, those who refer to washback validity (such as Morrow) are mistaken in that they “seek washback as a sign of validity”. The correct formulation for Messick is to “seek validity by design as the likely basis for washback”. Essentially this means that by attempting to improve validity the likelihood of achieving positive washback is enhanced “pragmatically”, the touchstone is to ensure that an
assessment adequately ‘represents the focal construct’ using formats that are ‘acceptably obtrusive within the practical constraints of feasible test administration and scoring, that is formats in which method variance is relatively minor and can be taken into account in scoring and interpretation.” (1996: 252).

In summary, Messick’s paper highlights four validity questions which are relevant in the investigation of test impact:

- Are the value implications of score interpretations empirically grounded and commensurate with the score’s trait implications?
- Do the scores have utility for the proposed purposes in the applied setting?
- Are the scores applied fairly for these purposes i.e. consistently and equitably across individuals and groups?
- Are the short- and long-term consequences of score interpretation and use supportive of general testing aims and are there any adverse side-effects?

The other main point in Messick’s discussion of washback is his concern for evidence in the validation process and the implications of this concern for impact research.

2.4.4 The influence of Bachman’s work

In Chapter 1 the influence of Bachman’s work on Cambridge ESOL’s approach was noted, especially the so-called “VRIP” features, including impact. Impact as a “test quality” emerged in the work of Bachman following his 1990 volume *Fundamental Considerations*. Although he refers to washback, he does not use the term *impact*: this appeared later in the period 1990 to 1995 (e.g. in talks, seminars etc.) and in his volume with Palmer (1996), *Language Testing in Practice*.

In his discussion of validation, Bachman (1990) reflects the influence of Messick’s work; he explicitly uses the term validation “as a general process that consists of marshalling of evidence to support a given interpretation or use, a process that is based on logical, empirical and ethical considerations.” This process is not a one-off event but a continuous one, involving both logical analysis and empirical investigation (cf. Messick’s view that validity is “a matter of degree not all or none” - 1989:33). The accrued evidence for the validity of test interpretations must be derived from a variety of sources and be gathered in a number of ways over time.

In an extended section on validation, he also discusses *the ethical or consequential basis for test use* (1990: 279-285), i.e. the aspects of validation which are of
particular relevance to this thesis. He covers these under the following “areas of interest”:

- The rights and interests of the test takers and of the institutions responsible for testing and making decisions based on those tests, and the public interest.
- Value systems that inform the particular test use.
- The practical usefulness of the test.
- The consequences to the educational system of society of using test results for a particular purpose (e.g. washback as “the effect of testing on instruction”).
- Alternatives to testing (cf. Spolsky’s *Use with care!* discussion cited above).

Bachman’s conclusion is that “it is our responsibility to provide as complete evidence as possible that the tests that are used are valid indicators of the abilities of interest and that these abilities are appropriate to the intended use, and then to insist that this evidence be used in the determination of test use”.

In 1996, the notion of test usefulness is explained more fully (see chapter 2, *Test Usefulness: qualities of language tests*) and is set out in what might be called a “utility function” (based on the earlier VRIP features):

\[ \text{Usefulness} = \text{reliability} + \text{construct validity} + \text{authenticity} + \text{interactivity} + \text{impact} + \text{practicality} \]

Usefulness captures Messick’s notion of “construct validity conceived comprehensively”, and so, in order to achieve usefulness, there is a need to balance all the qualities in relation to a specific language testing situation: i.e. “with a specific purpose, a particular group of test takers and a specific language use domain … in mind.”

Impact and practicality are both recognised as essential test qualities in this approach. Bachman and Palmer refer to the broader context of test use which includes the value systems of society, arguing that “it is essential to take a systemic view, considering tests as part of the larger and societal context”. While they acknowledge the concept of washback, they point out that language testers must be prepared to deal with phenomena that are “far more complex and thorny than simply the effect of testing on teaching” (1990: 30-31). Impact, therefore, is a quality of the test which operates at the two levels described above and referred to in Chapter 1,
i.e. at a micro level in terms of the individuals who are affected by the particular test use and at a macro level in terms of the educational system or society.

Another important point which emerges from their discussion is the role played by individuals as stakeholders “a variety of individuals will be affected by and have an interest, or hold a “stake” in the use of a given test in a particular situation. …… Stakeholders that are directly affected include the test takers and test users or decision makers. In addition, a large number of individuals … will be indirectly affected.”

Bachman and Palmer focus their attention on the stakeholders who are, in their view, most directly affected by test use - the test takers and teachers (cf. Bailey’s participants). They deal with impact on society and education systems and individuals who are indirectly affected only very briefly in just over one page (1996: 34-35). They recommend that test developers should think carefully about what might happen as a result of using a test for a particular purpose and they offer some practical advice (reminiscent of Hughes’ “imperatives” for achieving positive washback):

- List the intended uses.
- List potential consequences – positive and negative – of using the test in particular ways.
- Rank desirability of these outcomes.
- Collect information to determine likelihood of these outcomes.

They also provide some concrete examples and discuss the specific considerations for assessing the potential impact of test use and for taking these into account in the design and development of tests (1996: chapters 7 and 9).

In summary, Messick’s conceptualisation of validity is an important basis for impact research and Bachman and Palmer’s work provides important guidance on its application to language testing. The emphasis on the cognitive dimension in construct definition remains an important consideration and the need for an overarching term such as usefulness to capture the range of test qualities which contribute to overall validity is recognised. In addition, however, it is important to extend the focus of attention to the social dimension of assessment and this is taken up in the next section.
2.4.5 Impact and the wider social context (the milieu)

The social context within which assessment takes place and the relationship between micro and macro levels has emerged in this chapter as an important consideration in modelling the impact of language assessment. The ways in which assessment policy functions within educational systems was introduced at the beginning of this chapter; this dimension needs to be addressed in more detail taking into account the complexity of the educational systems themselves and the dynamics of the change processes which occur when policies are implemented, as discussed by Wall in her conceptualisation of impact.

As we have seen in the discussion of validity, the use of a test in a specific educational context creates a range of effects and consequences which are only in part related to the intrinsic qualities of the test itself. For Messick only those aspects which are construct-related can be attributed to the test. Likewise, Bachman and Palmer (1996) consider only two of the qualities to be intrinsic to the test — reliability and validity, restricting the term validity to the test construct, which they see as an essential measurement quality.

However, there is a growing concern for the scope of the validity argument as understood by Messick’s conceptualisation. Moss, for example, writing from an educational perspective expresses this concern and highlights the importance of the social impact of assessment practices in the following way (1998: 11): “the scope of the argument goes well beyond … test specific evaluation practices; it entails ongoing evaluation of the dialectical relationship between the products and practices of testing, writ large, and the social reality that is recursively represented and transformed”.

The journal Measurement: Interdisciplinary Research and Perspectives, launched in 2003 and for which Moss is one of the editors, has published a number of “focus articles” addressing this concern and are relevant to this discussion. See for example, Pellegrino and Chudowsky, (2003); Kane, (2004); and Moss, Pullin, Gee and Haertel, (2005).

McNamara and Roever in their volume on the social dimension of language testing also note that “integrating studies of washback and impact into a larger interpretive argument has the advantage of framing the significance of such studies more clearly in terms of investigating the policy assumptions involved in testing programs”. (2006: 37)
They argue that the wider social context in which language tests are commissioned and have their place “is still not adequately theorised” (although as noted above, the critical testing movement has provided a theoretical challenge to existing philosophies of testing by considering the role and function of tests in the broadest political light). The work of Shohamy cited above has also been helpful in understanding how tests might lead to unfairness or injustice at the societal level.

In considering the values and beliefs that inform assessment policies, we therefore need to consider the values of the stakeholders themselves, the institutional values in the educational system and the value systems of society at large within a particular national or regional culture. A concern for the social dimension of assessment, and the particular relevance of social context for language learning and language testing, and the complex network of relationships which exists among the many different stakeholders in educational processes need to be carefully considered in impact research. While language learners and teachers are most obviously affected by the language tests, many other stakeholders are also impacted in different ways. (For discussions of stakeholders in assessment see Rea-Dickens, 1997; Saville, 2003; Saville and Hawkey, 2004; Taylor 1999. See also the literature on Evaluation, e.g. Roy, 1998; Varghese, 1998; Weiss, 1998).

Rea-Dickens (1997: 305) attempted to draw up a comprehensive list of potential stakeholders in language testing and she came up with the following: “language testers, teachers, parents, administrators, teacher educators, sponsors and funding bodies, government bodies, the public, various national and international examination authorities, members of working parties and curriculum committees, test takers (and the larger group of learners of whom they form part ) … … test administrators ….. test users, for example university admission officers who need to interpret scores on national and international proficiency tests…..”

Since Rea wrote this the ‘stakeholder’ concept has become increasingly important in the washback and impact study literature (Taylor 1999). Figure 2.3 (adapted from Taylor, 1999) shows the wide range of stakeholders involved in various ways in the development, construction and administration of high-stakes tests such as those produced by Cambridge ESOL. As Saville (2003:60) indicates, this kind of “taxonomy of stakeholders” places demands on an international examination provider to ensure that systems are in place “to review and change what it does in the light of findings on how the stakeholders use the exams and what they think about them”. This in itself is can be taken as a justification for including test impact studies as part of the test validation process.
In considering the range of stakeholders impacted by assessment or who contribute to assessment processes and mechanisms, we need to look beyond the discussions in the language testing literature (cf. Bachman and Palmer’s discussion of stakeholders and Bailey’s participants). Both the conduct and behaviour of individuals, based on their knowledge, beliefs, attitudes and preferences, contribute significantly to the success of planned activities in society. This takes us back into the realms of sociology, education and politics introduced at the start of this chapter.

Returning to Merton’s law of unintended consequences, it has been observed that the causes of unintended consequences are not easily traceable to a single source, but are related to the inherent complexity of social systems, and result from one part of a system responding to changes in another part of the same system, or to changes in the wider environment. These changes cannot necessarily be predicted as they may be affected by random factors or rare events (Taleb, 2007). As Merton observed 70 years ago, the interaction between policy, the incentives that are created for particular actions, and the individual traits and behaviour of the people involved, can lead to both unintended and unanticipated consequences. In some
cases negative consequences are the result of “perverse incentives” created by the policy itself (as we have noted above). These kinds of phenomena have been discussed in the educational literature on change processes. (See in particular: Fullan and Stiegelbauer, 1991; Fullan, 1993; 1999).

Eckstein and Noah (1993: 243) outline eight persistent “dilemmas of examination policy” that examination authorities “have to live with”:

- How to retain the function of the examination as a stabilizing element in education and society.
- How to promote examination results that are comparable and understandable.
- How to maintain the value of the credential earned by examination success.
- How to ensure a sufficient degree of regional authority and school or individual autonomy.
- How to use the examination to select for subsequent education, training and jobs.
- How to incorporate the new information technologies in the examination system and secure the promised benefits of greater efficiency and cost savings while safeguarding teaching and learning from undesirable side effects.
- How to maintain and increase the professional autonomy of teachers.
- How to raise standards of performance.

“Each of these dilemmas represents a potential problem associated with use of the examination system to control either individual destinies or what occurs in schools.” (1993: 244). They caution that, “as the United States quite properly considers reform of the testing and examination system as a way to lever education to a higher level of performance, policymakers need to be constantly mindful of the likelihood of unintended, undesirable consequences of their decisions.”

The relationship between high-stakes tests and change processes in education has attracted much attention in recent years as a variant of the law of unintended consequences. Known as Goodhart’s Law (or Campbell’s Law in the USA), after the original author who first expressed the principle, this “law” has been used on both sides of the Atlantic to point out the danger of negative consequences occurring when high-stakes tests intended for measuring pupils’ achievement are also used for monitoring educational standards. Goodhart (1975) claimed that once a social or economic indicator is made a target for the purpose of conducting social or economic policy, it will lose the information that enabled it to play such a role. In other words,
the indicator becomes a target in its own right and this has the effect of changing the
processes it was intended to measure.

Merton also noted this danger when he pointed out that a peculiarity of human
conduct stands in the way of successful social prediction and planning: “... public
predictions of future social developments are frequently not sustained precisely
because the prediction has become a new element in the concrete situation, thus
tending to change the initial course of developments. This is not true of prediction in
fields which do not pertain to human conduct.” (1936: 902)

The English educational system has been characterised by “assessment-led” change
since the 1980s, and government policy has explicitly exploited the presumed effects
of assessment on schools (see Broadfoot, 1979; Gipps and Murphy, 1994; Mansell,
2007). However, educational researchers have shown that important consequences
of this policy have been contrary to what was intended. Wiliam (1996, 2008), for
example, has noted that the use of examination results to produce school league
tables, which in turn are used to monitor school standards in England, is an example
of Goodhart’s Law. The effect of the policy has been to focus attention on the targets
and as a result to narrow the curriculum, increasing “teaching to the test” without
bringing about the intended improvements in pupils’ achievement.

Similarly in the USA, some consequences of the No Child Left Behind (NCLB) act
have been seen as an instance of the same phenomenon: scores on the tests have
become the goal of the teaching process, thus losing their value as indicators of
educational standards and distorting the educational process in various undesirable
ways (Chalhoub-Deville and Deville, 2008; Melken, 2008). So if the NCLB requires
schools to show improvement in students’ test scores, the schools respond in various
ways to ensure that this improvement is demonstrated. Although the test scores may
rise, the intended improvements in standards may be threatened rather than
achieved (i.e. a perverse result).

The impact in the schools (e.g. the narrowing of the curriculum) is not principally to
do with the focal constructs of the tests themselves being deficient, but is a result of
the policies which lead to the uses and interpretations of the test results in society.
The perverse incentives illustrate how educational policies can create a tension
between educational objectives at the micro level (teaching and learning in schools)
and a requirement for accountability at the macro level. These examples point to the
problem of making policies without anticipating ways in which the outcomes may
differ from those which are intended (e.g. due to intentional or unintentional deviance
from the desired practices and behaviours). Such considerations are important in developing a model of impact which incorporates both micro and macro level features.

2.5 Summary and conclusions

2.5.1 Summary

This section summarises the points which have emerged in this chapter and the research questions to be addressed are posed.

The chapter introduced the concept of impact with reference to the ways in which it is used in social contexts, and the effects, consequences and influences of social action were discussed with specific references to side-effects and unintended, negative consequences. How examinations function within educational reform programmes was discussed and relevant literature in sociology, education, applied linguistics, and language testing was reviewed.

The chapter also plotted the development of washback research from around 1980 onwards, focusing on the emergence of “washback models” during the 1990s. Research methods were summarised and the review revealed several issues to be addressed in conducting impact research in educational contexts, e.g. the difficulty of establishing dependent variables (Green, 2007); the importance of qualitative methods, including use of case studies (Wall, 2005). A detailed review of the literature on case studies and meta-data was not included in this chapter but will be covered more fully in Chapter 3.

Table 2.13 summarises the developments of washback and impact models starting with Alderson and Wall (1993) and ending with Green’s washback model (2003).
Table 2.13 10 year trend in washback/impact model development from 1993

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Washback Hypotheses</td>
<td>Basic model - Washback</td>
<td>Basic model - Impact</td>
<td>Washback Methodology</td>
<td>Revised washback model</td>
<td></td>
</tr>
<tr>
<td>Washback metaphor</td>
<td></td>
<td></td>
<td></td>
<td>“Washback by design”</td>
<td></td>
</tr>
<tr>
<td>Washback hypotheses (15)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mainly limited to classroom contexts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Comparative analyses are important in washback research</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Ps:</td>
<td>Participants</td>
<td></td>
<td>Diffusion of Innovation – change as a process over time in phases (Henrichsen)</td>
<td>Complexity of washback</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Process</td>
<td></td>
<td></td>
<td>5 dimensions:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Products</td>
<td></td>
<td></td>
<td>Specificity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“learner washback”</td>
<td></td>
<td></td>
<td>Intensity</td>
<td></td>
</tr>
<tr>
<td></td>
<td>“program washback”</td>
<td></td>
<td></td>
<td>Length</td>
<td></td>
</tr>
<tr>
<td>Participants mainly teachers and learners</td>
<td></td>
<td></td>
<td></td>
<td>Intentionality</td>
<td></td>
</tr>
<tr>
<td>Others participants (mainly) related to the local context</td>
<td></td>
<td></td>
<td></td>
<td>Value</td>
<td></td>
</tr>
<tr>
<td>Limited focus on process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Antecedents Process</td>
<td>Focus on intended users and user systems</td>
<td></td>
<td></td>
<td>Factors mediating process: Test Prestige Personal Context (micro and macro)</td>
<td></td>
</tr>
<tr>
<td>Consequences</td>
<td>Characteristics of the innovation and the context</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Importance of the inter-relationships between variables</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amplified focus on characteristics of teachers and learners during implementation phase</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The first attempt to extend the washback concept and to develop a model of impact was made by Wall in her work in Sri Lanka in which she focused on language assessment within an educational reform process. The management of change and the need to monitor changes over time and for appropriate comparisons to be made was highlighted.

The shift of emphasis from washback to impact at the end of the 1990s was discussed in relation to validity theory, especially the theoretical framework...
developed by Messick who refers to the consequential aspects validity. In this thesis, impact is broadly located within this framework: how language constructs relate to uses and consequences is a crucial consideration, and validation involves the accumulation of validity evidence over time.

The final part of the chapter returned to the wider educational literature and a focus on the complex social systems (milieu) within which “ripples” of test impact are felt. The need to incorporate social dimensions of assessment into a model of impact was endorsed. Finally the importance of educators and assessment providers working together to address dilemmas and to anticipate potential impacts wherever assessment forms part of a wider policy agenda was highlighted.

In proposing an expanded model of impact to be used in the context of Cambridge ESOL’s work, the following issues emerge from this overview:

a). the role and nature of the language tests in education and in society – the purposes, uses, and intrinsic qualities, especially the nature of the focal constructs;
b). the nature of social and educational contexts – the micro and macro levels – and how language assessment processes fit within them;
c). the needs to focus on the wide range of people who are involved in assessment practices – i.e. the participants or stakeholders;
d). the importance of studying change and to understand how this happens over time e.g. change management processes and the importance of non-linear, dynamic systems (complex dynamic systems);
e). how to conduct impact studies by extending what has been learnt in washback research, including the appropriate roles of researchers, the research methods and ways of dealing with the research outcomes (e.g. as data for test validation purposes);
f). how to anticipate possible effects and consequences more effectively and how to implement procedures to deal with unpredicted consequence when they occur (especially those which are detrimental).

Whether a test or examination achieves its intended objectives in an educational context can only be ascertained if appropriate comparisons can be made, e.g. to compare people, contexts and occasions, or various complex combinations of all three. Hawkey (2006:11-12) discusses the spatial/temporal dimensions of washback and impact studies and includes charts which locate impact studies in programme and test developments (Table 2.14)
Table 2.14 Spatial/temporal dimensions of washback

<table>
<thead>
<tr>
<th>Before</th>
<th>During</th>
<th>After</th>
<th>Etc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1</td>
<td>Time 2</td>
<td>Time 3</td>
<td>Time 4 etc.</td>
</tr>
<tr>
<td>Place 1</td>
<td>Place 1</td>
<td>Place 1</td>
<td>Place 1</td>
</tr>
<tr>
<td>Place 2</td>
<td>Place 2</td>
<td>Place 2</td>
<td>Place 2</td>
</tr>
<tr>
<td>Etc.</td>
<td>etc.</td>
<td>etc.</td>
<td>etc.</td>
</tr>
</tbody>
</table>

Taking a diachronic approach recognises that change over time is important and that access to baseline information is a relevant factor, i.e. to compare what is going on before and after changes are made. The comparison of test taking/learning contexts (places) within educational systems at different levels is also important (cf. Figure 1.1). Macro level comparisons might be made between local, national, or international contexts. At the micro level, comparisons might be made in schools and classrooms, e.g. comparing types of classes, learners and teachers within the same school.

2.5.2 The research question

In light of the need for an expanded model of impact raised in Chapter 1 and supported by the literature review in this chapter, we will investigate the following research question from the perspective of Cambridge ESOL:

What are the essential components of an action-oriented model of impact that would enable the providers of high-stakes language examinations to investigate the impact of their examinations within the educational contexts in which they are used?

This question is addressed through the analysis of the three case studies in Chapters 4 to 6 where we take account of:

- the actions examinations boards need to take to ensure that their tests and examinations have desired influences, effects and consequences?
- the means by which examinations boards can work more effectively with other stakeholders in education and society generally to foster beneficial impacts of their tests and examinations?
• the potentially harmful effects to be avoided by examinations boards, or if they occur, how the negative consequences can be remedied by subsequent actions in the short and longer term?

The outcome in Chapter 7 is a proposed model of impact which incorporates a theory of action and which can be integrated into the working practices of examinations boards such as Cambridge ESOL.
CHAPTER 3  Methodology

3.1 Introduction

In this chapter we will look at two aspects of the methodology employed in the thesis. The first is the approach to formulating hypotheses and collecting data within the real world paradigm of research into which the impact research itself falls. The second concerns the use of case studies and case studies as meta-data.

3.1.1 Real world research

Impact research can be contextualized within the domain of “real world research” (Robson, 1993, 2002) and a range of methodologies commonly used in the social sciences. In particular we will argue that there is a need to develop a repertoire of research tools, which are both quantitative and qualitative in nature (as observed in washback research reviewed in Chapter 2).

The washback studies of the 1990s revealed the extent to which contextual variables interact to confound the interpretation of research outcomes, and so an expanded view of impact and its scope is likely to encounter greater challenges in this respect. The dual features of real world contexts and many complex variables mean that the research methodologies for investigating impact need to be robust enough and flexible enough to deal with these challenges.

3.1.2 Where do impact studies fit in?

In the past few decades there has been growing interest in real world research in education, as well as in many other fields, and a number of useful handbooks for social scientists and practitioner researchers have been produced which draw attention to the issues and complexities involved: “The ‘real life’ situation refers in part to the actual context where whatever we are interested in occurs, whether it be an office, school, hospital, home, street or sports stadium”. (Robson, 2002, 3).

Test impact studies are concerned with real life situations in which a common feature in the situation is the presence of assessment, and in our case, language tests or examinations.

While laboratories are ideal for conducting experiments in a controlled and closed environment, the real world does not allow for such control in the actual context where whatever we are interested in occurs, and even if it were feasible, such control
might be ethically questionable. Real world contexts tend to be complex, open systems which can only be partially or relatively poorly controlled and over the past 30 years or so, as this has become increasingly apparent, social scientists and practitioners have developed research methods to accommodate this fact of life.

One of the issues which this research community has had to confront was a generally held view that the “gold standard” for social research was the experimental approach (e.g. using randomised controlled trials) in fixed research designs. This tended to mean that there was a preference for research designs which were quantitative and supported by statistical analyses. However, the value of alternative approaches to social research has been recognised in many fields of enquiry (including education, health, social work, marketing, business management etc.), as well as in many academic disciplines (sociology, psychology, anthropology etc). Central to this changing view has been the growing use of qualitative research methods, flexible research designs, and designs which employ mixed methods allowing for both quantitative and qualitative approaches to data collection.

Some researchers have questioned whether these alternative methods provide a sound scientific basis for research leading to a debate with views expressed from two sides: while some quantitative researchers have queried the scientific credentials of qualitative approaches, some researchers from the qualitative paradigm would not wish their work to be seen as “scientific”. Robson acknowledges this divergence of views but argues persuasively that there is a place for both qualitative and quantitative approaches and “strong arguments for characterizing both fixed and flexible designs as scientific” (2002, 5). An important proviso, however, is that they follow “sound research principles” and are carried out in a systematic and principled way.

So where do impact studies fit into this methodological debate? Are there existing approaches to impact research in other fields which could be used to inform this debate, or useful similarities of approach in other evidence-oriented fields of enquiry, such as health care, which can be drawn on?

In applied linguistics research, both qualitative and ethnographic research techniques are increasingly being used by researchers in language teaching and learning projects. The strategies used in these approaches seem to be relevant for research into washback, as noted in Chapter 2. Watanabe (2004: 23) has pointed out that it is difficult to find fully controlled experiments being used in washback studies and suggests that this may be because the experimental approach is likely
to mask the reality that other techniques can more readily reveal. There is an assumption here that the more flexible approach will also employ qualitative techniques: “qualitative research stresses gathering data in “real”, that is, non-experimental settings”.

It is important, therefore, to discuss which methods are appropriate to impact research, taking into account:

a). research design;
b). methods of data collection;
c). approaches to dealing with the data and conducting the analysis.

The similarities and differences between impact research and related concepts, such as evaluation and monitoring, will be discussed with these elements in mind.

3.2 Research Design for Impact Studies

Robson (ibid) contrasts two approaches to research design: fixed vs. flexible designs. In the first he includes traditional experiments and quasi-experimental designs which are common in the “scientific approach”, and in the second he includes case studies, ethnographic studies and grounded theory research. He also looks at designs for specific purposes and in particular he focuses on evaluation research, action research and other approaches which treat practitioners as researchers. This last point is significant as it introduces the role that participants (or stakeholders) play in flexible design research. In this section, we will look at these designs and the relevance to impact research.

The traditional, positivist view of science (typical in the natural sciences in the 20th century) has led to a number of assumptions about the nature of scientific enquiry. Among others, these included the following:

a). objective knowledge (or facts) can be gained from experience or observation;
b). facts are value free;
c). quantitative data is the basis of science;
d). hypotheses are tested against facts;
e). science develops causal laws;
f). explanation is based on relating events to a general law.

These assumptions have been challenged by philosophers and researchers from disciplines outside the natural sciences where are positivist view are still held by some; as a result, the late 20th century saw a shift in the social sciences towards relativism and
constructivism, and a general tendency for pragmatic, less dogmatic approaches to research design.

In some branches of social research, there was a clear rejection of the positivist and associated quantitative orientation, for example, from emerging groups such as feminist and other post-modernist researchers (see Reinharz, 1992; Rosenau, 1992). Others have sought to salvage the best aspects of positivism and to reconcile a positivist legacy in the 21st century with the current attitudes and views of science. In this respect it can be argued that traditional experimentation is not dead in our field, but needs to be appropriately tailored to our needs, often in conjunction with other approaches (see below). Cronbach (1982) talks of a “reconciliation” between “scientistic” and “humanistic” research and welcomes “multiple-method approaches” combining qualitative and quantitative elements.

What has particularly changed is the prevailing view of social phenomena and what this entails for research designs. Social phenomena are no longer treated as facts, but are believed to be constructed through people and social actions. This in turn means that quantification and measurement as a means of collecting and analysing data come into question. A relativist position suggests that alternative ways of looking at the world should be described with an emphasis on the meaning of experience and behaviour in context. The process of conducting research, therefore, generates working hypotheses (not facts) and there is a tendency towards qualitative methodologies (e.g. focusing on the role of language as both object of study and the way that the world is represented and interpreted by the participants in the research. See Fletcher 1996, 114; Robson, 2002, 25). In this respect, aspects of the relativist approach are indeed relevant to researchers investigating test or examination impact.

In the post-positivist era it is recognised that researchers come with their “own baggage” (personal stance, theories, background knowledge, values) and that this will influence what is being observed. This does not mean, however, that objectivity is rejected, rather the researcher needs to be aware of the possible biases that he/she brings to the research process and be prepared to deal with them. Constructivism, a less extreme relativist approach, focuses on the assumption that reality is socially constructed, and researchers must allow for the possibility of multiple realities, rather than objective truth. This again has led to preferences for research methods such as interviews and observations.

Robson himself advocates taking an approach which he calls critical realism (first used by Bhaskar, 1989, and explained more fully by Sayer, 2000, 10-28). He suggests that realism provides a way forward which deals with the problems of positivism but avoids the implied “divorce from science” that relativism brought in. As such he sees this as a “pragmatic
approach” which successfully bridges the divide between the empirical tradition and “less thorough going versions of relativism found in some constructionist approaches”, (Robson, 2002, 42). This point is taken up again in Chapter 7.

3.2.1 Fixed-design: experiments and quasi-experimental designs

What does a realist view imply for social scientists operating in the real world and more specifically for us, are there implications for impact research? The answer to the question lies in the nature of the pragmatic approach itself, or in other words, in establishing “what works for whom and in what contexts”.

If “true experiments” using randomised control trials can be established, they certainly have a role to play in social research, including in test impact studies (for example, to investigate the effects of the test where learners in one group take the test at the end of their language course and learners in another group do not). However, the reality of most educational contexts is that this degree of control is difficult to achieve and that, at best, a “quasi-experimental” design will be feasible (see Robson 2002, 133 on “quasi experiments”). This is a design where an experimental approach is followed but where random assignment to the treatment and comparison groups cannot been achieved. A consequence of this is that interpretation of findings may be more difficult than in traditional experiments, but nevertheless, remain useful. Commonly used pre-experimentation designs (single group – post-test only; post-test only, non-equivalent groups, pre-test/post-test single group design) come into this approach, as do other group and time series designs.

These are all possible designs within a research agenda for test impact. However, because of the complex nature of real world educational contexts, flexible research designs are more likely to be of value in impact studies (as noted by Watanabe). Furthermore, in light of the move to realist thinking in research (as argued by Robson), it is now considered “respectable” to use designs based on methods which generate qualitative data. Importantly it become axiomatic in social research that “theory rather than data or methods to produce those data” must be central to explaining the reality of the world we live in.

3.2.2 Flexible designs

Most flexible designs have their origins in the social rather than the natural sciences and have always relied heavily on qualitative methods. It should be remembered, however, that these designs can incorporate quantitative methods of data collection. The nature of flexible designs means that the design is likely to “emerge” and develop during the data collection phase. This can be contrasted with fixed designs
which typically require a pre-specification prior to data collection which is consistent with strict control of the variables.

Robson again is helpful in summarising the research traditions which are commonly used in “good” flexible designs see Table 3.1.

Table 3.1 Summary of flexible designs

<table>
<thead>
<tr>
<th></th>
<th>Case study</th>
<th>Ethnography</th>
<th>Grounded theory</th>
</tr>
</thead>
<tbody>
<tr>
<td>Focus</td>
<td>Developing an in-depth analysis of a single case or multiple case</td>
<td>Describing and interpreting a cultural and social group</td>
<td>Developing a theory grounded in data from the field</td>
</tr>
<tr>
<td>Discipline origin</td>
<td>Political science, sociology, evaluation, urban studies, other social sciences</td>
<td>Cultural anthropology, sociology</td>
<td>Sociology</td>
</tr>
<tr>
<td>Data collection</td>
<td>Multiple sources – documents, archival records, interviews, observations, physical artefacts</td>
<td>Primarily observations and interviews during extended time in the field</td>
<td>Typically interviews with 20-30 individuals to “saturate” categories and detail a theory</td>
</tr>
<tr>
<td>Data analysis</td>
<td>Description, themes, assertions</td>
<td>Description, analysis, interpretation</td>
<td>Open coding, selective coding, conditional matrix</td>
</tr>
<tr>
<td>Narrative form</td>
<td>In-depth study of “case” or “cases”</td>
<td>Description of the cultural behaviour of the group</td>
<td>Theory or theoretical model</td>
</tr>
</tbody>
</table>

(Based on Robson 2002, 165, abridged from Cresswell 1998, 65)

The use of case studies in social science research has thrown off the image of a weak alternative to scientific methods which had previously been considered more robust. The publications of Robert Yin dating back to the 1970s have been particularly helpful in changing perceptions of the value of case study research and his handbooks have been used extensively in designing case study research over the past fifteen years. Yin (2003b) provides an extensive discussion of the applications of case study research and covers three main types of case study: descriptive case studies; explanatory case studies; and cross-case analysis.

Case study research typically involves multiple methods of data collection, and has been defined by Yin as a “strategy for doing research which involves an empirical investigation of a particular contemporary phenomenon within its real life context using multiple sources of data”, (Yin, 1994, 2003a, 2003b). Weiss (1998: 328) also provides a useful definition when she describes the case study as “a research
strategy that investigates a phenomenon in its natural setting using multiple sources of evidence”.

Ethnography derives from anthropological studies and is far removed from the scientific tradition of experimentation (e.g. anthropological research into exotic cultures in places like New Guinea). Blumer (1969) talks of ethnography “to lift veils” and “to dig deeper” and there are elements of the approach which are of interest when designing projects to investigate micro level impact (especially the effects of tests on individuals or within single institutions).

Qualitative and ethnographic research methods, although closely related are not the same thing. As Lazaraton (1995: 460) notes, qualitative approaches such as classroom observation and case studies found in applied linguistics research do not constitute ethnographic research: “Ethnography requires a deeper and broader philosophical and methodological commitment than does simple participant observation”. In this respect the ethnographic approach is considered by its practitioners to be unlike other forms of research (see Hornberger (2006) for an ethnographer’s view in applied linguistics). Features of the ethnographic approach which are most relevant to our discussion include:

a). the discovery of cultural meaning in the behaviour, actions, events and contexts;
b). the involvement of participants to gain “insider” perspectives;
c). the emergence of research questions during the research;
d). the collection of data over a prolonged period, often in phases.

These points are coherent with constructivist approaches which seek to “illuminate” what happens within contexts where change is taking place.

Grounded theory, like ethnography, can be used in test impact research. Robson (op cit) points out that a case study can be approached ethnographically or an ethnographic study can be approached by means of grounded theory. Some of the features of grounded theory which are useful in designing and carrying out impact studies include the following:

a). procedures for generating theory through the research itself;
b). flexible designs which also require systematic and well co-ordinated activities;
c). explicit procedures for analysing qualitative data.
All three approaches (case study, ethnography and grounded theory) stress the need for rigor in data collection and appropriate ways for data to be stored, accessed and summarised. Multiple techniques will often be used within the same study. Furthermore, the role of the researchers and their research skills are key elements in all three. Robson suggests that researchers need to develop specific skills including: question asking; good listening; adaptiveness and flexibility; grasp of the issues; and lack of bias. This is consistent with establishing “trustworthiness” in flexible designs which rely heavily on the behaviour of participants in the research context rather than the use of “instruments” to collect data.

However, these designs are not without their difficulties, especially threats to the reliability and validity of the research which occur if procedural aspects are not dealt with effectively (e.g. researcher bias, respondent bias, reactivity, etc.). Triangulation is a procedural strategy which can help mitigate such risks and enhance the rigour of the research, and four types are typically found in the literature (see Denzin, 1988):

- data triangulation (more that one method of data collection);
- observer triangulation (more than one observer);
- methodological triangulation (combining qualitative and quantitative methods);
- theory triangulation (using multiple theories or perspectives).

For researchers from a testing or assessment background, the first two are familiar ways to enhance the reliability of human judgements, e.g. for tests of speaking or writing involving human raters.

Several other approaches to verification can be used in addition to triangulation, including: prolonged involvement of researchers, peer debriefing, member checking, negative case analysis and audit trails. However it is in the nature of real world research for “grey areas” to persist, even after triangulation and other checks have taken place, e.g. discrepancies between various sources of data which cannot be reconciled. Therefore iterative cycles of research may need to be carried out if such ambiguities are to be addressed (see below re: action research).
3.4 Special research purposes and impact studies

3.4.1 Evaluation

Test impact studies, like other types of social research, are linked to an “action agenda” and should lead to improvement in the tests or in related contexts of use, (cf. the imperatives of Hughes, Bailey, Bachman and Cambridge ESOL’s maxims). The implication in conducting this kind of research is that something will change as a result, and that the change will make a difference to those involved. Evaluation studies are another instance of this kind of research and it is important to consider in what ways evaluation shares similarities with impact research.

There is an extensive literature on evaluation within the social sciences and especially in education (e.g. Agar, 1986; Cronbach, 1982; Parlett and Hamilton, 1972; Weiss, 1998). In broad terms, an evaluation sets out to assess both the effects and the relative effectiveness of something. This “something” can be an innovation, an intervention, a policy or a programme of some kind. In the literature, there is broad discussion of both the types and purposes of evaluations that may be carried out. (See Table 3.2). Evaluation is similar to other forms of research in that it is principled and systematic and the design, data collection and methods of analysis need to be carefully planned and executed. Designs can be either fixed or flexible depending on the purpose of the evaluation.

Table 3.2 Purposes of an evaluation and possible questions posed by sponsors

<table>
<thead>
<tr>
<th>Purposes of an evaluation and possible questions posed by sponsors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Purpose</td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>To find out if needs are being met</td>
</tr>
<tr>
<td>To improve (something)</td>
</tr>
<tr>
<td>To assess the outcomes</td>
</tr>
<tr>
<td>To find out how something is working (or why it is not working)</td>
</tr>
</tbody>
</table>

(Based on Robson 2000, 10)
An evaluation typically has an avowedly utilitarian function and tends to be both judgemental and action-oriented. Procedural and ethical issues are of high importance and so evaluations need to be conducted fairly, sensitively with appropriate regard to technical adequacy and suitability of the skills of those involved. In this respect, evaluation has aspects in common with impact research.

The question of accountability is another key factor in evaluation, and again, this is likely to be an important consideration in impact research from the perspective of an examinations board. Alderson and Buck (1993: 21), writing about the public examining system in the UK, complained that there had “not yet been a call for public accountability on the part of the examining boards”. Recognition of the need for such accountability led directly to Cambridge ESOL’s policy on impact research which was discussed in Chapters 1 and 2.

A related concept is that of monitoring, which brings with it the common understanding of the term (to monitor is to observe, record or check). Monitoring, as the name suggests, usually involves systematic observations and a reporting mechanism which allows judgments to be made about what has been monitored. Monitoring is not thought to be a research-focused activity and can be distinguished from evaluation by a number of key features.

Weiss (1998: 333) defines monitoring as: “an ongoing assessment of program operations conducted during implementation, usually by sponsors or managers, to assess whether activities are being delivered as planned, are reaching the target populations, and are using resources appropriately”. It seems from this, that monitoring is likely to be conducted by those in authority over a project and who will be held accountable for its success. It tends to be done on behalf of the funders and high-level officials to hold the programme or a project to account, and usually only takes place during the implementation. This allows for intervention to take place to address any problems which are revealed (such as misuse of resources or inability to meet project deadlines). Monitoring, therefore, is akin to what some writers refer to as “process evaluation”. In relation to impact studies, the activity of monitoring may form part of the data collection employed within a wider set of research objectives.
3.4.2 Processes or outcomes?

Because of the instrumental function of evaluation as an activity, the purpose of an evaluation study needs to be explicitly stated at the outset. This raises the question as to whether some types or approaches to evaluation are more suited to certain purposes.

Within educational evaluation studies, Scriven (1967) famously contrasted two types of evaluation, *summative* and *formative* evaluation. Summative evaluation concentrates on the overall effects and effectiveness of the project or programme, whereas formative evaluation is intended to help in the development of that project or programme while it is still going on. The distinction here is principally one of *purpose* and is not absolute. In fact, recent views on this suggest that an evaluation can have *both* a formative and a summative use, whether or not it takes place during or after the event.

Weiss (1998: 330) defines evaluation as the “systematic assessment of the operation and/or outcomes of a program or policy, compared to explicit or implicit standards, in order to contribute to the improvement of the program or policy”. For her both the *operations* and the *outcomes* are important and, as noted above, the results of the evaluation should be improvements in the policy or programme. Cronbach (1982) likewise suggested that “evaluations that focus on outcomes can and should be used formatively”; and Roy (1998:71) also notes that “a study of the product is expected to indicate the pay-off value while a study of the process is expected to indicate the intrinsic values of the programme. Both are needed, however, to find the worth of the programme.”

Like evaluation itself, impact studies which have been discussed in the literature on evaluation tend to focus on both outcomes and processes. There is, however, a tendency for the focus to be on the outcomes, particularly as they effect systems or individuals beyond the project itself and to address longer-term issues such as *sustainability and persistence*. Varghese (1998: 49) agrees with this view and contrasts *impact* studies with *evaluation* studies suggesting that the latter focuses more closely on the immediate objectives of projects rather than the longer-term development.

Weiss (1998: 331), in her book on evaluation, also addresses impact which she defines as “the net effects of a programme”. By this she means the gains achieved for the participants in a programme minus the gains for an equivalent group of non-participants. This is clearly a narrower view based on a form of matched-groups
research design to allow for such comparisons to be made. However, she does expand on this by suggesting that impact also refers to the effects of the programme for the larger community, and that generally speaking, impact in concerned with the *outcomes of the programme*. (1998: 334-335). This latter definition is more relevant to our conceptualisation.

Impact studies in social research generally seem to be concerned with *outcomes and wider effects*. In the field of education where impact studies have been conducted, they have focused on the effects of innovations, such as new teaching programmes, and especially the effects on the people participating in them. This view of impact is borne out in the definition from the UK’s Department for International Development (DfID) in the context of developmental education studies: “impacts (also referred to as effects) may be planned or unplanned; positive or negative; achieved immediately or only after some time; and sustainable or unsustainable. Impacts may be observable/measurable during implementation, at project completion, or only some time after the project has ended. Different impacts may be experienced by different stakeholders”. (See DfID, 1998).

In our discussion of test impact the *observed effects* on the *inner workings* of a testing project and how successfully it was carried out during implementation is of interest, as are the effects of the test on the wider community. The issue of how “different impacts may be experienced by different stakeholders” in the context of innovation and change has received considerable attention in the literature and this has led to approaches to participatory research involving the stakeholders, e.g. action research, and greater emphasis on qualitative methods of data collection.

In the early 1970s Parlett and Hamilton (1972) developed the concept of “illuminative evaluation” with a concern for “description and interpretation rather than measurement and prediction”. This concept was particularly influential in discussions of evaluation techniques throughout the 1970s and 80s and it helped to provide a greater focus on how innovation operates within educational contexts and how change can be effectively managed to achieve the planned objectives (see Murphy and Torrance, eds. 1987: 60-61). In the McKay and Treffgarne collection on *Evaluating Impact* (DfID) there is a section on “participatory approaches” to impact studies which locates impact studies within the realm of innovation and participatory research in general.
3.4.3 Action research

Action research is the last of the “specific purpose” types of research to be reviewed in this section and it usefully serves as a link to the topic of innovation and change management which was an important element of our discussion of impact in Chapter 2.

Lewin (1946) first used the term action research; he was interested in the way we can learn about organisations through trying to bring about change in them. Since the 1980s this has been a popular approach to conducting research in educational settings and the close relationship between the researcher and the object of research fits well with the flexible and more qualitative designs discussed above. A widely adopted version views the research as a cyclical and iterative process (Kemmis and Wilkinson, 1998) which involves a number of phases:

- planning a change;
- observing what happens;
- reflecting on the processes and consequences;
- planning further changes and repeating the cycle.

Wall’s work on innovation and change management in the context of the Sri Lankan study incorporates a model which is not dissimilar to the classic action research cycle.

3.5 Data collection in real world research

Having reviewed the design features of research which might be applicable to test impact studies, this section looks at the implications for data collection. Closely connected to the choice of research design for conducting real world research is the choice of data collection method. As we saw in Robson’s analysis, the choice of research methodology normally involves a consideration of the advantages and feasibilities of both quantitative and qualitative methods of data collection.

Weiss (1998: 335) defines the two general approaches as follows: quantitative research “examines phenomena that can be expressed numerically and analysed statistically”, whereas qualitative research “examines phenomena primarily through words, and tends to focus on dynamics, meaning and context” and in so doing “usually uses observation, interviewing and document reviews to collect data”.

Weiss (1998:284) discusses this when it is applied to evaluation; she summarises the distinction in the following way: “Quantitative evaluators tend to focus on whether and to
what extent change in x causes change in y. Qualitative evaluators tend to be concerned with the process that connects x and y".

In discussing the distinction between the two approaches with reference to research in applied linguistics and TESOL, Lazaraton (2001: 1) cites Larsen-Freeman and Long (1991) and in summarising their analysis, she draws attention to various aspects of quantitative and qualitative research methodologies. This is summarised in Table 3.3 according to a number of continua, and in relation to a number of key features of each approach, including the data collection method itself, the type and focus of data collected and nature of the findings (although the characteristics in the lists are not mutually exclusive).

Table 3.3  Characteristics of quantitative and qualitative research

<table>
<thead>
<tr>
<th>Quantitative Research</th>
<th>Qualitative Research</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controlled</td>
<td>Naturalistic</td>
</tr>
<tr>
<td>Experimental</td>
<td>Observational</td>
</tr>
<tr>
<td>Objective</td>
<td>Subjective</td>
</tr>
<tr>
<td>Inferential</td>
<td>Descriptive</td>
</tr>
<tr>
<td>Outcome-oriented</td>
<td>Process-oriented</td>
</tr>
<tr>
<td>Reliable</td>
<td>Valid</td>
</tr>
<tr>
<td>Particularistic</td>
<td>Holistic</td>
</tr>
<tr>
<td>'hard', 'replicable' data</td>
<td>'real', 'rich', 'deep' data</td>
</tr>
</tbody>
</table>

Lazaraton herself is an advocate of flexible designs using qualitative data collection methods (cf. her work on conversation analysis based on ethnographic principles, Lazaraton, 2002). Similarly, Duff (2002; 2006) notes how qualitative practices can enhance applied linguistics research. These include:

- conducting longitudinal research, when possible;
- eliciting participants’ perspectives on their own behaviour;
- using participants who know each other and have some familiarity with the researcher;
- looking for recurrent patterns across data sets;
- providing methodological, analytical and perspective / epistemological triangulation whenever possible.

Duff also proposes that qualitative studies should seek to yield interesting understandings of local contexts and to examine the interplay of variables and issues to a greater extent than quantitative research typically sets out to do. Again this reiterates some of the points which came out of our discussion of flexible research design and the relevance for impact
research. The implication of practices such as “eliciting participants’ perspectives on their own behaviour” or “using participants who know each other” is that there is need for an appropriate range of data collection methods which can be used by researches in these circumstances to capture such phenomena.

Robson (2000) is helpful in addressing this point and dedicates a chapter on what he calls research “tactics”. In this he lists surveys and questionnaires, interviews and observational methods as the main methods of data collection in real world research. He also has an appendix on the roles of “practitioner researchers”, a concept which, as we have noted, crops up frequently in current literature on social research.

Surveys and questionnaires, interviews and observational methods are all relevant to impact research and were employed in the case studies reported Chapter 4 to 6. The methods used in gathering, analysing and interpreting data are described in relation to each of those three projects, and so we will restrict ourselves here to a brief overview of these methods.

### 3.5.1 Surveys and Questionnaires

Survey methods are typically associated with large-scale data collection exercises requiring large samples and statistical procedures in the analysis. These methods can be useful in impact research to establish current practices (e.g. antecedents in relation to system and user characteristics) before a test or revised assessment practice is introduced, or else they may be part of the regular monitoring of how a test is used and how stakeholders feel about it (i.e. as part of a policy to collect evidence of validity). Construction of the instruments, sampling and analysis are important considerations especially when a wide-scale survey with statistical reliability and validity is required. This is likely to be the case in nationwide or international surveys relating to examination practices. It needs to be recognised, however, that real world research brings its own challenges in this respect and difficulties in obtaining representative samples or adequate response rates are likely to be faced by researchers. While care should be taken to address these issues it may be possible to deal with the data other ways, as Robson points out “even if statistical generalization is not legitimate, it may be feasible to use the kind of theoretical generalisation discussed in the context of flexible design research.” (Robson, 2002, 267).

Small scale surveys and questionnaires can also be used in case studies and in other kinds of narrowly-focused research. It is equally important, however, for care to be given to the construction of such instruments, even in the cases where, for
example, statistical analysis will be inappropriate. Potential problems which can occur with survey research include: a lack of clear aims, implicit rather than explicit theoretical input, un-established causal relationships, inadequate sampling, instruments containing invalid items, lack of triangulation through other data collection methods, and interviewer or researcher effects.

In summary, whether small- or wide-scale, questionnaires can provide insights into the formulation of impact hypotheses and can serve to bring into focus issues which need to be investigated in more depth using other techniques such as interviews or direct observations. The IELTS Impact Projects discussed in Chapter 4 made extensive use of questionnaires which were administered in survey-style data collection exercises.

3.5.2 Interviews

Survey techniques can be used in oral rather than written formats, e.g. for interviews conducted over the phone. Oral approaches involving interview formats are, however, more commonly used in face-to-face contexts, with an interviewer (researcher) asking questions and receiving answers from an individual or a group of respondents.

The extent to which this questioning is structured is a key point, ranging from scripted, fully-structured interviews to unstructured, open-ended ones. The former tend to be used in fixed design research and the latter (semi-structured or unstructured) tend to be considered “qualitative research interviews” as used in flexible designs (King, 1994). An alternative way of thinking about this is to make a distinction between respondent and informant interviews (Powney and Watts, 1987). In all types of interview, however, the role and behaviour of the interviewer is a crucial aspect of the method.

The relationship between the participants is also a consideration and this will vary depending on whether the interviewer is an “insider” or “outsider” to the context, and on whether the interview takes place in a one-to-one or group format. The dynamics of the event depending on these factors can determine the amount and quality of data obtained. Tellis (1997) suggests in relation to case study research that researchers need to “consider not just the voice and perspective of the actors, but also of the relevant groups of actors and the interaction between them”.

A particular instance of the group format is the focus group, much exploited in market research and by pollsters.
3.5.3 Observational methods

Real world research generally involves observation in a general sense. In particular, we are concerned here with the kind of observation carried out directly by human observers, either by being physically present at the time or through the use of recording techniques which produce a permanent record of the event which is to be observed (audio or video). In the case of the first, the observation could be conducted in an unobtrusive way (fly on the wall) or else the observer could become an active participant in the event (either in a flexible or in a structured way).

Participant observation is now widely used in flexible research designs which follow ethnographic principles (e.g. the work of the Chicago school of sociology). A key feature of this approach is that the observer seeks to become a member of the observed group, making this method particularly appropriate for certain kinds of research study: e.g. with small groups, for events which are relatively short, for frequent events, etc. Structured observation usually makes use of some kind of observation schedule and a coding system for capturing aspects of the event which are predetermined to be of interest to the researcher. Because subjectivity and inconsistency can be a problem with observations, it is quite common now to record what happens (using audio and video means) so that the coding and the analysis can take place later, possible with several observers so that inter-observer agreement can be addressed (e.g. to deal with possible threats to consistency and objectivity). When recordings are made as permanent records, these can be subjected to a variety of analytical studies (e.g. transcribed and analysed using discourse or conversation analysis).

3.6 Data collection within impact studies

In our conceptualisation of language test impact, we can choose to focus our attention on micro or macro levels of impact. This could be long-term research covering many facets of a testing system or a one-off study of a single context of test use. In other words, specific studies can be designed to explore the relevant phenomena. The methods of data collection and analysis techniques used will depend on the specific aspects of impact which are focused on within the overall impact model. This can be conceived of as a matrix of possibilities covering macro/micro levels and employing techniques to capture qualitative and quantitative data at both the macro and micro levels. For example, the following areas can be focused on:

a). The context and nature of teaching/learning activity in schools and individual classes.

b). The content and nature of teaching materials related to the test (including the school-based and commercially produced materials).
c). The **views and attitudes** towards the test of major user groups at different phases of the test development and operational cycles, including those who influence policy such as politicians, employers and parents.

d). Characteristics of the test-taking population and the **individual test takers**, including standard demographic data and socio-psychological and strategic characteristics.

It is also useful to place the research into a time sequence - see below for discussion of the timeline.

### 3.6.1 Mixed methods designs

Mixed method designs have become increasingly common in social research and a number of handbooks have been published, for example, Cresswell and Plano Clark, (2007). Cresswell and Plano Clark define mixed methods research as an approach to research design which guides the collection and analysis of both qualitative and quantitative data in many phases in the research process. Their stance is essentially pragmatic and they list four types of mixed method studies which can occur concurrently or sequentially, as shown in Table 3.4.

**Table 3.4 Mixed method designs (based in Cresswell and Plano Clark 2007)**

<table>
<thead>
<tr>
<th>Type</th>
<th>Triangulation</th>
<th>Embedded</th>
<th>Explanatory</th>
<th>Exploratory</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Timeline</strong></td>
<td>Concurrent</td>
<td>Concurrent/ Sequential</td>
<td>Sequential</td>
<td>Sequential</td>
</tr>
<tr>
<td>Both used at the same time</td>
<td>Qualitative within a Quantitative design</td>
<td>Quantitative followed by Qualitative</td>
<td>Sequential followed by Qualitative</td>
<td>Qualitative followed by Quantitative</td>
</tr>
<tr>
<td><strong>Variants</strong></td>
<td>Convergence; Data transformation; Validating quantitative data; Multilevel.</td>
<td>Experimental; Correlational.</td>
<td>Follow up explanations; Participant selection.</td>
<td>Instrument development; Taxonomy development.</td>
</tr>
<tr>
<td><strong>Examples of purposes</strong></td>
<td>Explain why/how data types converge. Gain insights from open ended questions to support survey data. Use qualitative data to explain or expand on experimental outcomes.</td>
<td>Gain insights from a large scale survey by conducting experimental work and seeking to explain the observations by conducting more in-depth studies.</td>
<td>Gain insights into what might be happening on a small scale and then follow up with the collection of data from a wider population to provide confirmation.</td>
<td></td>
</tr>
</tbody>
</table>
The timeline is considered to be important in mixed method designs and so particularly relevant for impact-related research of the kind described in the following chapters. Cresswell and Plano Clark (2007: 111) discuss the sequencing of “phases” in research processes which combine qualitative and quantitative methods. Insights gained in one research project can be followed up in a subsequent phase of research using techniques which were not used in the first instance, e.g. a quantitative survey followed up by interviews with participants. They describe how, in each phase, a number of considerations need to be addressed depending on which design is used and whether the “weighting” is predominantly qualitative or quantitative (e.g. sampling procedures, permissions needed, information to be collected, methods of recording data etc.).

They also discuss longitudinal designs which have a number of anticipated phases over an extended period (12 months or more) and illustrate this with an example of procedures for an exploratory instrument design study. The study has two main phases, the first lasting six months and the second four months. Again this is particularly relevant to the design of impact research.

### 3.6.2 Analysis of data and reporting

From the discussion above, it should be clear that the methodologies employed in test impact research will typically involve both quantitative and qualitative data being collected, often within flexible designs such as those in used in qualitative studies and in relation to many complex variables. This means that in any model of test impact must address the question of how the data is to be stored and analysed.

Weiss (1998:285) suggest a number of “basic analytic strategies” to be used in both quantitative and qualitative data analysis. These are as follows:

- descriptions;
- counting;
- factoring;
- clustering;
- comparing;
- finding commonalities;
- co-variation;
- ruling out rival explanations;
- modelling;
- telling a story.
However, she points to differences in the analysis strategies applied to qualitative and quantitative data because of the “nature of the data and the analytic intent”. (1998: 283). Quantitative research seeks a larger degree of generalisability and tends to use measures yielding numerical values (for surveys and questionnaires) or narrative data which is coded into numerical values that can be analysed and post-validated using statistical methods. As Weiss (1998:284) notes, in quantitative research the data analysis is determined at the beginning of the process. Once the data have been collected quantitative research focuses “on locating significant relationships among them”. The researcher uses statistical techniques to identify associations among the variables and the likelihood that these associations are real (and not merely the result of chance). While not always the case, analysis of quantitative data tends towards the positivistic aim of understanding cause and effect.

As we have already discussed above, qualitative studies tend to provide more holistic and dynamic information. This be data collected from smaller numbers of participants and be less “measurable” (e.g. field notes, narrative accounts of conversations, recordings of interviews or group discussions, documents, video-recordings of classrooms etc.). Data of this kind are less susceptible to statistical analysis and require qualitative interpretations and validation techniques (e.g. through triangulation – noted above).

Along with Lazaraton and Duff, Davis (1995: 446) provides a useful discussion of qualitative theory and methods in applied linguistic research. She suggests the following pattern of data analysis: “… a search for patterns of generalisation across multiple sources of data; and the establishment of credibility through systematic evidence from thick description”.

Davis’ “search for patterns of generalisation” seems to correspond to what Weiss’ calls “finding commonalities”; i.e. a basic strategy in data analysis which entails the identification of trends and significant effects. The investigation of aberrant cases is the corollary to finding commonalities and may be important in qualitative data analysis. Outliers may indicate problems with the data, or unexpected but nevertheless significant findings (e.g. unanticipated effects in the case of impact research).

From Weiss’s list of analytic strategies, some are of particular importance in test impact research, including: comparing, co-variation, modelling; and telling a story.
As noted in Chapter 2, comparison is of particular importance in impact studies. For example, we may be interested in comparing contexts where a particular test is used with contexts where a different test is used; or we may be interested in circumstances where changes are taking place to the examination system and it is important to compare the situation, before, during and after the introduction of a new test. Co-variation is also a key concept in impact studies because it is concerned with the way changes in one phenomenon relate to changes in another. Co-variance analysis can be used to measure such phenomena quantitatively if such data is available.

Modelling and telling the story include drawing inferences and providing a well-balanced final evaluation at the end of project (e.g. concluding which hypotheses were supported and which were not). Modelling includes the pulling together of findings, descriptions and explanations of key processes and outcomes. Telling the story in writing is an important aspect of post-modern approaches to social research; to engage fully with “the story” and its possible meanings, researchers must have understood the purpose, processes and outcomes of the study that they been engaged in. In telling the story, they provide the essential communication of the findings of the study to interested readers. It serves to clarify results, re-emphasises the importance of contextual factors, and above all, it is important in making recommendations for future developments.

Having reviewed the approaches to research design, data collection and analysis which are relevant to test impact studies, in the next section we turn to the use of case studies as meta-data and how this approach will be used as the central method of analysis in this thesis.

3.7 Case studies in social research

Case studies are particularly relevant in that the three impact projects covered in Chapters 4, 5, and 6 are treated as cases studies, and as such are the objective of investigation. The nature of case studies and how the data from them will be used is dealt with below, including a discussion of cases studies as meta-data.

As we have noted above, investigators in the real world are compelled by circumstances to cover complex, contextual conditions (not isolated variables) and to rely on multiple sources of evidence and not single sources. An ability to deal with complexity and contextualisation is at the heart of the case study approach.
Earlier criticisms of case studies (of both the method and the outcomes) arose from questions concerning the objectivity, rigour, generalisability and explanatory power. This led to a stereotypical view of case studies as somehow inferior to other, seemingly more rigorous, approaches which were to be preferred. Yin (2003b, 33) talks of a “caricature” in the social sciences where case study research is characterised as “soft” science. The protagonists of case studies, such as Yin, have argued that this view is simply wrong, and that when properly conceived and executed the case study as a research method has much to commend it.

This view is shared by researchers across a variety of disciplines and case studies are now being widely used in sociology, psychology, political science, social work and education, as well as in government, planning, public administration, public policy and management science (see Yin, ibid). In applied linguistics, case studies are now used in both first and second language acquisition research (Davis, 1995) and have been employed in washback studies discussed in Chapter 2. Duff (2006: 89) provides examples of how case studies and associated ethnographic approaches have sought to emphasise or foreground “contextualization, complexity and credibility” of studies of language education and learning.

The “case” under investigation in social research can be interpreted very widely; it can be an individual, a group, a single social setting, an organisation or indeed to a whole country. To be successful, therefore, a case study must clearly define the case itself and be based on well-articulated research principles, starting with “problem definition”, through the design phase, the data collection and data analysis phases and ending in appropriate reporting of outcomes. Yin makes it clear that a case study is not the “easy option” when it comes to research design and rigour. Yin suggests researchers should emulate the “scientific method” by “posing explicit research questions” and “developing a formal research design” (Yin 2003b, xviii). Robson also points out (2002, xiv) that “at the heart of it is the research question” which must be answerable “within your constraints of time and other resources”.

Cook and Campbell (1979), also argue that case studies should not be seen as examples of “low class experimental design”, but a fundamentally different research strategy with its own legitimate designs which provide an acceptable alternative to experimentation. Sloppy or incompetent applications of the approach which lack reliability and validity in their findings are not acceptable as these are essential elements of all properly conducted research.

Yin, Robson and others such as Miles and Huberman (1994), seek to bridge the positivist/relativist divide and so do not reject the idea of generalisation from single cases to wider contexts. They suggest that it is possible to enhance the generality of findings by
employing careful procedures in conducting the case studies (e.g. careful sampling for representativeness, repetition to create multiple case studies, aberrant case analysis etc.).

Duff, on the other hand, rejects the idea that case studies are concerned with generalisation in the traditional sense, and suggests that the aim is to enhance the potential for careful reasoning and insights, especially through the writing up of the research. She observes (2006, 89) that this works “by providing thick description, credible evidence, thorough data analysis, appropriate representation of contexts and data so that readers can learn from others’ experiences and draw their own conclusions (or inferences).” These points are particularly relevant to the meta-analyses of the three cases reported in this thesis.

Gerring (2007) provides a useful discussion of case study research designs. Table 3.5 based on Gerring’s work illustrates three different designs, each of which has six considerations: population, sample, cases, observations, times and variables.

Table 3.5  Case study designs

<table>
<thead>
<tr>
<th></th>
<th>Design A</th>
<th>Design B</th>
<th>Design C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Populations</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 Samples</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3 Cases</td>
<td>2</td>
<td>40</td>
<td>8</td>
</tr>
<tr>
<td>4 Observations</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>5 Times</td>
<td>1</td>
<td>1</td>
<td>1-5</td>
</tr>
<tr>
<td>6 Variables</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Design A is a case study data set with just two cases and 40 observations made at a single time in relation to three variables. Design B is a cross-case, cross-sectional data set with forty cases, again made at a single time in relation to three variables. Design C is a time-series cross-sectional data set with eight cases and the 40 observations made at five different times (in sequence) in relation to three variables.

Design B (cross-case, cross-sectional) and Design C (time-series, cross-sectional) demonstrate that there are two main types of variation in the design: spatial and temporal. The first has only spatial variation, whereas the second has both spatial and temporal. For both types, the comparison between cases is important (irrespective of whether there are few or many cases), but in the first the comparison is synchronic (a single snapshot) whereas for the second it is diachronic which allows for several snapshots to be taken at different times. This means that comparisons can be made not only between cases within
each snapshot (within time 1 and within time 2), but also between the times when the observations were made (i.e. for each case a comparison between time 1 and time 2).

Gerring (2007: 28) goes on to discuss this are greater length illustrating a covariational typology of research designs as shown in cells 1 to 10 in Table 3.5: There are typically five case study research designs, shown in the shaded cells.

### Table 3.6 Spatial and Temporal variation - reproduced from Gerring

<table>
<thead>
<tr>
<th>Cases</th>
<th>Spatial variation</th>
<th>Temporal variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>None</td>
<td>1. [Logically impossible]</td>
</tr>
<tr>
<td></td>
<td>Within-case</td>
<td>2. Single-case study (diachronic)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>3. Single-case study (synchronic)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4. Single-case study (synchronic + diachronic)</td>
</tr>
<tr>
<td>Several</td>
<td>Cross-case &amp; Within-case</td>
<td>5. Comparative method</td>
</tr>
<tr>
<td></td>
<td>Cross-case</td>
<td>6. Comparative-historical</td>
</tr>
<tr>
<td></td>
<td>Cross-case &amp; Within-case</td>
<td>7. Cross-sectional</td>
</tr>
<tr>
<td>Many</td>
<td>Cross-case &amp; Within-case</td>
<td>8. Time-series cross-sectional</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9. Hierarchical</td>
</tr>
<tr>
<td></td>
<td></td>
<td>10. Hierarchical time-series</td>
</tr>
</tbody>
</table>

Gerring notes that “…the evidentiary basis upon which case studies rely is plural, not singular. Indeed there are five possible styles of covariational evidence is a case study. Usually these are intermingled – different sorts of analysis will be employed at different stages of the analysis – so that it is often difficult to categorize a study as falling into a single cell ….. .”

Examples of case studies being used in impact-related research studies are described in Chapters 4 to 6. For example, in Case 1 in Chapter 4, the importance of case study methods emerged as a major finding and it became apparent that case studies needed to be incorporated into future research for the Cambridge examinations. Chapters 5 and 6 discuss the case studies which were carried out as part of the impact studies being reviewed. In Chapter 6, the project which is analysed is itself a case study carried out in the context of one school in the course of an academic year.

Case studies are clearly one of the most important types of research design available to the impact researcher. By focusing on single cases of individuals, groups or institutions, it is possible to investigate impact at a micro level. It is also possible to use multiple case studies to build up an understanding of impact at the macro level (see for example, Menken (2008: 52) and use of “pyramid project designs” in educational research).
3.7.1 Case studies as meta-data

The three projects which are to be analysed in this thesis are treated as cases of impact studies. Each project is a single case covering an extended period of time with a diachronic perspective – represented by cell 4 in Gerring’s matrix. Each case is investigated post hoc and provides the meta-data and the focus of the meta-analysis.

Meta-analysis, which is widely used in evidence-based medicine for example, dates back to the 1950s. Such analytical techniques have also been introduced into educational research, starting with the work of Glass in the 1970s (see Glass et al, 1981; Hunter & Schmidt, 1990). In 1976, Glass proposed a method to integrate and summarise the findings from any body of existing research studies and he called this meta-analysis or the “analysis of analyses”. “I use it to refer to the statistical analysis of a large collection of results from individual studies for the purpose of integrating the findings. It connotes a rigorous alternative to the casual, narrative discussions of research studies which typify our attempts to make sense of the rapidly expanding research literature”. (Glass, 1976: 3). Glass argued here that literature review should be as systematic as other primary research and should interpret the results of individual studies in the context of the overall distributions of findings body of research. Since that time meta-analysis has become an accepted research tool and is used in a variety of disciplines. While Glass’s original view of meta-analysis was largely statistical, it is now also viewed as an observational study of the existing evidence from previous individual studies.

The steps involved are similar to any other research undertaking: formulation of the problem to be addressed, collection and analysis of the data, and reporting of the results. Researchers engaged in this activity need to write up in advance the research proposal, clearly stating the research questions and objectives, the hypotheses to be tested, the sub-groups of interest, and the proposed methods for extracting and analysing information. In the case of this thesis, this has been achieved in Chapter 2.

3.7.2 The units and structure of the meta-analysis

Yin (2003a, 21) describes five design components that are important for case studies:

a). a study’s questions;
b). its propositions, (if any);
c). its unit(s) of analysis;
d). the logic linking the data to the propositions;
e). the criteria for interpreting the findings.

The “study questions” and relevant “propositions” have been outlined in Chapter 2. The “unit of analysis” for each of the three cases is a completed impact-related project with specified start and end points. All three cases are related to each other (as in a multiple case study design) but are separate in their own aims and outcomes. This allows for a cross-case comparison in drawing conclusions and recommendations.

The designs, data collection methods and types of analysis used in each impact project will be described and evaluated within the context of each case and generalities and insights will be drawn out. This falls into a qualitative approach to analysis and unlike other types of meta-analysis no attempt will be made to use numerical or statistical methods.

According to Yin (ibid) a case study must satisfy at least three key tenets of qualitative methodology: “describing, understanding, and explaining.” Likewise, Feagin, Orum, and Sjoberg (1990) see the goal of reaching a “holistic understanding” of cultural systems as a critical characteristic of case study research. It is perhaps because of this particular goal that case studies are used to investigate complex social phenomena that otherwise would be difficult to describe and analyse, and why it is appropriate to treat the three projects described in the following chapters as cases.

The logic for linking the data to impact propositions and the criteria for interpreting the finding of the meta-analysis are based on the following seven features which emerged in Chapter 2 from the literature on washback and impact research: the test features; the context; the participants; the outcomes; the roles of researcher; the research methods used; and the timeline.

These seven features are used in Chapters 4 to 6 to structure the analysis of each case and to develop the expanded model of impact, starting with Case One in the next chapter.
CHAPTER 4 - Case Study One

The IELTS impact studies – 1995 to 2003

4.1 Introduction to Case Study One

The first case study is the IELTS case, which focuses on the IELTS impact studies set up in the 1990s at the time washback models were beginning to emerge. The data used in the meta-analysis include project documents and records in the archives of Cambridge ESOL, as well as publications available in the public domain.

The IELTS case represents the first iteration of the impact model and is significant for the conceptualisation of impact and for the development of instruments for collecting data; these features are described and illustrated in relation to the IELTS testing system (details of which are summarised in Appendices 4.1 and 4.2).

The case study highlights the ways in which three phases of the IELTS projects contributed to a greater understanding of the impact model. The insights gained are discussed and ways in which they contribute to the developing model of impact and subsequent iterations of the research (described in Chapters 5 and 6) are suggested.

4.1.1 From washback to impact

By the early 1990s, models of washback had been established with a number of features which were central to this conceptualisation. Figure 4.1 and the explanatory notes summarise the main features of washback which had emerged by the time the IELTS impact projects were being planned and carried out.

Figure 4.1 Model - based on the washback models of the 1990s
The test features: Surface features of the test were the main focus, for example item types and formats (e.g. multiple choice). Content validity, especially in terms of authenticity, had become an important issue. In test validation (evidence of validity) the unitary concept of validity was beginning to be adopted, in particular through the influence of Bachman.

The context: There was one main context which was the focus of attention: the school and classroom (i.e. the micro context). The test taking context was typically not separated from the school context where the teaching and learning takes place. Although some wider contextual features (macro context) were starting to be discussed these were not yet a major focus.

The participants: The main participants were taken to be the teacher and the learners in the classroom/school context. There was a limited focus on other participants, such as materials writers, or participants from the wider context (e.g. parents).

The outcomes: Outcomes were seen as changes attributable to the introduction of the test: behaviour of participants – actions, activities, performance in the target language; views and attitudes of participants; decisions to make changes to the curriculum/ syllabus and to develop new materials and methods (products)

The processes involved in bringing about the outcomes were not well understood nor well represented in the model. For example, the processes whereby the test features influenced the content and methods of the teachers were not understood. Some evidence existed to suggest that content but not the teaching methodology was affected but when these effects occurred, how the actually came about and what factors influenced the strength of the effects was not included in the model.

The researcher: The researcher was typically an academic, not usually involved in the test development process as a participant, nor as a participant in the teaching/learning context itself (i.e. an outsider).

Research methods: No clear impact methodology, instrument validation procedures or validated instruments had been established, but qualitative methods were emerging in addition to survey techniques for data collection. The need to problematise washback in terms of hypotheses had been recognised.

Timeline: In the washback model, the timeline was implied but not explicitly focused on. The need for comparative data – before/after – had led to a focus on time-series designs and an appeal to insights from innovation theory. Innovation theory, as discussed in Chapters 2 in relation to Wall’s work using Henrichsen’s hybrid model of diffusion/implementation, suggests that each period of an educational innovation has its own antecedents, processes and consequences. The investigation of “antecedent conditions” are Henrichsen’s version of the baseline study (similar to Saville’s “situational analysis” in the ESOL test development model: Saville, 2003). The consequences, therefore, are the changes which are brought about as a result of the new processes which have been introduced.
4.2 The IELTS case and its historical context

The IELTS impact studies were originally designed to investigate the effects and consequences following the revision of the test in 1995. The intention was to establish a long-term project that would eventually lead to impact research becoming part of the routine validation procedures and was conceived as a longitudinal project which would follow a number of phases. The current author was instrumental in setting up and managing this project over the period under review.

IELTS dates back to the mid-1970s when important changes were taking place in the prevailing approaches to language teaching and testing in Britain. In particular, the communicative approach to language teaching became a major influence at that time, (see Hawkey, 2004). The construct of the original English Language Testing Service (ELTS) and its innovative features provided the basis for generating the initial “impact hypotheses”, as we will see below. The two major revisions of ELTS/IELTS in 1989 and 1995 were also used to inform these hypotheses.

Three main periods of development can be identified. “Period” is used here to refer to the main historical developments of IELTS; “phase” is used to sub-divide the period from 1995 to 2003 which is the main focus in this chapter.

Figure 4.2 IELTS timeline from 1980

<table>
<thead>
<tr>
<th>ELTS</th>
<th>IELTS</th>
<th>IELTS 95</th>
<th>further changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1980</td>
<td>1989</td>
<td>1995</td>
<td></td>
</tr>
</tbody>
</table>

The eight year period following the 1995 revision saw important changes to the testing system and an expansion in uses of the test, reflected in rapidly growing candidate numbers, and the concept of impact became an established feature of the validation programme by 2003.
• Period 1: 1975 to 1989. In this period the original development and implementation of ELTS took place, as well as the ELTS validation project conducted between 1982 and 1986.

• Period 2 overlaps with Period 1 and is from 1986 to 1995. In this period the development and implementation of IELTS took place.

• Period 3 defines the IELTS Case under review here. This period from 1993 onwards again overlaps with Period 2. The development and implementation of IELTS 95 and the setting up and implementation of the IELTS impact projects took place in this third period. Period 3 is sub-divided into three phases: from 1995 to 96; from 1997 to 2000; and from 2000 to 2003.

The development of IELTS over the three periods is described in detail in Appendix 4.2 and only the main points are summarised here.

4.2.1 Period 1 - the early developments – ELTS

The development of ELTS in Period 1 left a lasting imprint and significantly influenced the later developments. Three aspects are particularly noteworthy:

• the birth of the international IELTS partnership – the test sponsors and the roles and responsibilities for test development and administration;
• the development of the ELTS constructs – a task-based, communicative approach;
• main weaknesses emerged – e.g. poorly developed and disseminated specifications.

4.2.2 Period 2 - the ELTS Revision (1986-1995)

The second period is characterised by the review and revision of ELTS leading to the introduction of IELTS in 1989. The following points can be highlighted:

• the ELTS validation study;
• the ELTS revision project – outsourced to C. Alderson and colleagues in Lancaster;
• the establishment of the international partnership with IDP and the change of name to IELTS;
• the simplification of the academic modules and introduction of the GT module;
• the growing impact through increased candidature and wider uses after 1990.
Impact had not been a validation concern in the revision of ELTS. Although positive washback had been claimed for ELTS, no thought had been given to how this might be demonstrated. Moreover, no data were available from the first two periods of development which could be used to address this issue.


Period 3 began in 1993 with a review of IELTS focusing on issues and problems arising from feedback and validation data. The following points are noteworthy:

- the growing influence of Cambridge ESOL’s expertise within the partnership, e.g. the capacity to conduct research and validation studies;
- the introduction of routine validation procedures;
- the further simplification and improvement to systems in the 1995 revision;
- the commitment to regular review/revision based on VRIP approach;
- the establishment of the Joint Research Committee;
- the outlining of an impact research agenda within the research programme;
- the planning of the impact projects.

The wider context in education and society was provided by the views and attitudes which were prevalent in each of the three periods and which led to the test being developed and revised in specific ways. This notion of societal context will be referred to hereafter as the milieu.

In summary the IELTS projects were contextualised within a milieu which can be characterised as follows:

i. the evolution of ELTS/IELTS from the 1970s onwards, including the growing understanding of its innovative role in the assessment of languages for academic purposes;

ii. the influence of validity theory on assessment and especially in the examinations produced by Cambridge ESOL (then UCLES EFL);

iii. the models of washback which were appearing and the ways that they might be applied and extended in relation to IELTS research;

iv. the preferred research approaches and methodologies of those advising on the study of impact at that time.
4.3 Delimiting the IELTS Case

The unit of analysis in this chapter covers the eight year period (Figure 4.2) which began in 1995 as part of the revision process and lasted until 2003. The three phases of the impact studies, the timescale and principle focus of each activity are shown in Table 4.1:

Table 4.1 Phases of IELTS Impact studies 1995 to 2003

<table>
<thead>
<tr>
<th>PHASE</th>
<th>TIMESCALE</th>
<th>FOCUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase 1</td>
<td>1995-1996</td>
<td>The identification of areas to be targeted and the development of instrumentation to collect information which allows impact to be measured</td>
</tr>
<tr>
<td>Phase 2</td>
<td>1997-2000</td>
<td>The validation of the instruments prior to full-scale implementation</td>
</tr>
<tr>
<td>Phase 3</td>
<td>2000-3</td>
<td>Implementation of the instruments as part of a major survey Reporting</td>
</tr>
</tbody>
</table>

This timescale shows the actual time taken. In the early planning (in Phase 1) the subsequent Phases (2 and 3) were designed to have shorter timescales but in practice took longer than originally envisaged.

In 1995, the areas of impact to be targeted were identified and plans to develop instrumentation were set out. This was carried out through collaboration between Cambridge ESOL staff and Lancaster University. The Head of research and validation in Cambridge (Michael Milanovic) invited Charles Alderson (Lancaster University) to address ways in which Cambridge ESOL “might consider researching the impact and washback of the IELTS test”. As a result, Alderson submitted a paper entitled, “Ideas for research into impact, washback and IELTS.” This supplemented Cambridge thinking on the general principles of impact (as part of the VRIP approach) and provided a starting point for planning the IELTS impact studies.

Alderson had gained experience of washback studies through the Sri Lankan Impact Study (see Chapter 2), and the study of the washback of TOEFL on preparation courses (Alderson and Hamp-Lyons, 1996). Drawing on his experiences, he outlined the key concepts and issues and proposed some ideas for research. He emphasised the following points:

- the importance of predicting impact, then comparing predictions with data;
- the need for baseline data (thus implying a longitudinal approach);
- the need for attitudinal data;
the importance of systematising test impact research, including specific impact studies, as a contribution to continuing test revisions.

Importantly, he highlighted the need to distinguish between real impacts and coincidental changes, and pointed out the problem of identifying clear examples of classroom washback. This latter point has continued to be an issue as noted in Chapter 2, and a continuing concern for Alderson himself (Alderson, 2004). It was recognised from the outset that linear relationships would be difficult to prove and that it would be difficult to show that events in and around classrooms are actually effects caused by IELTS rather than the result of other coinciding factors. There was also awareness at the start of the projects that contexts and systems would become increasingly important in the thinking on the emerging impact model.

Alderson’s proposal fitted well with the Cambridge approach and together with the Lancaster group of researchers, four specific project areas were defined which fit within the time-delimited case under review here.

During Phase 1 a range of standardised instruments and procedures were developed to focus on the following aspects of test impact (Table 4.2).

<table>
<thead>
<tr>
<th>Project 1</th>
<th>Project 2</th>
<th>Project 3</th>
<th>Project 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>The content and nature of classroom activity in IELTS-related classes</td>
<td>The content and nature of IELTS teaching materials (including text-books)</td>
<td>The views and attitudes of user groups towards IELTS</td>
<td>The IELTS test-taking population and the use of test results</td>
</tr>
</tbody>
</table>

The first two projects were concerned with impact in the more traditional sense, i.e. the washback effect of the test on teaching/learning. The second two were concerned with the wider impact of the test on other parts of the IELTS assessment system, i.e. within the administrative and academic contexts where IELTS scores are used and on attitudes and behaviours of IELTS stakeholders, including those outside the classroom context.

All four projects fitted into the view of washback/impact which was prevalent at that time based on the Hughes/Bailey model of participants, processes and products. No attempt was made to fit the research into an extended conceptualisation of impact at the macro
level, and there was a clear understanding of practical problems in obtaining baseline data for a test in its third period of evolutionary development.

The project moved forward when a commissioning letter (from the current author - Saville, June; 1995) requested Alderson and his team to carry out the next stage of activity. This was to involve specifying four projects and the delivery of a set of instruments for capturing impact data. The commissioning letter set out the following objectives:

i. The design of a range of instruments and procedures for the collection of classroom data which would be applicable to the study of IELTS and eventually to other Cambridge ESOL exams, including:
   a. questionnaire for teachers;
   b. a questionnaire for students;
   c. an observation schedule for classroom activity;
   d. a procedure for producing summaries of critical classroom activity;
   e. a procedure for setting up focus groups and recording the data (questions, audio recording etc).

ii. An instrument for materials analysis which would focus on IELTS materials which were currently available (to be based on a number of sources including the Colorado experiment, work by post-graduate students at Lancaster and existing checklists, such as the ALTE content analysis checklists;

iii. The design of four questionnaires to collect views and attitudes about IELTS from the 4 main stakeholder groups:
   a. test takers;
   b. EFL teachers;
   c. administrators;
   d. lecturers in receiving institutions;

iv. An instrument for collecting information about the test-taking population and the use of test results to supplement the standard demographic data collected routinely. While this instrument was to focus IELTS candidature it was to be adapted from a range of existing questionnaires (e.g. UCLES/UCLA Language Learner Questionnaires, see Bachman, Cushing-Weigle and Purpura, 1993 – Appendix 4.3).

While these instruments and procedures were intended to be used for the IELTS-specific impact studies, it was also envisaged that it would be possible to adapt them for subsequent uses with other Cambridge exams (cf. impact studies reported in Chapters 5 and 6 below).

The early development work was completed by researchers in both Cambridge and Lancaster (Saville, 1996; Banerjee, 1996; Herington, 1996; Horak, 1996; Winetroube, 1997). The interaction between the two teams was positive and staff in both locations engaged in activities such as “brainstorming” and “prediction” activities which were used to decide the areas of impact which would be most relevant.
The use of external consultants was not restricted to the Lancaster team. During Phase 2, other academic consultants became involved, including Antony Kunnan and James Purpura (ex-students of Bachman at UCLA) who advised on the validation of questionnaires in Projects 3 and 4. Roger Hawkey joined the Cambridge team as an external consultant with specific responsibility for impact research projects and acted as principal coordinator for the Phase 3.

In addition to the Cambridge/Lancaster work, a number of related studies were taking place at the same time and collaboration was encouraged (by Saville) in a deliberate attempt to create a “network” of IELTS-related impact studies. These included:

- the IDP-funded research in New Zealand by Dr John Read and Belinda Hayes on *The Impact of the IELTS Test on Preparation for Academic Study in New Zealand* (a funded project with the IELTS research programme);
- the PhD of Tony Green at the Universities of Reading and Roehampton comparing the impact of IELTS-oriented and pre-sessional English language programmes. This led to Green’s “washback model” referred to in Chapter 2.
- a BC-funded survey of IELTS stakeholders in the UK to review attitudes towards IELTS (i.e. *IELTS Market Survey* 2001, commissioned by Hayde at that time). There was close involvement of Saville and Hawkey in setting up and collecting data for this survey and the results were reported to IELTS Partners.

### 4.4 Phases 1 and 2 - development and validation of the instruments

#### 4.4.1 Project 1

Project 1 was designed to investigate the context and nature of classroom activity in IELTS classes. Four instruments and associated procedures were developed, as follows:

- an observation schedule for classroom activity
- a procedure for producing summaries of classroom activity
- a questionnaire for teachers after teaching an observed lesson
- a questionnaire for students after taking part in an observed lesson.

Early versions of the instruments were submitted to small-scale trials in Lancaster (see Banerjee, 1996). More extensive feedback from experts with research interests in classroom observation was also gathered and analysed in order to produce a “final” version of the instrument for use in 2002 (in Phase 3). The practical difficulties did not emerge in Phase 1, but as the project moved into piloting in real world contexts, the difficulties in collecting classroom data from many contexts around the
world became apparent and were not easily resolved. As a result, this project began to fall behind the others. By considering these difficulties, however, a greater focus on the importance of case study and ethnographic methods emerged in the collection of classroom data. The notion of the “teacher as impact researcher” also began to emerge (cf. action research within language classrooms), in which the teacher is characterised as an “insider” within the context being observed and is “immersed in the research”. This project led to a view that teachers need to develop awareness of the research objectives and to become actively involved in the collection of impact-data as “participant-observers”.

It was not until Phase 3 (the extended data collection survey conducted with many participating IELTS teaching centres) that significant amounts of data were collected in this area. Even though these data were limited they proved to be extremely rich and time-consuming to analyse. In this Phase the relationship between external (outsider) researchers from Cambridge and internal (insider) researchers from within the teaching institutions also became an area of interest. The need for audio/video-based data to provide opportunities for “reflective analysis” of lessons and views of participants (either singly or in focus-groups) led to the development of more explicit procedures for capturing and storing these data (e.g. video data bases which allow for storage and retrieval of lessons or critical incidents on lessons). It became apparent that although these data are difficult to capture and analyse, they can begin to shed light on impact features which cannot be captured through other means (e.g. wide-scale surveys).

4.4.2 Project 2

Project 2 focused on the impact of IELTS on the content and nature of teaching materials, starting with the development of an instrument to collect relevant data. This instrument was designed to evaluate the impact of the test constructs on the content/formats of materials used to prepare learners at the upper-intermediate to advanced levels of proficiency (i.e. required for studying through the medium of English).

The work carried out to design and validate the instrument is described in detail in this section (see also Saville, 2000; Saville and Hawkey, 2004; Hawkey 2006). It is particularly relevant in this chapter as it illustrates important methodological aspects in developing and validating instruments for impact research. The discussion goes into more detail than for the other the projects and focuses on lessons learnt from the validation process.
There were three stages in the development/validation of this particular instrument:

i). the initial design and small-scale piloting of a draft questionnaire by the Lancaster team;

ii). small-scale validation through piloting of the instrument followed by detailed analysis of data and a first revision by the Cambridge team;

iii). further validation through additional "mini-piloting" leading to a second revision by the Cambridge team led by Hawkey.

In 1995, the development of the original instrument for the analysis of textbook materials (abbreviated to IATM) was conducted by Frank Bonkowski in Lancaster. Bonkowski drew on existing instruments for the classifications and lists included in the pilot instrument, (e.g. the ALTE checklists for tasks and examinations, 1995 - see www.alte.org) but extended these to cover contents and methodology of IELTS textbooks and teaching materials. The IATM was, therefore, explicitly designed to draw out the extent to which IELTS and the teaching materials overlapped in terms of the focal constructs (cf. Green’s model of washback).

An understanding of what is covered in IELTS materials such as course books and test preparation books was obtained by reviewing the materials themselves in a systematic way. It was noted that at least three categories of material were used to prepare students for IELTS: practice tests and specimen test materials; course books dedicated to the particular examination; course books not directly linked to the test but whose content and level make them appropriate for test preparation. The IATM was to be a structured questionnaire and the intention was to allow for a course or test preparation book to be “rated”. An analysis of the responses would then lead to insights about the influence of the test on the materials.

The development phase entailed several iterative cycles which were followed in all project areas, including: a literature review; the design of sections and items cross-referenced to the IELTS specifications and the focal constructs; consultation with Cambridge staff and researchers at Lancaster; drafting, piloting and revisions (with as many as six iterations in the drafting stage).

The target users of the IATM were applied linguists and practicing teachers using a particular course book or materials who would rate the book/materials using the instrument. The method of elicitation was designed to operate through open-ended comments, yes/no responses, multiple-choice and scalar items. The pilot IATM had eight parts - four focusing on the target textbook as a whole, four more the skills
components (listening, reading, writing and speaking). These were grouped under the following headings:

a). General Information - Baseline data on the textbook

b). Specific features of the Textbook - Items on organisation, media, support materials, assessment, plus an open general-comment section

c). General Description of Contents - Items on topics, timings, texts, tasks, language system coverage, micro-skills training, test-taking strategies

d). Listening - Sections headed: Input-texts; Speakers; Tasks - Items on listening component text length, authenticity, settings, topics, interaction, inter-relationships, accent, turns, syntax, micro-skills and functions, test techniques and conditions; plus open comment section on listening activity content and methodology

e). Reading - Sections headed: Input-texts; Speakers; Tasks - Items on reading text length, source, authenticity, topics, micro-skills and functions, test techniques and conditions; plus open comment section on reading activity content and methodology

f). Writing - Sections headed: Input; Task; Scoring Criteria - Items on text length, topic, source, exercise task type and length, language system coverage, micro-skills, test techniques and conditions; plus open comment section on writing activity content and methodology

g). Speaking - Sub-sections: Input; Task; Scoring Criteria -Items on interaction, topics, prompt types, exercise tasks, register, exercise conditions, scoring criteria plus, open comment section on speaking activity content and methodology

h). Evaluation of Textbook as a Whole and Summative Evaluation - Items on level, time pressure, task difficulty, test relationship to IELTS plus open comment section on textbook-test relationship.

At the start of Phase 2, validation of all instruments was the main objective and discussions between the Cambridge team and Alderson focused on the need to ensure that data collected would not be invalidated through the use of poorly constructed instruments. Alderson and Banerjee (1996), in a paper commissioned by Cambridge, pointed out that research instruments used in many fields often lack adequate validation; they made a clear distinction between “piloting” (which is often carried out) and “true validation” which, they claimed, is rarely carried out. Many of Alderson and Banerjee's recommendations were implemented, wholly or in part, in
Phase 2 and the use of both quantitative and qualitative methods proved particularly useful, especially for the IATM. The following validation were collected and analysed:

- A review of the clarity and usefulness of the instructions for using the instrument;
- Nine full analyses of textbooks carried out by trained/practising ESOL teachers: four raters used the instrument to evaluate one well-known IELTS-oriented textbook, *Passport to IELTS*; two raters used the instrument to evaluate a preparation book for another examination (TOEFL) namely *Cambridge Preparation for TOEFL*; two raters evaluated a general course book, *Upper Intermediate Matters*; one rater evaluated an IELTS preparation book, *Prepare for IELTS*;
- Completion of the instrument and provision of editing suggestions on the format by two ELT specialists;
- Provision of four data summaries using the instrument by an MA student in Lancaster (Yue W, MA dissertation, 1997);
- A taped discussion and transcription between two raters who had used the instrument to evaluate textbooks (Recorded and transcribed in Yue W. and submitted to Cambridge as an appendix to her dissertation);
- A recorded interview with a written summary of two authors discussing an IELTS-related textbook.

One IELTS preparation textbook was evaluated by four different raters, which proved useful for comparing *rater consistency* in their analyses. This analysis suggested a number of changes should be made.

Four textbooks were covered by one or more ratings; two of the books were explicitly for IELTS students, one was related to another proficiency exam (TOEFL), and one was a general text for upper-intermediate students of English (i.e. not intended specifically for international test preparation). These uses of the instrument provided important comparative data for validating the instrument in terms of convergent and divergent validity and led to suggestions for improvement. Five kinds of improvement to the instrument emerged in relation the items: removal; modification; merging; change of location in the instrument; and supplementation. This also led to rationalising and merging some of the checklists and classifications.

A qualitative example of validation was carried out by Yue Wu Wang in her dissertation. This validation exercise recommended by Alderson and Banerjee (1996:32) suggested that the data should: “provide insights into whether problems
were caused by the instrument and raters’ interpretations of wording or the raters’ interpretation of the textbook”. It was also realised that this approach addressed ease of use and practicality.

Hawkey, in the role of project coordinator wrote a paper entitled *Predicting Impact on Language Learning and the Classroom* (1999), which also provided input to the decisions made for revisions at this stage.

Overall the analyses suggested a need to shorten the questionnaire and improve its practicality through the deletion of redundant items and by sampling units in the textbook rather than by covering all of them systematically. It was also thought possible to reduce the workload of raters by moving objective information to “a baseline data section” at the beginning which could be completed in advance by researchers in Cambridge (i.e. before sending it to the rater). In addition, it was thought necessary to strengthen the coverage of teaching/learning and to include reference to indirect as well as direct impacts of the test on materials. At the end of this phase, a revised instrument was produced with the following changes:

- rationalisation of the whole instrument by seeking summary responses rather than separate responses for each unit in a book;
- the transfer of more baseline data to a section A for completion in advance;
- reduction in the main and sub-sections by merging;
- removal of sub-divisions under each skill;
- rationalisation through merging and deletion of most checklists;
- replacement of the 4-point numerical scales with item-specific categorisations (strongly slightly not really / always, often, occasionally etc);
- deletion of direct references to IELTS;
- addition and modification of items in all sections,
- enhancement of the open-ended comment sections;
- balancing of the number of items across the four skills.

As a result of the changes the revised instrument was considerably shorter than the initial version (14 pages), but nevertheless it remained time-consuming to complete and this was a cause of concern. It was agreed, therefore, that the instrument should be shortened again and that there would be another round of piloting; the aim was to streamline the IATM without losing crucial data. This pilot exercise took the form of focus-group discussion allowing for complex questions which had been raised in the first pilot to be addressed more effectively.
Hawkey arranged for two experienced EFL professionals to use the revised instrument to rate one IELTS-oriented course book and one general course book (not specifically related to IELTS) and then to meet him for an extended discussion (i.e. a full-day session). The procedure used here was an important part of the validation methodology; on arrival at the meeting the two participants were given a summary reiterating the purpose of the project and summarising feedback from previous phases. The summary also focused on the intended outcomes of the day: i.e. to provide feedback regarding corrections, deletions, additions, mergings, reformatting, rewordings or other changes which could lead to modifications before the instrument was used more widely. The discussion between the raters and the coordinator also covered a range of other issues, including the length and accessibility of the instrument for practising teachers and issues related to its complexity, repetition and redundancy. Suggested alterations were discussed and it was possible to agree alternative formulations during the day.

The interactive and immediate nature of the focus group suggested that uncertainties which might arise in completing questionnaires “at a distance” could be avoided by including a meta-commentary on its purpose and component parts. Ten comments were inserted to aid completion and to render the instrument more user-friendly than the earlier versions.

The revised instrument was seven pages long, half the length of the second pilot instrument, but it retained a capacity to elicit detailed information and evaluative comments about course books and other support materials. This third version of the instrument was used to collect data from a sample of teachers selected from IELTS-oriented teaching programs identified by a pre-survey administered in mid-2001 and reported in the final project report (2004) – see below.

4.4.3 Project 3

Project 3 was intended to focus on the views and attitudes of selected IELTS participants and stakeholders. It targeted students and teachers in teaching and testing contexts, as well as stakeholders in wider contexts such as receiving institutions. The three main contexts were:

a). test preparation contexts (e.g. in language centres and other teaching establishments);
b). test taking contexts at the BC or IDP examination centres;
c). contexts of test use, especially in the receiving universities.
Seven questionnaires were developed to explore the *views and attitudes* of the following categories of IELTS stakeholder:

- students preparing for IELTS.
- teachers preparing students for IELTS.
- teachers preparing students for academic study (post-IELTS).
- IELTS administrators.
- admissions officers in receiving institutions.
- students who have taken IELTS.
- academic subject teachers.

The initial development of the questionnaires took place in Lancaster in Phase 1 (see Winetroube, 1997) and piloting was carried out by Cambridge in Phase 2. As a result of the analyses (both quantitative and qualitative), the instruments were re-designed before proceeding with data collection.

The proposals for revising the questionnaires were discussed (end-1999) and, as in the case of the IATM, streamlining and rationalisation were recommended. The revised instrument was a *modular questionnaire* focusing on student characteristics and test attitudes (combining questionnaires 1 and 6 from Project 3 with the test takers characteristics instrument from Project 4).

### 4.4.4 Project Four

Project Four focused on the IELTS test-taking population - the test takers themselves.

Test takers are impacted in many ways, some of which are well known, e.g. if the test taker fails to get the required IELTS band score, his/her chances of getting a place at the chosen university will be affected. Other effects may be less apparent and apart from obvious demographic differences (age, gender, L1 etc), many individual difference characterise the IELTS test takers around the world. In order to understand IELTS candidature better (and to supplement routinely collected data), an in-depth candidate information sheet (CIS) was developed. The initial work was carried out by a graduate student working with Alderson (Herrington, 1996) and later extended by staff in Cambridge.

The draft instrument was adapted from a range of existing questionnaires for learner profiling, including the Language Learning Questionnaires (LLQs) (see Appendix 4.3) which had been designed to elicit information on a number of socio-psychological
and strategic constructs, (including attitude, motivation, effort, anxiety, cognitive and meta-cognitive features), as well as more detailed demographic data.

As we have see have seen from the other projects, the development of instruments tended to follow cycles of procedures which contributed incrementally to their validity and usability, as follows: operationalisation; prediction; brainstorming and review; piloting and trialling; qualitative and quantitative validation; rationalisation. In Phase 2, a single questionnaire was administered to a range of IELTS candidates and the results analysed in Cambridge in collaboration with James Purpura. As a result a revised instrument was produced for use in Phase 3.

As the project progressed, there was a growing realisation that the impact of the test on individuals would need to receive closer attention in future research. Individual characteristics need to be better understood to estimate the extent to which the test has an impact (positive or negative) and in what ways contextual features play a role. It was realised that the IELTS Impact Studies (as reported here) would lead to limited insights in this complex area; additional, in-depth studies would be needed if further progress was to be made.

### 4.5 Phase 3 – the impact survey

Phase 3 was the culmination of the sub-projects, bringing together the revised instruments and procedures into a single data collection exercise to survey a range of IELTS stakeholders in a representative range of contexts world wide. A summary of the instruments which resulted from Phases 2/3 is reported in Hawkey (2006; 67-72) and facsimile versions of the questionnaires are reproduced in the appendices of the same volume. They include:

- A student questionnaire (4 parts).
- A teacher questionnaire for teachers preparing students for IELTS (4 sections).
- The revised IATM.
- An analysis form for observing IELTS-related lessons.
- A receiving institution questionnaire (2 sections).

Hawkey became consultant to the project in autumn 1999 and completed the analysis and rationalisation of the instrumentation in Phase 2. As Project Coordinator, he then coordinated Phase 3 beginning with the IELTS impact pre-survey (May 2001) to update the information on IELTS preparation/test-taking around the world. A questionnaire was sent to over 300 institutions and test-taking venues in 41 countries and nearly 200 (about 65%) responded with the following details:
• the language tests for which courses were provided;
• the courses;
• the textbooks and other materials used;
• the student population (numbers, nationalities, etc);
• the teachers.

The pre-survey established direct links with the centres themselves and this proved helpful in seeking the active participation of the locally-based teachers and administrators in completing the instruments and collecting the observations in the classrooms.

At the time of Phase 3, electronic methods for distributing questionnaires and collecting responses had recently become available to researchers. The technology, however, was not well developed, and while e-mail was used to contact many participating centres, it was decided to use hard copy versions of the questionnaires and other instruments.

In preparation for the main survey, letters and e-mails were sent to 72 potential case study centres, mainly taken from those who responded to the pre-survey (as well as a few others who were contacted later) requesting participation in the survey. As this participation would require considerable time and effort on the part of the students, candidates and teachers, the pre-survey had determined the willingness of the centres to participate.

The main survey sought to collect data on the following aspects at each of the case study centres:

• The profiles of the candidates and teachers before the candidates take IELTS (i.e. in preparation for IELTS).
• The test preparation courses including the teaching materials used.
• The attitudes of candidates towards the test after having taken it.
• The impact of IELTS on the receiving institutions.

Signed consent was received by all students, teachers, textbook evaluators and administrators who took part in the survey and over a period of several months in 2002 responses were received from many diverse contexts across the world:

• 572 candidates who completed questionnaires before and after taking IELTS;
• 83 teachers who completed the teacher questionnaire;
• 43 teachers who completed the instrument for the analysis of textbook materials.
It was not possible to send researchers from Cambridge to interview or record the participants in other countries and so we relied on local support for the data collection. The collection of these data from teaching contexts (classrooms) proved most difficult. In the UK, interviews and focus groups were completed with stakeholders at selected “case study” centres; these involved 120 students, 21 teachers, 15 administrators at receiving institutions. In addition 12 “live” IELTS-preparation classes were video-recorded and analysed.

One of the key points to emerge was the need to develop a more robust storage and retrieval system for observational data of this kind within impact research (see Hawkey, Thompson and Turner, 2006 and 2007).

Insights which emerged from the survey covered the following areas:

- Test taker profiles
- The nature of preparation courses
- Teachers’ perceptions - the influence of IELTS on course content and materials
- Teacher and test taker perceptions of fairness
- Test taker likes and dislikes
- Test taker anxiety
- Likes and dislikes of receiving institutions

A final project report was produced by Hawkey (2004) and the project was incorporated into Hawkey, 2006. The report provided important information for the IELTS partnership on the general impact of IELTS which derive from the multi-faceted data collected in the survey. The growth of candidates had been particularly marked in the 8 year period under consideration and importance of IELTS had certainly increased (with potentially greater impact on more people and more systems).

In summary, Hawkey (2006: 132) concludes that:

a). IELTS was perceived as a communicative test covering 4 skills and this was generally believed to be a positive thing.

b). IELTS was difficult but mainly fair and suitable for undergraduate and graduate level students.

c). The content was appropriate and relevant covering a suitable range of micro-skills.

d). The reading and writing modules were perceived as more difficult, especially because of time-pressure.
e). As a high-stakes test, IELTS caused anxiety but not more than other tests of this kind and to some extent this was seen as motivating.

These points mainly relate to impact as washback, but the process of carrying out the impact studies also suggested ways in which the impact model could be expanded and how future impact studies carried out.

4.7 Analysis of the IELTS case and lessons to be learned

Table 4.3 summarises the IELTS case and addresses the following questions:

What are the main issues which emerge from a meta-analysis of the four projects?
What can be learned from this analysis to inform the development of the impact model?

Table 4.3 Summary of lessons to be learnt from the IELTS case

<table>
<thead>
<tr>
<th>Washback model</th>
<th>Analysis from the IELTS case</th>
<th>Issue for developing the revised and expanded impact model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The test features</strong></td>
<td>IELTS can be seen within the communicative tradition of the 1980s – the ELTS legacy. There was an explicit element of ‘washback by design’ – or in other words, predicted washback based on the test features such as authenticity of content, task types, etc.</td>
<td>The need to relate the features of test impact more explicitly and systematically to the underlying test construct.</td>
</tr>
<tr>
<td></td>
<td>The need to see validity within a unitary framework which includes consequences.</td>
<td>The need to collect evidence of validity over time and to monitor planned versus unplanned outcomes.</td>
</tr>
<tr>
<td></td>
<td>The need to follow all test centres and test sessions through routine (operational) data collection and analyses.</td>
<td>The need to focus on specific interactions between candidates and test features – e.g. greater focus on issues such as the differential functioning of test materials/procedures and potential test bias.</td>
</tr>
<tr>
<td></td>
<td>The need to monitor in what ways and to what extent the test has an influence on candidate performance, both in terms of taking the test and in using the language effectively in target situations.</td>
<td></td>
</tr>
<tr>
<td><strong>The contexts</strong></td>
<td>The worldwide use of the IELTS test made it a unique case in washback research at the time (IELTS had over 300 test taking locations in 120 countries).</td>
<td>The need to focus on contexts other than the teaching/learning context (macro contexts). What are these contexts and what are the relevant contextual variables in each?</td>
</tr>
</tbody>
</table>
The worldwide use introduced many different contextual features related to the test taking (and related teaching/learning) contexts in each country (i.e. locally).

The high-stakes use of test results both locally and in contexts in other countries was an important feature of the research. In other words, the language learning and test taking takes place in one country (often not a country where English is the main language) but the use of the test results takes place in another (UK, Australia, New Zealand etc).

The link between the test development context (involving an examination board and its technical staff) and the test taking and use contexts was novel.

The need to investigate how multiple individual contexts relate to each other. For example, what are the similarities and differences between the IELTS testing centres and their related “feeder testing schools” where the test preparation and other classes take place. What differences exist: a) between test taking contexts within a single country (34 in UK, 25 in Australia, 31 in China etc)? b) between countries?

The need to take into account how the test development and test use contexts relate to the teaching learning context.

<p>| The participants | IELTS contexts allowed for a focus on participants in test taking and test use contexts. The range of participants within these contexts is very varied. The individual test taker variables can be investigated both generally and in specific instances. The IELTS partnership allows for the sponsorship/development roles to be looked at in more detail, both from an historical and forward looking perspective (changing relationships). IELTS – Cambridge ESOL allows for the involvement of the test developer within the planning and conduct of impact research as part of test validation. Relationships with academics and consultants can be explored within the expanded view of the stakeholder constituency. | The need to focus on a broader range of participants, including those outside the teaching/learning context. The need to consider who the key stakeholders are and what roles they perform within the various relevant contexts. The need to see participants not only as group members and also as individuals. The need to monitor individual differences. The need to see the test sponsor and test developer as key stakeholders in conceiving and conducting impact research. The need to consider the role of the researcher in washback/impact research – e.g. as involved participant or distant observer? |
| The outcomes | The IELTS case allows for specific outcomes to be reviewed in the three phases of the impact research, and particularly in relation to the four sub-projects. Analysis of each sub-project provides insights into the research | The need to predict and monitor outcomes in wider range of relevant contexts The need to establish to what extent linear relations can be established. The need to take a dynamic systems approach with understanding of |</p>
<table>
<thead>
<tr>
<th><strong>The researcher</strong></th>
<th>The roles of the test developer as researcher can be looked at in the IELTS case.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The need to consider the role of the researcher in collecting and interpreting impact data e.g. to reconsider the role of researcher as a participant or stakeholder - as a teacher or as the test developer</td>
</tr>
<tr>
<td></td>
<td>The need to consider the relevance of insider/outsider perspectives in impact research.</td>
</tr>
<tr>
<td></td>
<td>The need to involve participants (academics and others) as researchers within their own IELTS-related contexts.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>The research methods</strong></th>
<th>The IELTS case allowed for the test taking and test use contexts to be investigated through collection of routine data; this was a result of the 1995 revision. Summary statistics and evaluations can be produced.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The IELTS research agenda allowed for impact to be prioritised.</td>
</tr>
<tr>
<td></td>
<td>The three phases of the IELTS case had a particular focus on the design, development and validation of instruments.</td>
</tr>
<tr>
<td></td>
<td>The data collection, storage and retrieval of data within a validation programme was a particular concern.</td>
</tr>
<tr>
<td></td>
<td>For an examination board (as part of a University), the ethical and legal consideration were of particular relevance.</td>
</tr>
<tr>
<td></td>
<td>The emergence of real world research methods was of particular relevance given the operational nature of the impact research and the practical considerations which face examination providers (time, cost, human resources etc.).</td>
</tr>
<tr>
<td></td>
<td>The need to use mixed-methods - quantitative/qualitative and multiple sources of data.</td>
</tr>
<tr>
<td></td>
<td>The need to contextualise the research within the social sciences and broader educational research – e.g. the need to use case study approaches.</td>
</tr>
<tr>
<td></td>
<td>The need to focus of the design and validation of appropriate instruments and procedures.</td>
</tr>
<tr>
<td></td>
<td>The need to focus on issues related to data collection and procedures for data storage/retrieval.</td>
</tr>
<tr>
<td></td>
<td>The need to develop practical guidelines and to be aware of ethical considerations (privacy, data protection, IPR etc);</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>The timeline</strong></th>
<th>Within Cambridge ESOL, test development had already been</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The need to monitor changes over time (i.e. the need to determine what conditions</td>
</tr>
</tbody>
</table>
characterised as an iterative process which needed to be modelled and documented effectively.

IELTS 1995 was one of the first cases in Cambridge ESOL.

The 1990s was a period of expansion of the uses of IELTS, in part brought on by the revision of 1995 (see growth of candidate numbers).

The growing importance of the stakes associated with the test was important in the timeline.

The period under review in the IELTS case was an important period in the development of impact as an area of concern in language test validation.

Both within the IELTS projects and elsewhere, insights began to emerge in terms of our understanding of washback in particular (especially the work of Green on the processes of washback). are like at different points before, during and after the innovation).

The need to address issues of forward planning in implementing longitudinal approaches to data collection, both in terms of developing and validating the tests in general, and specifically for conducting the related impact research.

The need to adopt a project-based approach to test development and validation with a detailed and systematic plan for addressing the impact dimension – e.g. to be articulated as an action-oriented plan similar to other aspect of a test development project.

The need to address the problem of collecting useful baseline data – either as the starting point prior to developing a new test, or as a “snap shot” of an existing state of affairs before revisions are made (part of the “situational analysis” - see Saville 2003).

The need to address real world problems of data collection in longitudinal studies (for example the difficulty of controlling local variables to allow like-for-like comparisons to be made over time).

4.6.1 Research design and procedures

The following issues were identified in relation to impact research design and procedures:

- the need for clear objectives to be set, with appropriate designs and explicit impact hypotheses from the outset (coherent with contemporary views of social research discussed in Chapter 3);
- the need for purpose specific instruments and procedures to be developed and validated to ensure that data can be collected effectively within the intended assessment contexts;
- the need for the storage and retrieval of data to be appropriately addressed;
- the need for an approach to the analysis and interpretation of multiple sources of data to be adopted allowing for triangulation of both quantitative and qualitative types of data drawing on best practice in real world research in the social sciences;
- the need for the effective management of impact projects including the adoption of project management and action planning techniques (a practical point which is particularly relevant to examinations board and public bodies involved in assessment);
the need to address legal and ethical issues (e.g. to gain appropriate permissions from participants).

4.6.2 Time-related factors and the iterative cycles

The timeline and longitudinal dimension were important factors in the IELTS case. The timing of the impact research as part of the 1995 revision (15 years after the launch of the original ELTS) and the period over which it was conducted were key features. Reflecting on these points provided important insights into ways in which impact research should be conceived and carried out, and the retrospective analysis revealed the need to take time-related factors into account more effectively, including:

- the point in the test’s life-cycle in which impact research is started, designed and carried out;
- the iterative cycles of the impact research, i.e. in planned phases such that each phase leads into and provides a basis for subsequent activity.

The following seven points can be made about IELTS across the three periods of its history:

i. At the start of the ELTS developments in the late 1970s there was no comprehensive understanding of all facets of validity. There was overemphasis on content validity at the expense of good practice in other ways (i.e. production of test specifications based on sound theoretical constructs; communication with stakeholders; adequate provision for test validation procedures; suitable systems for regular review and revision of the test; etc.). A fuller understanding of the validity developed over a 10 year period and validity issues were properly addressed as part of the second revision, starting in 1993.

ii. Views on face validity and washback were expressed at the beginning of the ELTS developments but there was a poor understanding of the mechanisms which underpin these notions. This was typical of the thinking which accompanied the development of several communicative language tests in the 1980s (see Hawkey, 2004).

iii. Many language teaching professionals expressed strong reactions against discrete point formats which had become common in international tests such as the TOEFL. Discrete point tests of the receptive skills (reading and
listening) and indirect tests of writing (and to some extent speaking) were believed to narrow the curriculum, leading to unproductive cramming and memorised learning of rules (including test taking strategies). ELTS was an ambitious attempt to introduce a communicative approach during the 1980s, but this was achieved at the expense of other assessment considerations. These needed to be addressed after the test was introduced, and regular revision cycles then became a feature of the IELTS system.

iv. Historically speaking, the early 1990s was the earliest period in which the question of the impact of IELTS could be effectively addressed; prior to that the fundamental issues of test validity had not been dealt with. When these issues had received adequate attention, the focus could move on to impact in a more systematic way (as part of VRIP).

v. Within the evolving IELTS partnership, the roles and responsibilities of the partners and the academic community had always been an intriguing feature of the IELTS system. In the early days, the distribution of roles and responsibilities led to a disjunction between the conception/development of the test and the live operation of the test; this separation continued until the early 1990s. In particular, the separation of the academic teams involved in test development and validation from those responsible for administration meant that inadequate attention had been given to routine systems underpinning test production and validation.

vi. Those outside of the test production/administration context were not immersed in the logistic and administrative concerns of the test. When the UCLES EFL team (and especially the Evaluation Unit) had been put in place the necessary leadership and resources existed to move things forward.

vii. The roles of insiders and outsiders in IELTS-related contexts became an important theme and the contexts which are most relevant are:
- the test development/production contexts;
- the test teaching/learning contexts;
- the test taking contexts;
- the test use contexts.
4.6.3 Better understanding of contexts and the roles of stakeholders

The understanding of contexts and the roles of various stakeholders increasingly becomes a major concern in impact research. The contexts chosen can be considered as types of context, given that there is significant variability within each of them, i.e. real world circumstances which exist locally in the numerous different places where the test is taken and used, and there is no unique set of features which defines each one of these. This relationship between the types of context and the many real world instances can be considered a kind of type/token relationship.

It emerged that the issues surrounding the types of context can only be more fully understood if greater insights can be derived from the study of many specific examples (tokens). The type/token analogy leads to a conclusion that “the tokens of context” should be considered as individual cases, each with unique combinations of complex variables – see below. For IELTS, the immediate concerns in 1995 were limited to the contexts which informed the early phases of the impact projects. This thinking was also reflected in various codes of practice available at that time which categorised stakeholders into three groups, assigning responsibility for fair testing to all three. For example the Code of Fair Testing Practice (1988) and the ALTE Code of Practice (1994) both refer to the stakeholders as follows: the test developer (examination provider), the primary users (candidates/test takers), and the secondary users (those who make use of scores to make judgements about the test takers).

In the IELTS case, there is additional complexity which is not covered by the limited number of contexts which were first envisaged, as shown in Figure 4.3. The threefold categorisation does not take into account all the stakeholders involved in and around the test development contexts, nor the potential for other secondary stakeholders around the learning and test taking contexts and in the wider test use contexts. The test taker, for example, is potentially a learner, a candidate and a user of the test scores. In that role he/she comes into contact with English teachers, test administrators and users of the scores, including admissions tutors, registrars and subject teachers in their university of choice.
The types of relationship between the participants in an organisation are represented in Figure 4.4. The examination provider is shown below the central line, including stakeholder groups within and around the institution which produces and administers the exams (the examinations board).

Figure 4.4 Organisational network

Schematic diagram showing the organisational network
The context of the exam provider is represented by the area of the figure below the central line, made up of a number of units which produce and administer the exam (as shown by the circle in the middle). In the case of IELTS this included staff in Cambridge together with the relevant management teams and other participants in the IELTS partner organisations. A network of professional collaborators working with the exam provider is responsible for functions such as writing items and carrying out assessments as oral and writing examiners. In the case of IELTS, the item writers are managed by Cambridge but the examiners are recruited and deployed locally by the other partners.

An examination provider may be required to produce an examination on behalf of a sponsor, such as a government department or regulatory body. In the case of IELTS, the policy group was composed of senior representatives of each partner organisation provides this function. Both the BC and IDP are also officially constituted bodies with a role to play in the assessment of English for academic purposes in UK and Australia. In addition to official regulators and sponsors, exam providers are influenced by professional bodies and pressure groups, introducing considerations related to self-regulation and professional standards (e.g. the ALTE and ILTA codes of practice).

In Figure 4.4, the area above the line represents the type of contexts which exist where the test is taken and where the results are used within a local network of relationships. Often the test centre is a school or college where language learning and test preparation courses are held and where the teachers have a dual function as teachers and as examiners.

The changing environment for English language in the past 15 years has meant that the IELTS stakeholder community has grown and the contexts of use have expanded with the growth of candidature. This has introduced greater diversity in the contexts and range of stakeholders. In addressing this during the IELTS projects, there was a growing recognition of the need to understand the wider issues surrounding the processes which involve many stakeholders beyond the traditional washback environment. The international dimension of IELTS introduced challenges of multiple international contexts and the difficulty of predicting and monitoring the impact in so many diverse contexts. A major conclusion is that, while aggregated survey data is useful for understanding what kinds of variation generally exists, more fine grained approaches, such as case studies, are appropriate to the understanding test impact in local contexts.
The list of participants (or stakeholders) begins to grow when the wider contexts are taken into account and when the events and places are considered in relation to the decisions and actions taken. As the context variable become more complex, so do the stakeholder categories which are relevant to those contexts see Table 4.4.

Table 4.4 People, events, places and actions

<table>
<thead>
<tr>
<th>The people:</th>
</tr>
</thead>
<tbody>
<tr>
<td>The stakeholders or participants the teaching/learning and test taking contexts:</td>
</tr>
<tr>
<td>English Language Learners</td>
</tr>
<tr>
<td>Candidates (learners as test takers)</td>
</tr>
<tr>
<td>Teachers</td>
</tr>
<tr>
<td>Materials writers</td>
</tr>
<tr>
<td>Administrators</td>
</tr>
<tr>
<td>- of courses</td>
</tr>
<tr>
<td>- of exams</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The events and places:</th>
</tr>
</thead>
<tbody>
<tr>
<td>The teaching/learning context (schools, language centres, etc..)</td>
</tr>
<tr>
<td>Events managed by teachers and administrators</td>
</tr>
<tr>
<td>- course design, materials productions, lesson plans</td>
</tr>
<tr>
<td>Events managed by the learners</td>
</tr>
<tr>
<td>- attendance and participation in classes</td>
</tr>
<tr>
<td>- self-study and homework</td>
</tr>
</tbody>
</table>

| The test taking context (test taking venues) |
| Activities for the candidate:               |
| - enrolling                                |
| - sitting the exam                         |
| - getting the results                      |

| Activities for the administrators:         |
| - taking and submitting entries           |
| - handling the test materials             |
| - organising the administration (before the day) |
| - conducting the administration (on the day) |
| - post-exam processing                    |
| - issuing results                         |
| - handling queries                        |

| The other actions taken:                  |
| The decisions taken locally on the basis of the test results |
| By the learner/candidate                   |
| By teachers                                |
| By course designers                        |
| By administrators                          |

| By secondary users elsewhere (e.g. staff in receiving institutions) in contexts where the test results are use |

| The changes brought about as a result of using the test both locally and elsewhere |

For IELTS with its expanding range of uses, Table 4.5 shows the categories of stakeholder which can be identified. We can begin to consider not just general
groups of stakeholders, but also the roles which *individual people* play in specific contexts.

Table 4.5  Broad categories and groups of stakeholders

<table>
<thead>
<tr>
<th>Broad Categories and Groups</th>
<th>Specific groups and names of individuals</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stakeholders within the test production network</strong></td>
<td></td>
</tr>
<tr>
<td>Owners of the test and their governance structure</td>
<td></td>
</tr>
<tr>
<td>IELTS partners</td>
<td></td>
</tr>
<tr>
<td>CEOs and Directors</td>
<td></td>
</tr>
<tr>
<td>Staff within the exam board</td>
<td></td>
</tr>
<tr>
<td>Senior Managers in operational positions</td>
<td></td>
</tr>
<tr>
<td>Test developers and validation staff</td>
<td></td>
</tr>
<tr>
<td>Administrative staff</td>
<td></td>
</tr>
<tr>
<td>Item writers and senior consultants (chairs of teams, chief examiners etc)</td>
<td></td>
</tr>
<tr>
<td>Examiners and markers</td>
<td></td>
</tr>
<tr>
<td>e.g. for speaking and writing tests</td>
<td></td>
</tr>
<tr>
<td>Team leaders and trainers</td>
<td></td>
</tr>
<tr>
<td>Consultants and specialist advisors including academics</td>
<td></td>
</tr>
<tr>
<td><strong>Stakeholders within the test administration network</strong></td>
<td></td>
</tr>
<tr>
<td>Test centre administrators and support staff</td>
<td></td>
</tr>
<tr>
<td>Marketing and Communications staff</td>
<td></td>
</tr>
<tr>
<td><strong>Primary test users – learners/test takers</strong></td>
<td></td>
</tr>
<tr>
<td>Language Learners</td>
<td></td>
</tr>
<tr>
<td>In high school</td>
<td></td>
</tr>
<tr>
<td>In vocational training</td>
<td></td>
</tr>
<tr>
<td>In employment</td>
<td></td>
</tr>
<tr>
<td>As undergraduates at university</td>
<td></td>
</tr>
<tr>
<td>As post-graduates at university</td>
<td></td>
</tr>
<tr>
<td>As pre-sessional students prior to university (under or post graduate)</td>
<td></td>
</tr>
<tr>
<td>As language school students</td>
<td></td>
</tr>
<tr>
<td>Candidates for the test</td>
<td></td>
</tr>
<tr>
<td>Language learners involved in the test taking process</td>
<td></td>
</tr>
<tr>
<td>From the decision to enter to the use of results</td>
<td></td>
</tr>
<tr>
<td><strong>Secondary test users</strong></td>
<td></td>
</tr>
<tr>
<td>Parents and other sponsors of the learners/candidates</td>
<td></td>
</tr>
<tr>
<td>Language Teachers</td>
<td></td>
</tr>
<tr>
<td>In private (language) schools</td>
<td></td>
</tr>
<tr>
<td>In state-funded schools</td>
<td></td>
</tr>
<tr>
<td>In language centres at universities</td>
<td></td>
</tr>
<tr>
<td>Private tutors</td>
<td></td>
</tr>
<tr>
<td>School Managers</td>
<td></td>
</tr>
<tr>
<td>School owners</td>
<td></td>
</tr>
<tr>
<td>Head teachers</td>
<td></td>
</tr>
<tr>
<td>Specialist course designers and materials producers</td>
<td></td>
</tr>
<tr>
<td>Administrative staff (secretaries, clerks, etc.)</td>
<td></td>
</tr>
<tr>
<td><strong>Other Users of test scores</strong></td>
<td></td>
</tr>
<tr>
<td>Academic bodies (universities and colleges where English is the medium of instruction)</td>
<td></td>
</tr>
<tr>
<td>Admissions tutors</td>
<td></td>
</tr>
<tr>
<td>Registrars</td>
<td></td>
</tr>
<tr>
<td>Academic tutors and course directors</td>
<td></td>
</tr>
<tr>
<td>Vice-chancellors and Deans of universities</td>
<td></td>
</tr>
<tr>
<td>Marketing and recruitment staff</td>
<td></td>
</tr>
<tr>
<td>Government agencies</td>
<td></td>
</tr>
<tr>
<td>Immigration departments – Home Office, DIMA, CIC etc.</td>
<td></td>
</tr>
</tbody>
</table>
The broad categories should be broken down so that the views and attitudes of specific groups and individuals can be captured and analysed (i.e. to fill in the right hand column of Table 4.5).

This analysis of the contextual parameters and organisational systems suggests that language assessment must be considered as one inter-related component within a number of other systems (i.e. interdependent elements forming a complex whole - see Larsen-Freeman and Cameron, 2008: 26). The complexity of interactions between contextual variables in local test taking and test use contexts means that new systems are likely to be created. Some of these will not have been anticipated and so their desirability or otherwise will not be known by the test producer. Over time, these systems may themselves evolve and potentially diverge from each other so that separate sub-systems will be created.

The contexts of learning and test preparation which develop in relation to a specific testing system are likely to be particularly variable for international examinations. The participants in each local context bring their own knowledge, views and behaviours into that context. Access to relevant information, training and materials may be variable and depend on a range of contextual factors, including the efficiency of the test provider in communicating information about the test to stakeholder groups in an
equitable way. The newly emerging systems will be difficult to predict in advance and so need to be monitored as they evolve over time.

The review of the IELTS history provides support for a conceptual perspective which is “emergentist” and consistent with the view that validity itself is an emergent property of assessment within dynamic educational systems. Effective planning and test design build the intrinsic properties of the test (content, formats, procedures, etc.) but the overall validity of the test can only “emerge” when the test is functioning in real contexts. In order to address validity of this kind, real world contexts of test production, test-related teaching and learning, test taking and test use need to be better understood. In other words, the validation programmes of examinations boards need to collect context-specific data to investigate whether the test is “fit for its intended purpose(s)” and is used in ways which were anticipated or which have subsequently been shown to be appropriate.

4.7 Conclusions and implications for the revised impact model

The basic washback model illustrated in Figure 4.1 can now be extended based on the insights from the IELTS case:

i). The test features.

It is not adequate to focus on surface features of the test. Construct based approaches to test design and development need to build in test taker and contextual models so that part of the validity argument includes the expectation of impact at both individual and systemic level. In terms of test validation and evidence of validity, the unitary concept of validity needs to be adopted and extended (e.g. as suggested by Mislevy et al) so that evidence can be accrued when a test is being used. Facilities for appropriate data collection and analysis needs to be part of the test design.

ii). The context.

While the focus of attention needs to remain on schools and classrooms (i.e. the micro context), it also needs to be extended to wider “macro” contexts. Researchers need to take into account the fact that assessment has a social dimension and that language tests fit within educational systems, which can be characterised as complex and dynamic. The nature of such systems means that it is not apposite to seek linear or causal relationships. This understanding needs to influence the types of research methods which are used to investigate impact.
iii). The participants.
In the expanded model a much wider range of participants needs to be taken into account which goes beyond the classroom/school context. The interaction between participants needs to be seen within a social model within which the validity of a test is constructed. This will allow for the possibility of impact being positive and as predicted for some participants and negative and unpredicted for others. Understanding how these mechanisms work should be part of the test validation process.

iv). The outcomes.
The introduction of a new or revised test should have intended outcomes which are achievable by designing the testing system effectively. These objectives should be based on a clear understanding of the current situation (the existing and developing milieu), as well as clearly understood and well-defined processes and procedures to be introduced as part of the innovation (i.e. the anticipated change process). However, the outcomes in terms of the consequences – intended and unintended - cannot be wholly predicted; this is because the validity of the examination can only emerge when it has been “installed” into its real-life contexts of use. These contexts will not be uniform and will be constantly changing as a result of socio-political and other factors. Outcomes and changes which might be attributable to the introduction of the test need to be monitored in relation to: the behaviour of the many and various participants (actions/activities, etc); the views and attitudes of those participants; the decisions which are taken leading to changes in educational and societal contexts where the test plays a role, including the impact on curriculum/syllabus, new materials and other products; the processes involved in helping to bringing about the intended outcomes need to be well represented and focused on in the expanded model. This approach will require in-depth study and analysis of particular cases as well as aggregated information from many contexts.

v). The roles of researchers.
The role of the test provider and other participants as co-researchers is an important feature of the expanded model. While it remains important to involve academics (as in the IELTS projects), it is equally important to include those involved in the test development process as participants (i.e. ab initio), as well as those involved in the teaching/learning contexts when the test “goes live” (i.e. as insiders). A greater understanding of insider/outsider roles will become important as the impact model develops.
vi). Research methods.

The need to problematise impact in terms of hypotheses linked to aspects of validity (especially the focal constructs) is a critical feature of the expanded model. The IELTS case provided insights into impact methodology, instrument design and validation procedures. Mixed methods with a greater use of qualitative techniques emerged as essential tools to supplement more traditional survey techniques. In-depth case studies are likely to be required to provide richer insights into the complex interactions between individual and systemic variables in different contexts where tests are used. Impact researchers need to be prepared to develop or adapt appropriate instruments, and they also need to consider improved ways to store and access the complex data which are generated by large-scale, mixed method designs. Ways need to be found to make such data easily accessible and susceptible to analysis. Further insights and guidance are likely to be required from disciplines such as innovation theory and other branches of social science research.

vii). Timeline.

In the expanded Impact model, the timeline needs to be explicitly focused on, suggesting a longitudinal approach involving iterative cycles. The need for comparative data is also likely to lead to a greater use of time-series designs.

In light of these findings the model can be revised along the following lines as shown in Figure 4.5.

The IELTS case confirmed the importance of the four key points which Alderson introduced at the start: the importance of predicting impact, then comparing predictions with data; the need for baseline data; the need for attitudinal data; the importance of systematising test impact research as a contribution to continuing test revisions, including specific impact studies. An additional dimension represented in Figure 4.5 is that of iteration and the longitudinal nature of impact research. The IELTS case demonstrates clearly how the potential for various kinds of impact emerge along with the evolution of the testing system itself over time. In particular, the growth of IELTS candidates and the increase in stakeholders of various kinds meant that many potential effects and consequences also increased proportionately.
In the next chapter, the second case is reviewed.
CHAPTER 5 - Case Study Two

The Progetto Lingua 2000 Impact Study

5.1 Introduction to Case Study Two

The second case study is the next iteration in the development of the impact model. Dating from the late 1990s, the research under review is referred to hereafter as the Progetto Lingue 2000 (PL2000) Impact Study.

This project explored the application of the impact research procedures within a single state educational context, namely Italy. It was of particular relevance to the emerging model as it covered both macro and micro aspects of impact. The macro context is provided by the Italian educational system as a whole, and more specifically, the specific reforms which were taking place during the 1990s of which the PL2000 was a part. Against this backdrop, a number of micro contexts, represented by a seven case study state schools, provided the major focus of the impact study in question. The data collected from participants in these schools in the form of questionnaires and interviews are an important feature of this project.

The PL2000 was an Italian government project to improve language education throughout the country which included external assessment and certification as an integral component, including Cambridge ESOL examinations. This was an ambitious reform project aimed at introducing some quite radical innovations in foreign language education in schools to meet the communication and certification needs of students. These needs were defined in terms of the Council of Europe’s Common European Framework of Reference for Languages (CEFR, 2001) and this represented one of the first attempts in Europe to bring about curriculum reform based on the CEFR and its level system.

The key element of the PL2000 was that extra-curricular language courses were to be delivered to homogeneous groups of learners in modules making use of contemporary educational technologies. As far as language assessment was concerned, an unprecedented policy was funding to encourage the use of external certification of learners’ language proficiency using the examinations of internationally recognised providers (e.g. Cambridge exams for English, and similar ones for French, German and Spanish from the international providers for those languages).
5.1.1 The PL 2000 Impact Study

During the 2001-2002 school year, with the encouragement of the Ministry of Education (then known as the Ministero della Pubblica Istruzione – MPI), the author and other researchers in Cambridge and Italy participated in the planning and implementation of a limited study of the impact of the PL2000; this involved the collection of data from a range of stakeholders in seven case study schools, including students, parents, teachers, heads, government officials and language testers. This context is represented schematically in Figure 5.1:

The impact study itself was only a small element of the overall PL2000 and was carried out by an external agency with the approval of the Italian authorities. The willingness of the Ministry officials to permit the study was evidence of an interest in obtaining feedback from a variety of sources, and not just from the Ministry’s own monitoring systems (referred to as monitoraggio in Ministerial documents – see Appendix 5.4). This is usually carried out by INVALSI – the Istituto Nazionale per la Valutazione del Sistema dell’Istruzione – the national agency for educational evaluation. The information from the impact study was made available to Ministry staff in order to contribute to their own evaluation of the project and for the development of future initiatives in school foreign language education in Italy. Although small-scale, the PL2000 Impact study produced findings in many of the areas specified by the Ministry at the launch of the project for its own monitoring requirements, including information on the students involved in participating schools.
and analyses of the teaching and learning contexts covering content and methodology, use of new media and the success of the learners in reaching the stated objectives linked to the CEFR.

The PL2000 Impact Study has contributed to the development of the impact model in various ways, but in particular through the focus on a single national context, and on the relationships between key stakeholders and language teaching/assessment systems. It is also significant that the study provided an opportunity to investigate the innovative processes which were at work in implementing major, nationwide reforms in the teaching of languages in school settings (cf. the work of Wall in Sri Lanka and Cheng in Hong Kong cited above in Chapter 2).

In summary this study provided data on:

- attitudes to external English language exams at different levels of the CEFR (a major innovation in the Italian context) collected from students, teachers, school heads and parents;
- the relationships between teaching, learning and assessment within the context of a major reform project, including the perceptions of students and teachers, as well as observational data collected in classrooms in case study schools;
- comparisons between the examination performance of students attending schools taking part in the PL2000 and those in schools not involved with the Project.

The use of external certification emerged as a major impact of the project. The analysis showed that students, teachers, Heads and parents were motivated by the external examinations and that course design, teaching methods and materials were influenced in various ways by the PL2000 policies including the examinations.

### 5.2 Delimiting the PL2000 Case

In the PL2000 Case the unit of analysis is a period of five years from the beginning of 2000 to 2005. This period is set against the wider socio-political milieu and the programme of reforms within the Italian educational system which date from the early 1990s, and especially from around 1995/6 when a new administration came into government. The impact study itself covers a period of about three years and can be divided into three research phases starting in year 2000:
Phase 1 - August 2000 to August 2001
    Planning and instrument development
Phase 2 - October 2001 to May 2002
    Data Collection
Phase 3 - June 2002 to March 2003
    Analysis, reporting and evaluation

See Hawkey 2006: 49, figure 3.3 - project plan model.

Aspects of the PL2000 Impact Study were closely examined by Hawkey between 2003 and 2005 in preparation for his volume on impact (2006, chapter 7). Some effects of the PL2000 continued after 2005 but are considered out of scope and do not form part of the analysis of this case.

The Italian educational system (Vertecchi, 2001) and the reforms of the 1990s are described in Appendices 5.1 to 5.3. Key features of the system are set out allowing comparison with other national systems, the approach to foreign language study in the 1990s, and details of the reforms, especially the Berlinguer Reform from 1997 onwards. The importance of the political dimension is highlighted and the changes of government are summarised.

5.3     Main objectives of the Progetto Lingue 2000

The PL2000 was one aspect of policy related to the autonomy of schools (autonomia) which had begun to influence educational structures and funding arrangements by year 2000. In particular, the PL2000 grew out of Law 440 (December, 1997) and government action carried out by the Ministry of Education under Minister Berlinguer. The mechanisms for implementation depended on key individuals within the Directorate General and especially the “nucleus” for PL2000 under Ispettore R. Sanzo who played a pivotal role in the Ministry leading up to and during the implementation of the PL2000 (1999 to 2002). Sanzo skilfully led the small team which launched the project and provided central guidance and coordination throughout Italy.

Details of the PL2000 are outlined in Appendices 5.4 to 5.6, including original documentation in Italian. The systemic and integrated nature of the innovations envisaged by the Ministry was an important and unusual feature of the project. The assessment dimension and the potential impact of the examinations therefore had to be evaluated in relation to the other innovative “segments” of the overall project. Figures 5.2 and 5.3 summarise how the various segments of the project were designed to fit together to support the anticipated educational goals.
As Figure 5.3 shows, there were four key segments of the project within a unified attempt to achieve a number of educational goals related to language improvement:

i). Overarching curriculum guidelines - based on CEFR including the levels and the communicative approach;
ii). Central and local resources and support for teachers and learners;
iii). Teacher development and support;
iv). Assessment of learning outcomes including the external certification of proficiency.
These goals targeting language education should, in turn, be seen within the overall context of the reforms of the educational system, especially school autonomy (“autonomia scolastica”).

By setting up the PL2000 in this way, the Ministry chose to bring the resources of the examination providers and cultural institutes to bear to support their aims. The long-standing relationship with cultural bodies and experience in previous projects meant that the Ministry anticipated a positive response to the PL2000, especially from the French, German and Spanish institutes, as provision of language examinations forms part of their cultural mission. The PL2000 was an opportunity for them to extend their activities and increase support for learners of their languages in Italian schools.

It was not as clear cut for English; the British Council (BC) provided teaching and cultural support but not its own English examinations. The BC provided a list of boards which were accredited by the Qualifications and Curriculum Authority (QCA) to offer publicly funded exams in the UK. Each English board had its own centre network (which included some BC offices in the case of Cambridge ESOL) but they had no direct relationship with the Italian Ministry. One of the first impacts of the PL2000 was the increase in promotional activity by the English boards in order to provide relevant information to the schools and to differentiate the various kinds of examination on offer. This included new websites and informational materials in Italian aimed at teachers in state schools. There were also attempts to ensure that the demand for English would not put too much strain on local administration systems; initially these systems were not equipped to work in the state sector and the capacity to provide examination centres and examiners for speaking tests throughout the country needed to be built up.

Of the two English boards, Cambridge and Trinity College London (TCL) had significant numbers of candidates prior to the PL2000. Cambridge had already worked with universities to offer PET (B1 level) as part of the Campus Project (see CRUI website: www.fondazionecrui.it/CRUI_Foundation) and had seen a growth in candidature in the mid 1990s. The PL2000, however, was to have a much greater impact on candidate numbers and geographical distribution.

A key point, which was left to the examinations board to deal with in their interaction with the schools, was the validity of the link to the levels of the CEFR and the extent to which the examinations matched the PL2000 aim of implementing a communicative approach covering all skills (Reading, Listening, Writing, Spoken production and Spoken interaction).
In developing impact hypotheses in the context of PL2000, the Cambridge team needed to demonstrate that the core constructs of the exams overlapped sufficiently with the educational goals of the PL2000. The Cambridge main suite exams (KET, PET, FCE, CAE, CPE) and other certificated exams which were offered (YLE, BEC, CELS), had been successfully used in other parts of the World with school-age learners and positive washback effects had been claimed in those contexts. The IELTS Impact studies were still going on and were already providing the basis for investigating these effects, and of course, the approach to validation more generally, incorporated the concept of impact, as noted in Chapter 1.

5.4 The PL2000 Impact Study

The decision to adapt aspects of the IELTS impact projects to investigate the impact of the PL2000 was taken soon after the first agreement (protocolo) with the examinations boards was signed by the Ministry in January 2000 (i.e. in Spring/Summer 2000). The opportunity to work within the context of the Italian state sector offered potential benefits for the researchers in Cambridge, in particular an opportunity to relate the macro educational context, including specific, government-backed innovations, to the micro contexts of individual schools and classrooms where the teaching and assessment was to take place, (including though not exclusively, the use of Cambridge ESOL examinations). The potential for positive washback on teaching/learning in schools, and the wider impact of the PL2000 on attitudes to language learning and assessment were recognised and the PL2000 offered an opportunity to investigate these effects.

The PL2000 Impact Study action plan was discussed between staff and consultants in Cambridge and received active support from the Chief Executive; the final draft was agreed in August 2000 and Roger Hawkey, the external consultant for the IELTS projects, took a coordinating role. It was, therefore, implemented with Hawkey as Co-ordinator working with the current author (Director, Research and Validation) and Liam Vint, the Cambridge Development Manager in Italy. Additional support was provided by relevant departments in Cambridge ESOL, by staff in Italy and by several Local Secretaries (especially those in Novara, Rome and Taranto where the case study schools were located - see below).

The investigation of impact in Italian schools could not be carried out without informing the education authorities and seeking the necessary permissions. It was a priority, therefore, to secure the support of the Ministry before proceeding. This was done by contacting Ispettore Sanzo (referred to above), the coordinator of the PL2000 at the Ministry. A preliminary meeting took place in September 2000 between Sanzo and Cambridge representatives to outline the objectives and to seek approval. Sanzo was supportive and agreed that it could go ahead, provided that permissions were obtained at local levels. In addition to providing
guidance from the outset, Sanzo was interviewed later as part of the impact study (October 2001) and he remained enthusiastic and supportive. Other Ministry officials also showed interest in the study when the final report was presented (January 2003) to Silvio Criscuoli, the Direttore Generale delle Ordinamento Scolastico, and Giuseppe Cosentino, the Direttore Generale delle Formazione del Personale della Scuola, (previously Sanzo’s immediate superior - 1999 to 2001).

The role of the central Ministry officials and other regional protagonists (illustrated by the Ministry in Lombardy in Appendix 5.11) proved to be a noteworthy feature of the impact study and informative in developing the impact model (cf. Henrichsen’s factors within the “intended user system” which hinder or facilitate change, as discussed by Wall, 1993).

Having set out the broad objectives and received the go ahead, the first step was to refine the research questions and related impact hypotheses and to engage in the necessary project planning based on lessons learnt in the IELTS projects.

5.4.1 The Research Questions

The overall aim of the PL2000 impact study was to identify, describe and explain the impacts of the PL2000 programme on a range of participants within the Italian context. From the Cambridge perspective and in light of the developing impact model around IELTS, an emphasis on the assessment dimension was predicted. More specifically, five main research questions were specified, as follows:

i). What impacts did the PL2000 have on the pedagogy, materials, and media for language teaching and learning?

ii). What impacts did changes in language teaching and learning pedagogy, materials, and media have on the performance and attitudes of the students and their teachers?

iii). What impacts did the PL2000 have on language assessment?

iv). What impacts did the PL2000 have on educational managers, including Head teachers?

v). What impacts did the PL2000 have on support for teachers in-service, resource centre developments, and test-awareness programmes?

Within these broad questions, the role that Cambridge exams played as part of the assessment provision, and the extent to which specific features of the Cambridge approach introduced observable effects and consequences were, of course, particular concerns.
Bailey’s washback model (1996) can be adapted to account for the PL2000 case, as shown in Figure 5.4. The blue box represents the innovative impetus from outside the schools related specifically to the PL2000 policies and actions, and the added red box represents the change processes within the school and its local/regional environment which were necessary to bring about the intended improvements. The CEFR principles and the external certification are represented as key drivers for change from the macro context. The red box, initially empty, is where data from the micro context within the schools was needed to provide insights into what was happening. This was an important part of the impact study and an opportunity to seek insights into the interaction between factors in the macro and micro contexts.

Figure 5.4 Extended version of Bailey, 1996

The existence of the CEFR as a shared reference document meant that the participants were able to address important concerns, such as the definition of proficiency levels and the implementation of a communicative approach, within a common framework. This facilitated communication and understanding of key concepts which are crucial in the “diffusion of innovation”. The existence of the common framework, albeit innovative and only partially understood by many participants, meant that all segments of the overall project could be linked; for example, materials and curricular documents followed the same guidelines as the teachers, the teacher trainers, the examination providers and the researchers.
5.4.2 Design and Data Collection

The PL2000 impact study posed a particular problem for the Cambridge team in terms of the scope and availability of resources. National agencies for educational research (e.g. INVALSI in Italy or NFER in UK.) normally address research questions of this kind by undertaking large-scale surveys on behalf of the national authorities. In so doing, they base their research on the principles of survey design (e.g. sampling and multi-level analysis techniques) and can bring to bear significant resources. The PL2000 impact study, however, was to be conducted by an “outside” organisation with a limited mandate to intervene in educational institutions where the data were to be collected, and so practical considerations and resource limitations were imposed on the project from the outset. The scope did not allow for comprehensive survey data to be collected, although quantitative data were to be collected wherever possible as part of Cambridge’s routine validation procedures. “Baseline data” consisted of historical data on Italian candidates’ performance in Cambridge exams which could be compared with PL2000 data, as well as general demographic data for all candidates taking the examinations. In line with the emphasis on mixed designs in impact research, it was decided that significant insights would be gained from “in-depth” data collection and analysis procedures; data would be collected from a number of case study schools and the focus would be qualitative (i.e. case study based). The quantitative data coming from the routine validation procedures would provide additional contextual information for interpreting the qualitative investigations conducted in case study schools and provide opportunities for triangulation (noted in Chapter 3).

In delimiting the scope of the project it was decided that the impact study would cover one school year (2001-2) using an opportunity sample of case study schools from three Italian regions. Permission was obtained from Head teachers and support was provided by English teachers and PL2000 “referenti” at the following seven schools:

- Scuola Elementare Statale “Maraschi”, Oleggio
- Istituto Comprehnsivo Statale “Verjus”, Oleggio
- Istituto Tecnico Commerciale Statale “Alessandro Bermani”, Novaro
- Scuola Media Statale “Circonvallazione Tuscolana”, Rome
- Liceo Ginnasio Statale “Virgilio”, Rome
- Liceo Ginnasio “Aristosseno”, Taranto
- Scuola Media Statale “Leonida da Taranto”.

In addition, the teachers of English from the Liceo Scientifico Statale “Alessandro Antonelli” provided input. The three geographical locations are represented
schematically in Figure 5.5.

Figure 5.5 Geographical locations of case study schools

These schools represented a useful range of schools across the cycles (elementary, middle and upper secondary) with several post-16 types of school included (Liceo/Istituto tecnico) and coverage of three geographical regions (urban areas only): north west - Oleggio, Novara; centre – Rome; south east – Tarranto (see school profiles in Appendix 5.8).

The project introduced a longitudinal dimension which had not been present in the IELTS projects: a repeated ‘snap-shot’ design was used, with the data collected on two visits to the selected schools at key points in the Italian scholastic year – October 2001 and then in April 2002. These visits provided in-depth evidence of changes in views, attitudes, approaches and experiences, as well as opportunities for verifying initial findings (see Hawkey 2006: 49-50, figure 3.4 – chronology and data collection). The case study model also provided a focus on the “participants’ perspectives” on their own behaviour (cf. Duff, 2000). The participants in each school knew each other and gained a level of familiarity with the principal researcher (Hawkey); their perspectives were captured through interviews, focus groups and from questionnaires. While these data were not systematic across all school contexts, in the analyses attempts were made to look for “recurrent patterns across data sets” and to seek triangulation (see Cohen and Manion, 1994: Denzin, 1988, Miles and Huberman 1994: Wall, 2005).

Table 5.1 summarises the overall participation in the case study schools, i.e. the type and number of participants.
Table 5.1  Overall participations in case studies

<table>
<thead>
<tr>
<th></th>
<th>Lessons videoed</th>
<th>Head teacher interviews</th>
<th>Teacher interviews, focus groups, etc.</th>
<th>Parent interviews, focus groups, etc.</th>
<th>Student questionnaires completed</th>
<th>Teacher questionnaires completed</th>
</tr>
</thead>
<tbody>
<tr>
<td>TOTALS</td>
<td>22</td>
<td>10</td>
<td>40</td>
<td>20</td>
<td>228</td>
<td>11</td>
</tr>
</tbody>
</table>

The qualitative nature of the data which were collected in the schools is shown in more detail in Table 5.2.

Table 5.2  Reproduced from Hawkey, PL2000 Impact study, Final report, 2003

<table>
<thead>
<tr>
<th>Schools</th>
<th>Class videos</th>
<th>Interviews, focus groups, written contacts</th>
<th>Questionnaires</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Head teachers</td>
<td>Teachers</td>
</tr>
<tr>
<td>School types</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10/0 1</td>
<td>4/02 1</td>
<td>10/0 1</td>
</tr>
<tr>
<td>Scuola elementare</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Istituto comprensivo</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Scuola media (1)</td>
<td>2</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Scuola media (2)</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Liceo (1)</td>
<td>1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>Liceo (2)</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Liceo (3)</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Istituto tecnico commerciale</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8 / 5</td>
<td>9</td>
<td>10</td>
<td>5</td>
</tr>
</tbody>
</table>
These tables show that a combination of classroom observations, interviews, and questionnaire data were collected from the two main participant groups - the students and the teachers. There was also interview data (some of it collected on more than one occasion), from school Heads and parents. This meant that the study did produce a significant amount of in-depth information, with opportunities for cross-checking and triangulation as originally intended.

The students in the case study classes at the seven schools were videoed during lessons in class and completed the student questionnaires covering language background, language learning experiences and use, attitudes, future plans and life goals. These data were supplemented by information from internal language tests, samples of students’ work, teachers’ comments, results from the Oxford Quick Placement Test (QPT – OUP, 2000), and Cambridge ESOL’s LLQs (although these instruments were not as widely used as had been planned at the outset and so the data were limited).

Other school personnel who participated in the study included six school Heads, and nine English teachers who were videoed in class, eight of them also interviewed, most re-recorded on the second school visit (April 2002), in addition to the 17 other teachers interviewed at the case study schools. Some of these teachers volunteered written comments on PL2000. Hawkey also received and responded to email assignments and comments from learners in two case study classes via their teachers. The information collected from teacher focus groups, interviews and questionnaires was compared with views expressed by teachers.
who took part in the Cambridge-Bell School essay competition (2001) in response to the question “What the PL2000 means to me as a teacher”, to which Hawkey was given access.

Through the offices of the Heads, the Impact Project researchers were able to meet and video-record twenty parents of students from the case study classes. This was a significant aspect of the project: although only a small sample, the parents represented a key stakeholder group who had provided significant support for the PL2000. They encouraged their children as language learners and provided support for the teachers and schools by endorsing the setting up of the courses and the provision of the certification. The concept of “verification of learning” based on external assessment proved popular with many of them, and in many cases they contributed towards the cost of the external examinations.

5.4.3 Summary of the outcomes

It will be remembered that there were five research questions in the PL2000 Impact Study focusing on the following:

i). changes in language pedagogy, materials, and use of media;

ii). changes in the behaviour and attitudes of the students and their teachers
    (including performance in external assessments);

iii). use of language assessment and changes needed to examination systems to
    accommodate the PL2000 school contexts;

iv). views of stakeholders in the wider educational context including managers and
    Heads teachers;

v). developments in communication and support structures related to the PL2000
    objectives (including in-service training, resource centre developments, test-awareness programmes, etc.).

Full details of the outcomes can be found in the project report which was presented to the Ministry in January 2003 (Hawkey, 2003). As Hawkey points out in this report, data reduction followed the model of qualitative research (discussed by Duff, 2000), taking the form of quantitative description, for example in the school profiles in (Appendix 5.8) and some of the questionnaire data, and “telling cases”, as in the interviews or classroom analyses. In keeping with the case study approach, descriptive statistical analysis was carried out on the data; where there were descriptive analyses of this kind (e.g. in summarising student questionnaire responses), the analysis was carried out by staff in ESOL’s Research and Validation group and the results were presented in the report by Hawkey.
In this next section the findings will not be discussed in detail but certain aspects will be picked out for discussion as meta-data.

a). Impacts on the classroom: changes in pedagogy, materials, and use of media

The communicative approach to language teaching and related pedagogy seemed to have been communicated effectively to the teachers, and this was particularly evident in the interviews (11 teachers and colleagues in case study classes and 37 teachers who were interviewed). Their comments revealed awareness of the communicative approach and possible ways to implement it in their classes. This included use of pair work, a focus on oral/aural skills, use of authentic reading materials, writing for communicative purposes such as correspondence, appropriate practicing for the external exams. An awareness that the external examinations focused on or incorporated speaking (both for the TCL exams and Cambridge ESOL) meant that speaking seemed to be equated in the minds of some respondents with the communicative approach. This can be considered a potential impact of the exams, in this case as washback. The increased attention given to speaking was possible as the same construct of spoken language ability was found in the learning objectives and in the criteria for success in the English exams; by focusing on interactive speaking skills, the teachers felt able to take “convergent action”, supporting both the learning goal and the examination preparation. Examinations with no speaking component would not have had the same impact.

The PL2000 did not specify curricular objectives in a taxonomic way, but proposed an alignment of objectives with the target levels of the CEFR and recommended a “modular organisation of curricula” in order to structure the learning load. This was to entail the division of the annual hours into short modules of learning, designed according to entry levels and target achievement objectives. The analysis revealed some of the consequences of this approach. For example, it was felt that these features gave the teachers and their students greater direction and they were more aware of the targets and communicative objectives.

Direct evidence of classroom pedagogy was gathered from 20 video-recorded language lessons in the case study schools. In most lessons that were analysed, there was evidence that the communicative approach, broadly in line with the aims and objectives of the PL2000, was being implemented. The teachers appeared to be motivated to implement classroom activities that were relevant to the communicative needs of their students and were prepared to admit that innovation was a direct
impact of the PL2000. Hawkey was able to demonstrate this in his report using extensive comments from replies to teacher questionnaires (compiled in April 2002) and supporting evidence from teacher interviews in the case study schools.

The “modular approach” to language teaching introduced explicit and manageable objectives emphasising successful learning, individualisation and prompt remediable action where appropriate. These aspects of the approach placed a high demand on the teacher to carry out appropriate lesson planning and class management and successful implementation depended on how well the teacher was able to handle these features. In some of the video-recorded lessons, Hawkey observed that on occasion “unproductive disorganisation set in” when the teacher was attempting to implement communicative activities and it seemed that the teacher may have devised an appropriate idea for a lesson but without adequately planning tasks or activities to ensure that learners got the best out of the lesson, or that the lesson fulfilled its role within the modular programme. Where such planning was apparently lacking, Hawkey noted the following kinds of problem: lesson fragmentation; stop-start episodes; overly long or muddled task organisation; students being inactive or not engaged (or perhaps bored); little opportunity for students to interact with each other; inconsistent opportunities for communication across the class; less than optimal monitoring; uneven use of teacher assistants (where present); an absence of round-ups, reinforcement, and consolidation. There was strong evidence, therefore, that lesson planning and classroom management are crucial aspects of the communicative approach. Knowing about the relevant principles is a necessary starting point, but in itself not adequate; the difficulties teachers experienced in practice suggest the need for additional in-service support programmes to help them to deal with these activities more effectively. This suggests that, even when there is potential for positive washback from the examination due to construct convergence with the learning objectives, unless training and experience of the teachers allow for appropriate lesson planning and management, the benefits may not be realised. A conclusion here is that examination providers might play a more active role in supporting teachers in this respect.

The teacher interviews, questionnaire data, and classroom observations raised a fundamental question of the appropriate balance between language functions and language form, an unresolved issue since the inception of the communicative approach in the 1970s. In changing to an approach based on the CEFR, a greater emphasis on functional language would have been anticipated by the Italian teachers in the PL2000. This obviously raised questions in the minds of some teachers as to the appropriate place of grammar and the extent to which a focus on form was still
appropriate. The findings on this question also reflected some of the classroom and planning issues discussed above. The project's objectives were explicitly stated: to develop communicative competence in reading, written and oral interaction and production. But problems in the form/function balance were noted in nearly half of the classrooms observed. The teachers' comments indicated that *linguistic competence* (lexical, grammatical, semantic and phonological competences) may require more systematic attention than the communicative approach might imply. As Hawkey noted, there was a tendency for teachers to be inconsistent and to correct some pronunciation, grammar and vocabulary errors but not others while students were in the process of communicating. In particular, this kind of correction seemed inconsistent with regard to the communicative impact of the errors, and other considerations such as the level of the students concerned. In conclusion, this question remained crucial to the achievement of the project and further professional support is needed in this area.

The PL2000 guidelines also referred to “use of diversified materials in place of textbooks” suggesting that this might lead to: textbook writers and publishers developing/revising books in line with the objectives; teachers and students using materials in place of or in addition to the traditional textbooks; use of examination-oriented books related to the external certification of students on PL2000 courses. Hawkey reports that there was evidence of all these tendencies. The language teaching textbooks specified by the case study schools on the school profile form seemed to indicate a preference for books taking a broadly communicative approach. The name of an external examination also occurred in a significant number of titles (e.g. *FCE Gold*, *PET Practice Plus*). But teachers also appeared to look for materials from a variety of sources including photographs, information-gap handouts, wall charts, examination practice cassettes, past papers, etc.

The use of IT and new media appeared as a separate recommendation in the PL2000 guidelines, i.e. “the integration of information technology into the teaching and use of diversified materials in place of textbooks”. Out of the twenty video-recorded lessons half used technological support for the lessons, for example: use of audio cassettes in class (chants/stories for use at elementary level; sample listening test cassettes for external exams at A2 level; “jigsaw listening” practice cassettes at A2); use of individual PCs (for student-to-student and student-to-teacher emails in a B2 level class); use of the school computer laboratory (for a variety of activities with a B2 class). Most of the IT examples in the observed lessons suggested a positive impact of the PL2000 and this was supported by the teacher interviews where
reference was also made to appropriate use of the conventional (audio cassettes) and newer technologies (PCs, internet, videos, etc.).

b). Impacts on perceptions, views and attitudes of students, teachers, school heads and parents

It will be remembered that data on possible impact on students in the case study classes came from several sources: the detailed student questionnaire (October 2001 and April 2002), the classroom observations (March 2001 to April 2002), and interviews and focus groups with the teachers, Heads and parents. A small scale application of the LLQ (also used as part of the IELTS Impact Studies) was used with 34 students. The motivation for learning English was investigated in October 2001 and April 2002 using the student questionnaires; in October, 118 students responded to the open-ended question: 'What are your opinions about learning English?' In their responses they exemplified some of their reasons for learning the language, the top three reasons being related to the status of English as an international language (51 responses); the potential usefulness of English in work (48 responses) and the benefits of English to their “future” prospects (36 responses). There were also mentions of travel, university study, and a liking for the English language and/or its culture. Table 5.4 summarises responses to a multiple choice item, on the student questionnaire (April 2002): “main reason for learning English” with 110 responses (some students selected more than one of the reasons given).

<table>
<thead>
<tr>
<th>Reason selected</th>
<th>Numbers of selections</th>
</tr>
</thead>
<tbody>
<tr>
<td>To get a better job</td>
<td>49</td>
</tr>
<tr>
<td>To communicate with more people</td>
<td>48</td>
</tr>
<tr>
<td>For travel overseas</td>
<td>37</td>
</tr>
<tr>
<td>Because I like it</td>
<td>25</td>
</tr>
</tbody>
</table>

This suggested the perceived importance of the English language in Italy as a means of international communication and to enhance professional opportunities. These reasons suggest a strong instrumental motivation and there was also an indication of integrative motivations, in the “because I like it” kind of comments.

The students were also asked to rate the frequency of their activities in using English outside class (e.g. ‘reading books’, ‘reading newspapers and magazines’, ‘writing letters’, ‘watching TV’, ‘using email’, ‘using the Internet’ etc.). Table 5.5 summarises these responses.
Table 5.5 Activities out of class

<table>
<thead>
<tr>
<th>Activities</th>
<th>Responses in October 2001</th>
<th>Responses in April 2002</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Never</td>
<td>Almost never</td>
</tr>
<tr>
<td>Reading books</td>
<td>43</td>
<td>16</td>
</tr>
<tr>
<td>Reading newspapers</td>
<td>38</td>
<td>30</td>
</tr>
<tr>
<td>Reading magazines</td>
<td>33</td>
<td>31</td>
</tr>
<tr>
<td>Reading newspapers and magazines</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Writing letters</td>
<td>32</td>
<td>36</td>
</tr>
<tr>
<td>Radio</td>
<td>38</td>
<td>23</td>
</tr>
<tr>
<td>TV</td>
<td>13</td>
<td>26</td>
</tr>
<tr>
<td>Movies</td>
<td>40</td>
<td>26</td>
</tr>
<tr>
<td>Going to shows</td>
<td>37</td>
<td>27</td>
</tr>
<tr>
<td>Using email</td>
<td>22</td>
<td>34</td>
</tr>
<tr>
<td>Using the Internet</td>
<td>30</td>
<td>19</td>
</tr>
<tr>
<td>Talking with visitors</td>
<td>30</td>
<td>20</td>
</tr>
</tbody>
</table>

There were some indications here that students by the year 2001 were already using the Internet as their main access to the language outside of the classroom (although broadband access had not yet become widespread in Italian households).

In response to the questionnaire item on the amount of time spent “studying English outside class, the 107 students responded as follows: (April 2002) as shown in table 5.6 showing that only limited language learning was taking place outside of the classroom for many students.
Table 5.6  Time spent learning English out of class

<table>
<thead>
<tr>
<th>Time</th>
<th>Numbers of selections</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than one hour</td>
<td>51</td>
</tr>
<tr>
<td>One to two hours</td>
<td>33</td>
</tr>
<tr>
<td>More than two hours</td>
<td>17</td>
</tr>
<tr>
<td>No time</td>
<td>6</td>
</tr>
</tbody>
</table>

In general, the student data on motivation and attitude were supported by the perceptions of the teachers who were questioned and interviewed. The open-ended comments made by teachers (April 2002) to the following question were particularly revealing: *Please list here and comment on what you see as the advantages and disadvantages of the PL2000 for your students and you.*

There were numerous comments referring explicitly to “motivation”, “interest”, “enthusiasm”, and increased “involvement” as advantages:

- “More motivation for students and teachers” (scuola elementare French teacher)
- “More interesting: children and teachers improve their use of the foreign language.” (scuola elementare English teacher)
- “More motivation, more involved in the lessons, more variety of lexis”. (scuola media teacher)
- “We prepare students, in a limited number, of course, good students. They are more stimulated, they are more involved in their activities.” (scuola media teacher)
- “Students are gaining motivation, extra hours, interest for languages.” (liceo teacher)
- “Students’ performance improves especially orally; students show more interest in different languages.” (liceo teacher)
- “Students are extremely enthusiastic; their speaking / listening / reading abilities have greatly improved.” (scuola media teacher)

The case study teachers were generally positive about the impact that the PL2000 was having on them and on their work. The following points from the teacher interviews (some relevant to the pedagogy recommended by the project) illustrate a *positive impact* on teacher attitudes.

- PL2000 brings “standardisation” (more teachers know what they are supposed to be doing) and motivation. (liceo teacher, April 2002)
- PL2000 means that “teachers know where they are going”; previously FL teaching objectives were vague, the emphasis on the academic aspects of literature, with the
basics of language forgotten.

- Teaching is now more rewarding because you can see the results.
- PL2000 emphasis on communication is appropriate and “more challenging for the teachers” (all from the liceo teachers’ focus group, April 2002).
- PL2000 has brought good participation by teachers, students and parents; the project is a good motivation for improvements from teachers. (scuola media teacher, October 2001).
- Big changes in teaching approaches bring very positive impact on students and teachers: NB the focus on language learning strategies and activities that engage the students. (liceo teacher of German, April 2002).
- Many teachers are motivated and feel “more able to comment”. (scuola media teacher, October 2001).
- Student motivation shown by the fact that some [C1] students are teaching English in primary schools voluntarily. (istituto tecnico teacher, October 2001).
- “Students are also more interested in international exchange projects nowadays”. (liceo teacher, April 2002).

Not all comments were positive in all respects. For example, the following from the teacher interviews:

- Language teachers are sometimes de-motivated by the extra administrative work entailed by PL2000.
- There may be resistance from other subject teachers to the scheduling of PL2000 afternoon classes. (liceo teachers’ focus group, October 2001).
- Difficult to motivate all the students in a mixed-ability class; “It is important for a teacher to stay with the same students over more than a year” (scuola media teacher, April 2002).

It was interesting to note that of the possible negative impacts, most related to the practicality of using the exams and to administrative difficulties which were encountered, including fitting in extra work preparing for the exams (both for the schools and for students). These kinds of impact, while not related to construct issues, are significant and often go unnoticed in the washback literature. Work load and practical concerns are important for the successful take-up of innovations in schools (such as the use of new examinations) because are they always important considerations for teachers.
The following responses to the same teacher questionnaire were also relevant in this respect:

- "They sometimes have difficulty doing homework if they remain at school in the afternoon". (liceo mother-tongue teacher)
- "Attendance during the course: students have many subjects to study and a lot of homework" (scuola media teacher)
- "Extra time devoted to PL2000 can create homework problems in other subjects" (liceo teacher).

These kinds of reaction are predictable in educational innovations; whenever significant changes take place various kinds of disruption also occur and this often results in extra work for the participants and an increase in uncertainty.

However, the overall impression from this analysis was that the teachers were well motivated by the PL 2000 and were keen to pursue its aims in a professional way. In general they also seemed to have been encouraged by their management and this may have also contributed to the positive attitudes. In the ten interviews with school Heads (between March 2001 and April 2002 - eight of which were conducted in Italian but translated into English), they made relevant points on the impact of PL2000 on motivation and on student attitudes.

From the Head of a scuola elementare:

- Very positive [impact] since the Project highlights the EU and the need for foreign languages for communication;
- The Head has noticed an increase in students’ enthusiasm, pride in their foreign languages, willingness to interact in them, and understanding of their importance (both, October 2001)

From the Head of a comprehensive scuola media:

- PL000 has achieved a greater degree of student participation in foreign language learning;
- The impact of PL2000 has been extremely positive, participation in it being necessary to reinforce the position of the school locally and more widely;
- PL2000 also helps the school to achieve a greater degree of open-ness.
- The parents seem happy and understand the objectives of PL2000 language teaching. The school has received no complaints and, sometimes, is congratulated on its work. Its chosen policy seems to have been accepted positively.
Foreign languages are important as borders and barriers have fallen and communication is facilitated between European and non-European countries. The students have more opportunities of communicating; a knowledge of foreign languages is fundamentally important for students completing secondary education as well as for continuing education at higher levels. (all five, October 2001).

From the Head of a second scuola media:

- PL2000 is in line with pressure from parents who want to enrich the students’ educational experience, and
- PL2000 impact is positive in terms of motivation for the students. Thus it is good for the school, but time is needed before final judgement can be made. (October 2001)
- There are now more students (and parents) wanting to participate in PL2000 courses, numbers rising from 31 to 50 over the year.(same Head, April, 2002).

From the Head of an Istituto Tecnico:

- PL2000 gives a new impetus; opportunities to strengthen the language education offered by the school, and
- Teachers are enthusiastic, prepared to work beyond the objectives the school sets; the school’s and PL2000’s objectives are the same! (both March 2001)
- The training has improved their English, given them different kinds of English (for different purposes), brought conferences, films, discussions and debates in English made them better teachers. So the PL has given the teachers the opportunity to broaden their knowledge, use better materials, made them more aware as teachers, and
- The enthusiasm of the PL has been positive for teachers and students (both April, 2002).

From the Head of a Liceo:

- PL2000 is to be seen not only as an important instrumental tool but also as a tool to enhance knowledge of civilisation, (March 2001)
- The school, as a leading school in the region, is very happy with the PL.
- The parents and students welcome the PL with open arms for English, French, German and Spanish. English is in the highest demand (both October 2001).
- PL is accepted at all school levels in this region. The students are now used to PL methods, so progress quickly.
- The PL is great from the teacher’s point of view given the training and satisfaction they get out of this kind of projects. School will have conferences and meetings open to locals, to discuss projects of this kind and get suggestions from them. and
• PL is a project of the highest importance, more than any of the many projects in which the school has participated over the years. Foreign languages are important not only for what they teach but as something the students will use in their studies and their jobs, thus vital for society, (same liceo Head, April 2002)

Innovation theory emphasises the need for positive top-down support in ensuring take up of new ideas in educational contexts (see Nicholls, 1983). The points from the Heads’ interviews provide good evidence of this in the PL2000 context. Their comments are notable for the emphasis put on the importance of foreign languages in broader educational contexts, as well as for utilitarian uses. There was clear recognition of potential benefits related to motivation and the greater interest the PL2000 had generated in students, teachers and parents. Importantly for school management, there were implications for enhanced reputation that participation in the project might bring to their schools.

As a result of what seemed to be excellent dissemination of information surrounding the PL2000, the parents who were spoken to in the impact study tended to be aware of the PL2000 and its objectives. Their support and encouragement was, therefore, an important factor and likely to have provided a positive endorsement for the school managers in adopting the PL2000.

From the school Heads, the following comments were made about the parents:

• PL2000 represents parents’ and students’ views re the importance of foreign languages, the English language in particular. Foreign languages are a first priority for parents and students. The parents requested more foreign language teaching directly, and
• Students and families know PL2000 is important for EOP, EAP. (Head of an istituto tecnico, March 2001).
• The Project has led to the construction of a language lab, with the assistance of parents.
• The parents seem happy and understand the objectives of PL2000 language teaching. The school has received no complaints and, sometimes, is congratulated on its work. Its chosen policy seems to have been accepted positively. (scuola media Head, October 2001).
• Good parent : school relations, but, of course, it is the most motivated families who send students to the afternoon LL, IT, arts and sports activities, about 10-15%. Not necessarily the best students, but the mid to high ability ones. The students really needing help do not come; their families tend not to trust the schools. (Head, scuola media, October 2001).
• The Preside is determined to expand and enrich the PL200 programmes in English and French to attract students (and parents) who are not so highly motivated as the initial groups (same Head, April 2002).
• Meetings with parents were part of the 2000 survey [into communication needs] (Head, istituto tecnico, March, 2001).
• The parents requested participation in PL2000. (Head, liceo, March 2001).
• The parents have supported the project fully; they had already financed various projects to enable their children to learn foreign languages (same Head, April 2002).

It was a design feature of the impact study that stakeholders from the wider community should themselves be involved, and especially parents wherever possible. Between March 2001 and April 2002, interviews or focus groups were therefore carried out with 21 parents who had children at elementary, middle and high school levels. These parents spoke of the impacts of the PL2000, both on their children and on themselves, and it was apparent that their comments were overwhelmingly positive. As a volunteer group, this might have been expected; it was likely that only parents with positive views would have accepted invitations from the school Heads to be interviewed in this way by an outsider. However, their views (translated from the Italian) are of interest and go beyond the expected project impacts to some extent: 23 parent comments referred to English as a tool for their children’s academic or working futures; 14 comments focused on the way English is or should be taught and learnt, and the value of having competence in the language.

• Favourably impressed, especially as his children can learn English from an early age. (father of twins studying at scuola elementare, October 2001).
• Good that the school is promoting foreign languages. (mother with a son learning French at the same school).
• Major impact on spoken English; the parents’ generation are unable to understand or speak when English language speakers speak.
• With mother tongue teachers, the children get real practice, and learn the pronunciation and culture; also there are a lot of TV ads in English, and the news subtitles in English on TV2.
• We are now living in a multi-cultural society; you need English everywhere as a medium, as for holidays, exchanges. The level of English on PL2000 course is higher than on curricular courses.
• PL a positive effect but the children need to go to English speaking countries for months to really learn it.
• Government should use both mother tongue and Italian teachers.
• Because of the EU and for universities, jobs and travel, English is important.
• English also opens horizons for the study of languages and other subjects in English (all from a scuola media parent focus group, October 2001).
• Very positive; son very enthusiastic, worries it will end, never misses a lesson.
• Enthusiasm not only for something new but because it is a different language, culture, all skills, a whole new experience, mother tongue teachers, new materials, communicative approach, the students to speak as much as they can. (parents of children at the same school, April 2002).
• 
• “[My daughter’s] enthusiasm to improve her knowledge of English, now seen as a second language for all, has been passed on to her mother, father and younger brother….. This has strengthened our interest in the study of English.” (scuola media parent, whose daughter is now at high school).
• “[My daughter] is very satisfied with her language competence and likes learning English; she has been abroad twice, the first year to Ireland, the second to London. She enjoyed the experience, especially using English as a language.” (second scuola media parent at the same school, both October 2001)
• “The PL2000 is a good project. I am glad my children are involved”.
• PL2000 methods are good as the students practice what they have learnt. The teachers have the background and expertise, which give parents confidence in PL20000.
• When [my husband] learnt a foreign language the teachers knew the language but couldn’t speak it! (C1 student parent, istituto tecnico, March 2001).
• “There should be a lot of computer use in foreign language learning”. (father of the same student, March 2001).
• The mother felt that the PL2000 had given her son positive experiences in terms of content, methodology, the students’ confronting a different reality and maturing in the views of different cultures.
• The father felt that his daughter was enthusiastic about her current course.
• The mother said that the use of IT, Internet and computers allowed a focus on individual learning and helped the students to mature.
• The father speaks to his daughter in English now and they can have a real conversation, with the daughter helping him with his vocabulary. (all re parents of B2 students at the same school, April 2002).


c) Impact on issues related to assessment

In addition to the general impacts of the PL2000, it was very important for the impact study to focus on the role of assessment, including the views and attitudes of the stakeholders towards the examinations which were available. The following comments were made by teachers in relation to the Cambridge exams; these are taken from the teacher interviews and cover both general points and specific Cambridge exams. They only cover English exams from Cambridge as the case study schools had all chosen to use Cambridge ESOL as their external exams. The comments reveal that the teachers believed that the exams had features that were
coherent with the objectives of the PL2000 and added value for learners. The following four points summarise the kinds of comments which were frequently made:

Recognition and currency: the well-known Cambridge certificates add value, especially PET and FCE.

Progression: the exams enable the learners to set targets, show progress across proficiency levels; they also provide continuity from school to university and/or work.

Four skills: the Speaking and (to a lesser extent) Writing components emphasise the communicative goals of language learning.

Motivating for learners: Cambridge examinations offer important benefits for learners and increase their motivation, even when the instrumental purposes are less obvious, as in the case of CAE which lacked recognition in Italy at the time.

The following are examples of the general comments which were made about the examinations which had been adopted:

- UCLES is the more recognised Board because their exams cover the four skills, and
- Cambridge exams can provide a progression throughout school life (two teachers at a scuola media, October 2001)
- UCLES is better known than other boards, has a broad examination suite of exams covering all four skills and whose certification does not expire, is better known by receiving organisations. (Istituto tecnico teacher, March, 2001)
- UCLES exams are motivating, as well as accepted for university entrance, and
- UCLES exams may not be sufficiently ESP, (both from the liceo three-teacher focus group, October 2001)
- The Cambridge materials, programs, exams and certificates are quality tools. (Liceo mother-tongue teacher, February 2002 email)
- Some students find the Cambridge tests too ‘cognitive’, i.e. with items testing more than language competence. The teacher responds “If you have to think and can still communicate, you know the language quite well” (teacher from the same liceo, April, 2002).
- “English for Life” and PL000 at the school are based on taking a Cambridge examination after 1, 2 and 4 years (KET, PET, FCE). (Three liceo teachers, March 2001)
- The school aims at KET for year 1, PET for Year 2, FCE for Year 4, for students, nearly all of whom go on to university, aiming to become doctors, engineers,
lawyers, interpreters, teachers etc. (two mother-tongue teachers of English at the same liceo, October 2001).

- Teachers want more details re their students' UCLES examination performance, e.g. marks for each examination part, averages, suggestions for improvement; the existing profile chart may be "misleading",

- The timing of some of the external exams is inconvenient and thus demotivating. (both from one istituto tecnico Italian and one mother-tongue teachers of English to B1 level students, October 2001)

The following comments are examples of views expressed about particular Cambridge exams:

1. **KET related (A2)**
2. “KET may be more suitable for secondary school?” and
3. **KET provides an “impulse for learning”,** (both from two teachers at a scuola media, October 2001).
4. The KET programme impact is positive; good tests lead to good teaching. KET speaking is OK, but the writing test is difficult. KET grammar could be improved, (scuola media comprensivo teacher, October 2001).
5. **KET speaking tasks seem to reflect PL2000 objectives.** Target language letter-writing may not be a realistic communicative activity for students who write only notes (same teacher, April 2002).
6. Only selected students take the external examination (KET) (young A1 level teacher at the same school, April 2002).

**PET related (B1)**

1. PET can provide students with global, widely recognised, consistent communicative targets and certification, and
2. PET is not recognised as an EFL qualification by some universities, and
3. Teachers and students need more cassettes, more details of the meaning of results, e.g. ‘good’, ‘with merit’, ‘narrow fail’ and why not let students see their marked scripts? (istituto tecnico Italian C1 teacher and mother-tongue teacher-assistant, March, 2001).
4. Is PET good as a university entrance qualification? (scuola media teacher, October 2001).

**FCE related (B2)**

1. FCE is not only about language competence but also attitude. (three-teacher focus group, liceo, March 2001)
2. The FCE examination has changed, now more real, with a good format, no longer “old-fashioned” (istituto tecnico teacher and mother tongue assistant, March 2001).
3. Exams such as FCE encapsulate CALT, the four skills, especially speaking and writing, and the testing of real communicative abilities (three-teacher focus group, liceo, October 2001).
4. “Students should be aware that a large amount of hours are necessary to reach an FCE level; curricular teachers should themselves not only be aware of what an FCE level is, but also co-operate to select FCE candidates within their class groups” (liceo teacher, email April 2002).

CAE related (C1)
1. CAE is not a well-known or required examination, is taken by students for integrative motivational reasons (istituto tecnico teacher, April 2002).

d) Impacts on school management and teacher support
Various types of impact of the PL2000 project were observed on the managerial and administrative systems within the educational system. These impacts related to the need to adapt to the PL2000 requirements (as set out by the Ministry) within schools and to provide adequate leadership and support for the following:

- The integration of PL2000 programmes into institutional curricula.
- The formation of the homogeneous learner groups, especially for after school lessons.
- Additional resourcing including finances and payments (e.g. for the external exams) sometimes obtained from a variety of sources (the Ministry, regional funds, parents).
- Provision of teachers and facilities for the after school lessons.
- Appropriate distribution of responsibilities for referenti within the schools and school districts.
- The gathering of information and dissemination to relevant stakeholders.
- Adequate liaison with the resource centres (CRTs) and with the professional development programmes for teachers.

(NB. The importance of adequate communication was noted in the maxims of Milanovic and Saville, 1996, as outlined in Chapter 1).

In 2000, the CEFR itself was new to most language teachers and to most other stakeholders; many needed to understand the levels (A1 to C2) as well as the inherent “action-oriented approach” to teaching/learning. Also the role of assessment in the project and potential choices needed to be explained; the teachers, parents and students needed to know what was available; what the differences were between exams and what was a suitable option for them in their own context (level, skills tested, availability, cost, etc.).
Professional support for the teachers was important because of this and because of the generally innovative nature of the reforms of the PL2000 which put emphasis on “the communicative approach” with an explicit focus on real and relevant communication (that should be taught and learnt through the most up-to-date approaches and making use of contemporary media).

This need for professional support for teachers (required by the Ministry) also had an impact on the selected examinations board. For example, in the case of Cambridge ESOL, it led to a change in both the focus and the scale of the Cambridge seminar programme. Changes were made to reflect the specific needs of Italian state-school teachers who constituted a new stakeholder group for the Cambridge exams. Completely revamped and extended seminars for KET, PET, FCE, CAE and CPE were run from 2001 onwards to help the teachers bridge the gap between their more traditional methods and the communicative approach incorporating assessment of the four skills.

To meet the growing demand for Cambridge seminars in support of the PL2000, a team of 25 experienced presenters was set up throughout the country and they were trained to support the specific needs of the project. By mid-2001 this team had already run nearly 150 seminars (free of charge for the participants), usually lasting about three hours and aimed to meet the needs of local schools or the requests of the regional resource centres (CRTs - Centri Risorse Territoriali). Many of the larger language teaching meetings or conferences included two or three such seminars.

Print and web-format support materials were developed in Italian for teachers, parents and students. Italy was the first country to have a specific Cambridge website, Italia On-line. All handbooks and sample examination papers could be downloaded directly from the site. In 2000/01, Cambridge materials distributed included examination handbooks, speaking test videos, CDs for teachers and leaflets for students and parents. Cambridge ESOL also appointed a part-time PL2000 Coordinator (Gaby Forgione), based in Rome to support the full-time Country Manager for Italy. Both worked closely with about 130 Local Secretaries (i.e. those responsible for organising the examination in each Cambridge centre). The Cambridge Senior Team leader (STL) for the Speaking Tests worked with 43 Team Leaders in order to coordinate over 1,000 Oral Examiners and the Inspections team conducted visits of the new centres where the exams were being administered.

The impact of the PL2000 on the examination providers in their attempt to meet the Ministry’s requirements is an important finding. It supports the view that complex
interactions develop between examinations and other processes within educational systems.

5.5 The Cambridge exams within the PL2000

Cambridge ESOL’s routine validation procedures, supplemented by adapted instruments to collect candidate-related data, were the basis for the quantitative analysis. This allowed the performance of all candidates taking a Cambridge examination to be monitored and included item responses and scores (overall and by component). Background features of the candidates were also captured by the routinely administered Candidate Information Sheets. These were supplemented by procedures to capture information pertaining to the PL2000, including the school and class of the candidates, which was necessary to report results and performance trends in the Italian school system to the Ministry. Additional PL2000 data capture was, therefore, put in place for this purpose. The analysis carried out by Cambridge validation staff proved particularly useful in providing evidence of overall take up and success rates at the different levels, and revealed cities or regions where there was particularly high take up of the Cambridge exams (e.g. Lombardy). Possibly for the first time in the Italian context proficiency levels could be compared objectively across learners in different schools and informative comparisons were made with similar test takers in other regions or countries (e.g. Lombardy compared with Piedmont, Baden-Wurtenburg, or Catalonia).

Graph 5.1 shows the number Cambridge exams taken in Italy overall and the proportion taken as a result of the PL2000 over a four year period. Table 5.7 shows how these were distributed across the exams at the CEFR levels.

Graph 5.1 Number Cambridge exams taken

![Graph 5.1 Number Cambridge exams taken](image-url)
Table 5.7 Candidates for Cambridge Exams at all CEFR levels

<table>
<thead>
<tr>
<th>Examination</th>
<th>CEFR Level</th>
<th>1998/99</th>
<th>1999/00</th>
<th>2000/01</th>
<th>2001/02</th>
</tr>
</thead>
<tbody>
<tr>
<td>YLE Starters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>A1</td>
<td>792</td>
<td>1,853</td>
<td>3,229</td>
<td>4,883</td>
</tr>
<tr>
<td>LE Flyers</td>
<td>A2</td>
<td>244</td>
<td>882</td>
<td>1,512</td>
<td>2,357</td>
</tr>
<tr>
<td>KET</td>
<td>A2</td>
<td></td>
<td></td>
<td>11,163</td>
<td>13,945</td>
</tr>
<tr>
<td>PET</td>
<td>B1</td>
<td>1,658</td>
<td>3,654</td>
<td>9,78</td>
<td>12,02</td>
</tr>
<tr>
<td>BEC 1*</td>
<td>B1</td>
<td>82</td>
<td>885</td>
<td>2,00</td>
<td>604</td>
</tr>
<tr>
<td>FCE</td>
<td>B2</td>
<td>6,909</td>
<td>8,378</td>
<td>11,870</td>
<td>13,113</td>
</tr>
<tr>
<td>BEC 2*</td>
<td>B2</td>
<td>60</td>
<td>122</td>
<td>255</td>
<td>260</td>
</tr>
<tr>
<td>CAE</td>
<td>C1</td>
<td>1,499</td>
<td>1,471</td>
<td>1,682</td>
<td>1,812</td>
</tr>
<tr>
<td>BEC 3*</td>
<td>C1</td>
<td>10</td>
<td>9</td>
<td>28</td>
<td>111</td>
</tr>
<tr>
<td>CPE</td>
<td>C2</td>
<td>952</td>
<td>750</td>
<td>725</td>
<td>744</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td></td>
<td>20,200</td>
<td>31,063</td>
<td>53,566</td>
<td>72,928</td>
</tr>
</tbody>
</table>

It is clear that the PL2000 had a significant impact on the activities of the examination providers for all the languages included in the project. By 2003, the PL2000 entries accounted for a large proportion of all entries for the Cambridge exams in the country: 68% of entries at A1, 81% at A2, 64% at B1 (the level with the most Cambridge examination candidates) and 49% at B2 level.

A concern from the outset was to ensure that learners should attempt to take exams which were suitable for their proficiency level. At A2 (KET) level, the pass rate for PL2000 candidates in the 2001/2002 academic year was 70% which was just below the rest-of-the-world average (71%), but not quite as high as the figure for non-PL2000 candidate (79%). Further investigations revealed some particular difficulties which the PL2000 candidates experienced in taking this examination in comparison with other Italian candidates. In particular, lack of experience in Listening – both in class and in examination contexts – appeared to be the main problem, especially for the younger middle school students.

For PET and FCE the examinations appeared to be well-suited to the level of the candidates who were entered and the PL2000 pass rates compared favourably with other
cohorts. At PET (B1) level, it was 82% which was slightly higher than the world average (81%) and the same as the non-PL2000 candidates in Italy. The FCE results for 2001/02 showed a higher pass rate for PL2000 students than non-PL2000 or the rest of the world.

As anticipated, most school-age students were not able to reach the highest CEFR levels (C1/C) and so entries for the C1 and C2 level exams remained relatively low, (although several of the older students in the case study group did attempt CAE and CPE – i.e. 126 for CAE and 3 for CPE).

The need to expand the capability and capacity to administer more exams in more places in the country was a major impact and there was a large investment of staff time and other resources by the examination providers. It is interesting that some of them partially adapted their exams to meet the specific objectives of the PL2000 while still offering the international recognition and currency which the Ministry required. Cambridge ESOL and TCL introduced modifications to their English exams; the Goethe Institut introduced new versions of their German examinations at the A1/A2 levels (Start Deutsch); and the Instituto Cervantes adapted their Spanish examinations (DELE) for the Italian scholastic context. The enhanced professional support programmes run by the boards proved important in communicating the aims of the project to a large number of participants, encompassing more people and a wider geographical spread than the Ministry would have otherwise achieved in the time available. The roll out of the PL2000 across Italy depended heavily on the preparedness of the regional ministries and local schools, and the experience in Lombardy emerged as a particularly interesting example of how the change process evolved. The “Lombardy experience” is summarised in Appendix 5.11 and illustrates how the quantitative validation data were used to evaluate the project by local Ministry staff.

5.6 Follow up to the Impact Study

The PL2000 itself came to an end in 2002/3 when central funds were no longer made available. The four segments of the project were intended to carry on after the PL2000 which had been intended to “kick start” the innovations within the reform programme. The legacy turned out to be long lasting in many parts of the country, surviving several changes of government. The continuation of the project allowed for late adopters to take part having seen how others had already managed the changes successfully (cf. Rogers, 1962 - confirmation phase of the diffusion curve; also Fullan, 1991 – continuation phase of the innovation process).

The impact study report by Hawkey was presented to the Ministry as planned, although the extent to which it was used or acted upon is not known. From a Cambridge ESOL point of view, two further developments came out of the PL2000 impact project:
• A project to investigate the impact of the PL2000 on higher education, i.e. those institutions who received students from the schools after participation in the project. This became a collaborative research project between Cambridge ESOL and the University of Siena, Language Centre.

• A project to investigate the language learning gains of individual students preparing to take examinations linked to the CEFR levels. This became a collaborative project with the British Institute of Florence and formed the basis of the third impact study case described and analysed in Chapter 6.

5.7 Insights and lessons learnt from this case

5.7.1 General points

The PL2000 impact project applied lessons learnt in the IELTS studies; it adapted the instruments and techniques and introduced new features of data collection. For example, it demonstrated that it was possible to apply the approach successfully within a national context and introduced case study schools involving school visits and interviews with participants in their local settings.

From the point of view of Cambridge ESOL, it demonstrated the possibility of successfully matching the learning objectives within a national curriculum project to the Cambridge examinations by referring to a shared, neutral framework of reference (in this case the CEFR). This provided the basis for the shared constructs which are necessary to achieve intended washback (as discussed by Green, 2007).

It was not possible in the PL2000 impact study to collect base-line data or to perform a systematic before, during, and after analysis. It was possible, however, to monitor how the project developed over the timeline of the PL2000 case (Figure 5.1). As in the IELTS case, the timeline is important aspects of the analysis.

Figure 5.6 illustrates the many local contexts and specific roles of participants which had to be taken into account in order to understand the effects and consequences that were attributable to the PL2000 innovations as a whole, and those which might have been attributable to the specific tests used.
5.7.2 The Context

In Chapter 4 the relationship between an examination provider’s own organisational structures and local stakeholder groups in the many test-related contexts around the world was discussed. Figure 5.7, based on Figure 4.1, highlights the principal concern in this chapter, namely the local test preparation and test taking contexts in Italy.

Figure 5.7 Test preparation and test taking contexts

Schematic diagram showing the organisational network
In the IELTS case, local contexts varied greatly across many countries but the PL2000 project provided an opportunity to focus on a single country - shown inside the red box above the central line. The macro context was the school sector at a time of political and educational change and the importance of this socio-political milieu has been discussed and exemplified, as well as the importance of the individual stakeholders in various roles. Their perceptions, views and attitudes were examined in relation to the actions taken (including the use of the English exams).

We have located impact research within contemporary approaches to educational systems (illustrated by Figure 1.1 in Chapter 1) and have noted relevant research which has focused on innovation and how change is managed successfully (see also Kennedy, 1998: 332 for a discussion of the hierarchy of inter-relating sub-systems in which an innovation has to operate). In the PL2000 case we have observed the interplay between sub-systems and “cultures” and how individuals were critical in implementing policy and in bringing about changes. The importance of evidence on which to base analyses and to support arguments/claims about the nature of impact was also central to this approach.

Table 5.8 summarises the main contextual features of the Italian milieu at international, national, regional and local levels. As Figure 1.1 suggests, variations and contextual differences increase as you go “down the list” from national to local level.

<table>
<thead>
<tr>
<th>The international context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe – EU and other institutions (Council of Europe)</td>
</tr>
<tr>
<td>European policies: Mother-tongue plus 2 White Paper</td>
</tr>
<tr>
<td>Lisbon strategy; job creation, structural reform and social cohesion</td>
</tr>
<tr>
<td>Languages for competitive advantage: the knowledge based economy in Europe</td>
</tr>
<tr>
<td>Mobility, skills development, use of multi-media</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>The country - Italy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cultural and intercultural concerns</td>
</tr>
<tr>
<td>Attitudes towards language learning in Italy</td>
</tr>
<tr>
<td>Views of suitable target languages and cultures</td>
</tr>
<tr>
<td>Political system</td>
</tr>
<tr>
<td>Political climate (e.g. 1990-2003)</td>
</tr>
<tr>
<td>Ministerial structures: individual politicians and civil servants</td>
</tr>
<tr>
<td>Administrative structures</td>
</tr>
<tr>
<td>Educational system: structures and systems, resources, philosophical approaches, e.g. towards language education</td>
</tr>
<tr>
<td>Linguistic concerns: the relationship between Italian as the L1 and learning of target languages – “genetic” distance between L2 and source language</td>
</tr>
<tr>
<td>Availability of target language within national life (e.g. on TV, in print media and popular culture, etc.)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Regions (Lombardy, Piedmont, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geographical location (north, central, south)</td>
</tr>
<tr>
<td>Socio-economic development</td>
</tr>
</tbody>
</table>
5.7.3 Processes at the macro level

There are few impact projects which use case studies in national contexts, and the work of Wall in Sri Lanka stands out and is of particular relevance. Wall (2000) posed two questions: whether impact can be predicted and whether it can be controlled. In the PL2000 case, we have seen that the impact could be partly predicted (e.g. through impact hypotheses) based on policy documents issued by the Ministry (in Henrichsen’s terms “the resource system” and “innovators” responsible for setting up the project). These predictions included the impact that external assessment might have on other aspects of the project. It is clear, however, that the precise nature of the impact could not be controlled, mainly because the options available to the users within the explicit policy of school autonomy allowed for many potential outcomes, including a choice of which examinations to use. We can conclude, therefore, that socio-political factors at work within the macro context (nationally and regionally) provided some of the necessary conditions for the intended impacts to be achieved, but the autonomy of action which participants enjoyed locally meant that these impacts could not be easily predicted, and certainly not controlled.
The PL2000 case confirms the usefulness of innovation theory and the applicability of process-based models of diffusion in the context of impact research; such models allow the impact researcher to categorise the many factors at work in the milieu and help to illustrate how they interact with one another within a specific educational context. Wall (1999; 246-248; see also 2005) makes 14 observations about innovation and the management of change within education, summarised as follows.

Innovation:

- is different from other sorts of change in that it seeks to bring about planned improvements to a system;
- is a long, complex process consisting of many stages;
- has many participants, each with their own needs and limitations;
- has different meanings for every individual involved;
- may require change on different levels – content, methodology and attitudes (values etc);
- will reach different levels of implementation (cf. Beretta 1990, Bangalore Project); some participants will respond as planned and others may fail to take advantage of the changes;
- is difficult to measure where change is open-ended or is related to awareness
- has characteristics which both hinder and facilitate change;
- involves the analysis of context to judge whether it will be adopted – not just the classroom but the educational, political and cultural systems;
- will have a take up rate determined by many factors;
- can be implemented using a variety of models of change which are appropriate to the context;
- cannot be evaluated too early in the process as it may take time for the changes to be brought about.

She concludes with the following recommendations for investigating the impact of innovation in assessment contexts:

a). Analyse the "antecedent" situation to ensure that the change is desirable.
b). Involve teachers and possibly students in the stages of planning.
c). Incorporate stakeholder representatives in the design team to ensure that the test is both acceptable and comprehensible to those involved.
d). Provide draft test specifications for all key stakeholders, and carefully pilot the new test before its introduction.
e). Build on-going evaluation into the implementation process.
f). Do not expect an instant impact on instructional practices, or the precise impact anticipated.

It will be remembered from Chapter 2 that Henrichsen’s hybrid model has three components which provide the sequence of phases used by Wall in structuring her study: antecedents, process and consequences. The antecedent and process factors are useful in analysing contextual parameters, especially the macro context in a case such as PL2000. The three components of the model can be broken down as follows:

1) Before the innovation: the Antecedents
   Characteristics of the intended user system
   Characteristics of the intended users (e.g. teachers, students)
   Traditional teaching and assessment practices
   Experiences of previous reformers (e.g. what did/did not work before)

2) During the innovation: the Process
   Innovator (source of the innovation)
      Innovation (Message):
         plans and strategies:
         channels of communication
   Receiver – awareness, interest, evaluation
   Factors which facilitate or hinder change:
      Within the innovation itself (e.g. originality, complexity, explicitness, practicality, etc.).
      Within the resource system (e.g. capacity, structure, openness, harmony).
      Within the user system (e.g. location, centralisation of power and administration, size of unit, communication structure, education philosophies etc.).
      Inter-elemental factors (e.g. compatibility, linkage, reward, proximity, synergism).

3) After the innovation: the Consequences
   A decision is made to adopt or reject the intended changes:
      Adoption: this may lead to longer term continuation or to discontinuation at a later date.
      Rejection: the rejection may be continued or abandoned leading to adoption at a later stage.
   The consequences may be:
      immediate or delayed;
      direct or indirect;
      manifest or latent;
      functional or dysfunctional.
Wall’s analysis (2000: 506) using this framework confirmed that test impact is complex and that it is not reasonable to assume that all the factors will work the same in every context or that they will lead to the same outcomes. Based on her adaptation of Henrichsen’s model, she claims it is possible to make informed decisions about “the amount of risk that will be involved in any new testing venture” and that helps educators “to think about the kinds of characteristics that tests need to display if they are to create positive impact in difficult circumstances”.

The proactive use of such a model by the Italian Ministry planners would indeed have been helpful during the planning phase of the PL2000 (1997-1999). However, the nature of the Ministry’s own planning and decision-making processes and the dynamics of the events leading up to the launch of the project (e.g. the very short timescale for implementation) would not have enabled this. Interestingly, the Ministry planners, perhaps unwittingly, did implement aspects of the approach, and this may have promoted the successful implementation of the proposed changes. For example, key members of the Ministry team had been involved in language education over many years and had experienced many “special projects”; this provided continuity of experiences from previous projects which was brought into the planning and implementation phases. Their network of contacts nationwide and experiences of working with the cultural bodies (and examination providers) meant that the Ministry team had realistic and potentially achievable expectations.

Wall made a number of points about the Sri Lankan case which resonate in the Italian context, especially in the analysis of qualitative data collected in the case study schools. She commented that the ability and commitment of the teachers themselves was a critical factor, and traditional ways of teaching were not easily affected without provision of additional teacher training and monitoring related to the specific innovation; even well-trained teachers had difficulty introducing new ideas within their own existing classroom contexts. This might have been predicted from a fuller analysis of the antecedent factors and she recommended that, in future test impact studies, a baseline study should be conducted at the outset in order to understand as much as possible about the context that a test will be introduced into (see also Weir and Roberts, 1994, and Appendix 13 in Wall, 2005).

A greater emphasis on the antecedent factors helps determine whether a change is desirable, and if so, what the scope of the innovation might reasonably be in order to achieve the planned objectives. As argued above, however, the concept of a baseline study, while useful is often not always feasible in practice, at least not in the rigorous way Wall envisages. The time required for conducting baseline studies is
often at odds with the political agendas of the politicians and officials who have to implement reform programmes.

What is important for the impact model, is the need for a clearer understanding of the milieu into which assessment procedures are introduced and a recognition that there will inevitably be complex interactions between components of the system. Wall’s analysis of the process component of her project confirmed this view (see Wall, 2005: 269-272: Table 12.3, summary of the characteristics of the user system; Table 12.4, summary of the characteristics of the users – teachers; and Table 12.5, summary of the characteristics of the users – students). She showed that certain aspects of the innovation itself made it difficult for the participants to work with it and to cope with the changes, but more importantly, she notes that it was the system as a whole which had difficulty coping. In other words, it was the interaction between the factors within the system (i.e. the specific educational context) which either hindered or facilitated the process. Wall points to the following:

- there was not enough support materials for teachers;
- there was too little teacher training;
- there was not enough time for teachers to deal with the planned syllabus;
- there were too few links between participants in the wider user system – i.e. between teaching associations, colleges, universities, publishers, etc);
- there was ineffective exchange of information and expertise within the system;
- there were too few rewards for teachers in implementing the changes.

In her case study, these problems could not be easily addressed and resolved because the responsibility for addressing them was not located in a single individual or institution. Similar problems and issues emerged in the PL2000 case although there were localised factors which allowed some issues to be addressed more successfully at regional/local levels (e.g. in Lombardy).

5.7.4 Processes at the micro level

It was anticipated that the seven volunteer schools would provide only limited data which could not be generalised to the whole country; i.e. a small “stratified, purposeful sample” with limited coverage (Miles and Huberman 1994: 28) providing a micro level focus on individual learners and their teachers. However, the school-based analysis allowed the researchers to make some general observations about the nature of the innovative processes and the influence of the Cambridge exams as one “segment” of the integrated PL2000 initiatives. The schools in question were considered “early adopters” where participants were eager to take advantage of the
additional resources and support for their English language teaching. The choice of
the exams was a key part of their plans and, as shown by the interviews, there were
clear views on the benefits of external exams in general, and of the Cambridge
exams in particular.

The case studies provided invaluable insights into what actually occurred in the
classrooms during the PL2000 project. The possibility of investigating potential
mismatches between the stated intentions (e.g. to implement the communicative
approach) and what actually occurred in classroom contexts was probably only
available as a result of this impact study through the classroom observations and
interviews with the participants. No similar research was conducted by the Italian
authorities themselves.

The comments made by teachers and other participants about the Cambridge exams
were generally positive (as discussed above); the link between the features of the
exams (the underlying constructs and formats) and the objectives of the project were
generally well understood and acknowledged. The potential for positive washback
therefore existed in the sense intended by Green and discussed in the earlier
chapters. The impact study, however, confirmed a view put forward by Alderson (op
cit), that the potential for washback is not enough to ensure that good practice results
in classroom contexts. Hawkey was able to observe classes which were intended to
enact communicative methodologies but which were only partially successful,
probably due to limited training or experience on the part of the teacher.

One might conclude from this, that the potential for intended washback is “latent”
within a testing system (largely related to construct overlap) and that it may or may
not be “released” due to the interaction with contextual factors. These factors might
be termed “promoters” and “inhibitors” which either facilitate or hinder the intended
impact or influence the direction and intensity of washback (as in Green’s model). In
other words, an assessment system carries the potential for specific effects and
consequences deriving from the system as a whole, including the availability of
information and support materials for participants, the collection of adequate data
and the resources for implementing a validation programme, etc.. The extent to
which the latent potential is released will depend on contextual features which are
often determined by national or local politics and features of the educational system
(many of which will be beyond the direct control of the examination provider). An
alternative assessments system with different design features will hold the potential
for different effects and consequences depending on the interaction with the same
contextual features.
5.7.5 The participants (stakeholders)

In the PL2000 impact study it was possible to identify more precisely than in the IELTS case the key stakeholder groups and individuals and to analyse the complexity of the interactions between them in relation to PL2000 objectives. In the qualitative data in particular, it was possible to “hear the voices” of these participants.

Table 5.9 shows the broad categories of participants in the left hand column and specific groups and individuals are shown in the right hand column (in contrast to the similar table for IELTS – Table 4.5).

Table 5.9 Broad categories and specific individuals

<table>
<thead>
<tr>
<th>Broad Categories and Groups in the Italian Educational Context</th>
<th>Specific groups and names of individuals relevant to the PL2000 and the Impact Study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stakeholders within the test production and delivery networks</strong></td>
<td>Examinations board chosen by the MPI</td>
</tr>
<tr>
<td>• Owners of the tests and their governance structure</td>
<td>Individual members of staff allocated to the project including those working in Italy</td>
</tr>
<tr>
<td>• Staff within the examinations board</td>
<td>Staff working with Italian data and producing analysis, reports etc</td>
</tr>
<tr>
<td>Senior Managers in operational positions</td>
<td>NS/RJ and Impact team</td>
</tr>
<tr>
<td>Test developers and validation staff</td>
<td></td>
</tr>
<tr>
<td>Impact research team</td>
<td>Senior team leaders and Speaking examiners in Italy</td>
</tr>
<tr>
<td>Administrative staff</td>
<td>Seminar presenters – both in Italy and from UK</td>
</tr>
<tr>
<td>• Item writers and senior consultants (chairs of teams, chief examiners etc)</td>
<td></td>
</tr>
<tr>
<td>• Examiners and markers e.g. for speaking and writing tests</td>
<td>The 103 Cambridge examination centres working with state schools. Examination centres for other examination providers</td>
</tr>
<tr>
<td>Team leaders and trainers</td>
<td>Event organisers, including publishers examination centres, USRs etc</td>
</tr>
<tr>
<td>• Consultants and specialist advisors including academics</td>
<td></td>
</tr>
<tr>
<td><strong>Stakeholders within the test administration networks</strong></td>
<td>Students at all level of school enrolled in PL2000 courses</td>
</tr>
<tr>
<td>• Test centre administrators and support staff</td>
<td>Students choosing to take an examination related to the PL2000</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Primary test users – learners/test takers</strong></td>
<td></td>
</tr>
<tr>
<td>• Language Learners</td>
<td></td>
</tr>
<tr>
<td>In elementary, middle and high schools</td>
<td></td>
</tr>
<tr>
<td>In vocational training</td>
<td></td>
</tr>
<tr>
<td>As language school students</td>
<td></td>
</tr>
<tr>
<td>• Candidates for the test</td>
<td></td>
</tr>
<tr>
<td>Language learners involved in the test taking process</td>
<td></td>
</tr>
<tr>
<td>From the decision to enter to the use of results</td>
<td></td>
</tr>
</tbody>
</table>
### Secondary test users
- Parents and other sponsors of the learners/candidates
- Language Teachers in state-funded schools
- Language Teachers in private schools
- Language teachers in private language schools
- School Managers
  - School heads
  - Local representatives for projects
  - Course designers and materials producers
- Administrative staff (secretaries, clerks, etc.)

### Policy makers and support agencies
- Italian government and regulatory authorities
- Prime Minister and Minister of Education
- Ministry of Education
  - Legislative framework
  - Curriculum planning decisions
  - Central budgeting and distribution of funds
- Communication and dissemination of information
- Liaison with other agencies such as the cultural institutes and the examination providers
  - Project staff – central ministry
  - Project staff - regional Ministries and key staff such as local inspectors
- Regional resource centres
- GUI
- Consulates and embassies

### International cultural bodies
- Professionals and professional bodies and associations
  - Teaching associations
  - Regional and national associations
  - Specialist academics and experts in language testing
    - ALTE, etc
  - University departments and language centres
  - CRUI
  - Employers’ organizations

### Parents of students/candidates
- involved in PL2000 lessons and exams
- Language teachers in many schools contexts involved in PL2000
- Head teachers in schools adopting PL2000 and those choosing not to
  - PL2000 Referenti
  - Teachers and others working on new materials for the PL2000
- Staff involved in organising the courses and exams in the schools
  - (rooms, teachers, media, other resources, payments etc)

### 5 different governments (2 parties)
- 4 Prime Ministers
- 4 Ministers of education
- Berlinguer and Moratti reforms
- PL2000 nucleo - Sanzo
- Funds made available in the PL2000 directives
- MPI and other bodies including examination providers responsible for providing information
- Cultural bodies in Rome and their regional offices
- Small number of staff in MPI - PL2000 coordinator and PL2000 nucleo
- Regional PL2000 coordinators in USRs CRTs
- English, French, German and Spanish diplomats and cultural counsellors and staff in related bodies (British Council etc. as in Appendix 5.8)
- Local representatives of LEND (Lingue e Nuova Didattica), ANILS (Associazione Nazionale Insegnanti di Lingue Straniere), TESOL, IATEFL etc.
- Individual experts and consultants employed to work on aspects of PL2000
- Individuals at Italy’s 70+ universities and other HE
Stakeholders within the educational industry
- Materials writers and authors
- Publishers
- others

Other stakeholders
- Academic researchers – local/international
- The media – press, TV etc.
- Regulator and legal bodies

Competitors and imitators
- other assessment providers

Text book writers for international publishers and local publishers supplying Italian schools
Publisher representatives and sales staff

Individuals within Italian universities (e.g. at Siena University) or in other countries (e.g. the Cambridge ESOL R&V team)

Critics of PL2000 and the use of international exams

Critics of the CEFR and claims of linking of international examinations

Public interest reporters in national, regional and specialist press and TV (e.g. S Pontani, Europa Vicina)
Anti-trust commissioner in Italy

Examination providers not accredited by the MPI

Named individuals who were actually involved in the PL2000 or had a relevant stake in the outcomes (some of whom were also participants in the impact study) are listed. We can put names to the participants in the following roles:

- Ministry individuals at the MPI (Rome).
- Regional ministries and support staff (at the USRs).
- Head teachers and senior school staff in the schools adopting the PL2000 and especially in the case study schools.
- Language Teachers and Assistants in the PL2000 classrooms and especially in the case study schools.
- Parents of the students in the PL2000 schools, especially the case study schools.
- Students (as learners and candidates for the exams) in the PL2000 and especially in the case study schools.

The impact study was able to focus on the learners/test takers by monitoring their performance in the examinations (quantitative data) and by listening to their views in the qualitative data. However, the extent to which their learning experiences and the
outcomes achieved were enhanced by taking *specific* examinations did not emerge in the data. There was evidence that the macro context and conditions created by the PL2000 increased the motivation of teachers and learners to achieve their stated objectives, but the research was not able to tease out the individual differences between the learners and the extent to which examinations impacted on their success in learning the language, and if it did, in what ways this was manifested. In order to address these issues a closer investigation of the learner, the learning context and the factors influencing progress would be needed. This is taken up in Chapter 6.

The role of the examination providers proved to be an interesting dimension of the PL2000, both as institutions and as individual participants in PL2000 activities. Several Cambridge ESOL staff participated in delivering goals of the PL2000 under the terms of the memorandum with the Ministry and also took part in the impact study.

Two memoranda were signed between the examination providers and the Ministry in January 2000 and January 2002; the document shown in Appendix 5.9 is the second. It is important to note that none of the boards was directly involved in planning the assessment “segment” within the PL2000; discussions prior to announcing the project had been conducted by the Ministry staff with the cultural bodies in Rome through membership of a coordinating committee (referred to as a “comitato di coordinamento” in Ministry documents). The cultural bodies recommended which examination providers should be contacted to sign the memorandum of understanding (*protocollo d’intesa*). In the case of English, several providers were proposed by the British Council, none of which was considered the “official” UK provider, although all boards had been accredited by the QCA in England, i.e. Cambridge ESOL (then UCLES EFL), Trinity College London (TCL), City and Guilds International (Pitman) examinations, Edexcel (London examinations) and the English Speaking Board (ESB).

The examinations boards were obliged to take an *ex-post facto* perspective (Kerlinger, 1973: 379) on the PL2000. Likewise the Cambridge researchers in the impact project started from the “current state of affairs” and had to work back to establish possible explanations and causes for what was happening. All the boards responded to the Ministry by offering existing examinations which appeared to fit the requirements at the time. The CEFR made it possible to put forward arguments to justify and explain their use in Italian state schools.
As “outsiders” and in the timeframe set by the Ministry, the examination providers had no opportunity to develop new tests to match the intended context and uses within the PL2000. In Messick’s term, they could not “seek validity by design as a likely basis of backwash” prior to entering into the agreement with the Ministry (Messick, 1996; 252). Evidence from the impact study suggests, however, that the examinations did fit into the aims of the project and that there was evidence of “reverse impact” on the examinations during the implementation phase (i.e. where some examination providers adapted their systems to fit the Italian context - noted above as “impact on the test providers”). Many of the teachers interviewed seemed happy to “teach to the tests” with an understanding that this was convergent with their other teaching goals. So, despite the lack of prior matching of the examinations to the context and purpose, the washback was generally positive (cf. Wiliam, 1996 cited in Chapter 2). This was true for the teachers using Cambridge examinations in the case study schools and may have applied more widely (e.g. to other examinations used in PL2000 schools).

5.7.6 Conclusions from the PL2000 case

The PL2000 project started with specific educational reforms and an explicit recognition that assessment (and external examinations) would be one segment of the overall objectives. The exact details of examinations to be used were less important for the Ministry planners than the general nature of the assessment to be introduced and they relied on recommendations from the cultural institutes.

The introduction of external examinations from other countries represented a major innovation and a major impact. Stakeholders needed to understand and accept the implications of such a change, not least because of the potential for the schools and teaching to be evaluated on the basis of the outcomes achieved in the tests (the possibility of ‘league tables’ being compiled). This placed a responsibility on the examination providers to deal with the communication challenge and to ensure that misunderstandings and possible fears could be dealt with effectively. There was ample potential for resistance from stakeholder group, not least from those who held political and ideological beliefs which opposed the use of public funds for ‘foreign exams’ or saw external examinations as usurping the traditional role of teachers. While anecdotal evidence of these attitudes emerged in the Cambridge study, they were not found in the case study data as the schools involved were broadly in favour of the PL2000 provisions.

From the Ministry’s point of view, the anticipated washback effects at school-level were changes to teaching and learning linked to the principles of the CEFR and an
acceptance that learning outcomes should be linked to externally verifiable criteria. The linking of curriculum objectives to CEFR levels and the wider dissemination of the communicative approach using up-to-date methods and materials (building on earlier projects dating back 20 years) were clearly stated objectives.

Political issues are always important in educational policy, not least because of the ideological concerns and the changing political landscape. Governmental changes and subsequent changes of personnel in key positions (e.g. civil servants) often mean that projects which need time to be fully ‘installed’ are undermined and long lasting effects can be “watered down” or the outcomes perverted. Short-term political expediencies and manoeuvres within government and within relevant ministerial departments can influence timescales for implementation of projects, as we saw in the case of the PL2000. The Ministry team had to work rapidly to convert policy into concrete action in the first year of the project (1999/2000) with budgets and funding already in place. The timescale was particularly short in setting up the external certification and the examination providers were given 10 weeks (November to January) to respond to the Ministry’s request for examinations to be used later in the same academic year. The impact of the policy and the practical constraints in delivering it within the time allowed meant that the examination providers had to work hard to adapt their systems and to communicate effectively with potential test users (the “reverse impact” noted above). The deliberate decision by the Ministry to create an “open market” for the certification (especially for English) meant that the examination providers had to differentiate the key features of their tests (both the underlying constructs and the practical features and benefits) from the others. This was particularly noticeable for the Cambridge and the TCL exams. The need to provide relevant information and materials to the “school market” meant that the Ministry benefited from a large amount of additional resources in disseminating information about the PL2000 in general, and in relation to the key tenets such as the criterion-levels of the CEFR. A major impact of the project was the extent to which the CEFR became known in Italian language teaching circles compared with other European countries.

The interface between national priorities and international concerns was noticeable in this Italian example, and the long-standing debate on the role of language and language learning in society became linked with transnational policies in Europe. There was strong support for European integration, the Council of Europe’s initiatives, and the EU agenda on multilingualism – the so-called Mother Tongue plus 2 policy. The early adoption of the CEFR in Italy (before it was officially launched in 2001) created conditions which later became a major point of contention across
Europe; Ministerial authorities required the examinations boards to link their own exams to the level system, but did not stipulate guidelines on how this was to be substantiated. The currency attributed to the CEFR through its official endorsement meant that a market place for educational services (such as teaching materials and examinations) was created where the CEFR link was an essential part.

For the Cambridge examinations, informational materials were produced which focused on construct-related matters (such as a focus on four skills and a task-based or action-oriented approach) and practical issues which were essential to ensure that the teachers and their students knew how to prepare for examinations which operationalised these features.

The context-specific modifications made by the examination providers reveal the importance of the “two-way effects” of initiatives like the PL2000; i.e. the ways in which the PL2000 changed the assessment systems, and the ways in which the examinations influenced the educational system into which they were introduced. It could be argued that the ideal would be if the two sets of changes converged to bring about predetermined outcomes on both sides. However, analysis of the PL2000 suggests that this is unlikely happen in the politically charged circumstances of national reform projects. Under real life conditions, the actual outcomes are more likely to emerge in a less predictable way; i.e. they will be “created” as a result of the project in ways which cannot be fully predicted in advance.

In developing our understanding of the impact, this confirms the need to anticipate what will happen through careful design and implementation of the testing system, including the “new realities” which emerge when the system is operational and becomes partly “owned” by other stakeholders. Davies (1997: 335) has noted the limits of responsibility of examination providers for the effects and consequences of their examinations: “In my view … the apparent open-ended offer of consequential validity goes too far. I maintain that it is not possible for a tester as a member of a profession to take account of all possible social consequences”.

Of course, some areas of accountability related to the situated use of tests need to be addressed by examination providers and this has been a finding from the PL2000 case. Assuming appropriate test design (e.g. by aligning construct-related features to the learning context), these include the need to tailor provision of information about the examination and its administrative systems to specific stakeholders and user groups, and to install adequate procedures to monitor what happens during and after implementation (and then periodically over time). This should include routine
procedures and the capture of data which are specific to the new test taking contexts.

5.8 Summary and implications for the revised model of Impact

5.8.1 General comments
The IELTS case confirmed the importance of several points in the basic washback model which have also been of relevance to the PL2000 case: the importance of predicting impact and comparing predictions with outcomes; the question of baseline data and the extent to which this can or should be collected; and the need for attitudinal data to be collected effectively. An additional dimension (represented in Figure 4.7) was that of iteration to allow for predicted impacts (intended changes) to be monitored over time and pointing to the importance of the longitudinal nature of impact studies. The IELTS case demonstrated how various kinds of impact “emerge” along with the evolution of the testing system itself over time.

The seven features of the model introduced in Chapter 4 are revisited here using insights from the PL2000 case.

i). Test features
The IELTS case confirmed that a construct-based approach to test design provides the potential for positive impact. However, the test development process also needs to incorporate an explicit “test taker model” and contextual features related to the user system to ensure that the latent potential for positive impact has a good chance of being realised in practice.

The PL2000 case showed that impact predictions based on broad construct definitions can be achieved in practice. The use of the CEFR enabled the examination providers to make claims of alignment to the level system and implementation of a communicative approach.

An important impact of the examinations was the independent nature of the assessment procedures (external to the school) linked to internationally recognised standards.

ii). Context
The IELTS case confirmed that impact research must focus on schools and classrooms (i.e. the micro context) but also should be extended systematically to include macro contexts and the wider social dimensions. This has been the main theme of this chapter: the PL2000 case examined how language examinations fit
within a specific educational system during a process of reform. The meta-analysis confirmed the complexity and dynamic nature of such systems and that it is not realistic to expect causal relationships to be established (as noted in Chapter 2).

In educational reform projects baseline data will be difficult to capture because of the timescales operated by Ministries of Education, but this does not mean that attempts to capture data early in the process should not be attempted as comparisons need to be made to provide evidence that anticipated changes are taking place. Models of diffusion of innovation, such as Henrichsen’s model are useful in this respect.

Italy’s educational system and the politically motivated reforms demonstrated the importance of understanding the inter-relating sub-systems in which educational change takes place. An important feature of the Italian milieu was the political landscape and the changes in government which occurred at the time of the PL2000, shown in Table 5.10. Research methods to investigate impact in such contexts need to reflect the socio-political nature of the activity.

<table>
<thead>
<tr>
<th>Party A – centre-left Government 1 Prime Minister 1</th>
<th>Party A – centre-left Government 2 Prime Minister 2</th>
<th>Part A – centre-left Government 4 Prime Minister 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minister of Education 1</td>
<td>Minister of Education 2</td>
<td>Minister of Education 4</td>
</tr>
<tr>
<td></td>
<td>Minister of Education 3</td>
<td></td>
</tr>
</tbody>
</table>

iii). Participants
In an expanded model of impact, a wider range of participants needs to be taken into account beyond the classroom/school context. The PL2000 case identified both the categories of participant as well as the specific individuals who were instrumental in implementing the objectives.

In assessment contexts where new examinations are introduced as innovations into educational systems, it is important to identify who the “change agents”, “earlier adopter” and “project champions” will be; their active engagement with the
examinations will be a factor in successful diffusion. Effective communication between participants and dissemination of information is vitally important, with the need to “tailor” messages to the interests of target audiences (participants with different political, administrative, educational, or institutional roles).

The PL2000 classroom data revealed that awareness of and interest in the project objectives was not adequate to ensure that these were implemented in practice; for example, in the absence of training, teachers experienced difficulties in changing their teaching practices to implement communicative methods in their classrooms. This area of the model needs to be expanded by looking in more detail at the impact of examinations on individual learners at the micro-level.

iv). Outcomes
Outcomes should be based on a clear understanding of the current situation and the anticipated changes. The consequences of the changes may be expected to appear over an extended period of time and in different ways (immediate or delayed: visible or hidden). Not all participants adopt changes at the same time or to the same extent, and some participants are more successful than others in understanding the requirements and in delivering the desired outcomes.

This means that outcomes in terms of the consequences of using examinations – intended or unintended - cannot be wholly managed or controlled; the validity of the examination in real-life contexts of use will emerge only after it has been fully “installed”. These contexts will not be uniform and will be subject to constant changes as a result of socio-political and other influences. Examination providers, therefore, cannot be held to account for all the effects and consequences of the examinations they deliver. On the other hand, it is possible for them to play an active role in promoting good practice by communicating effectively with stakeholder and by monitoring the processes of change when an examination is being introduced into a new context.

It can be argued that the impact of a newly introduced examination should be anticipated so that changes which might be attributable to it can be effectively monitored, as demonstrated by the quantitative analysis and case study data in the PL2000 case. It was possible to monitor the behaviour of participants – their decisions, actions, views and attitudes - and observable changes to curriculum/syllabus, new materials and other products. In order to develop this approach, a capacity to conduct in-depth case studies is required and this is one of the main themes in Chapter 6.
v). Roles of researchers
A greater understanding of insider/outsider roles has become an important issue as the impact model develops. This is also taken up in more detail in the next chapter.

vi). Research methods
The PL2000 Impact Study confirmed the need to “problematise” impact in terms of hypotheses linked to anticipated outcomes. By establishing these hypotheses, specific research questions were drawn up and investigated using the approach developed for IELTS Impact projects; this provided procedures and a “toolkit” which was readily adapted to the new context (i.e. impact research methodology, instrument design and validation procedures). Mixed methods of data collection were employed including qualitative techniques and case studies, as well as more traditional survey data and the collection and analysis of demographic and score data. The case studies (conducted longitudinally) provided insights into the complex interactions between individual and systemic variables within different contexts where Cambridge examinations were being used as part of the PL2000.

This confirmed that impact researchers should develop or adapt appropriate instruments. It is likely that guidance in this area will come from disciplines outside of language testing, drawing for example on innovation theory in education and other branches of the social sciences.

Data collection of various kinds continued (in the case of Cambridge ESOL) after the end of the Impact Study but further case studies were not widely carried out. The cost and complexity of such research remains a difficulty and the data which are generated do not always get fully exploited. Ways need to be found to make such data more easily accessible and susceptible to analysis.

vii). Timeline
Again the need for the timeline to be a focus in impact research was confirmed in order for comparative data to be collected (e.g. in time-series designs) and for changes to be monitored over longer periods than in previous studies.

The next chapter examines Case Study 3 which follows on from the PL2000 Impact Study.
CHAPTER 6

Case Study Three

The Florence Language Learning Growth Project

6.1 Introduction

The focus in this chapter moves to the micro context of a single institution and the classes of learners and teachers within it. This is represented in Figure 1.1 - "context in education" - by the "exploded" red dot at the centre.

Figure 1.1 from Chapter 1 (also discussed in Chapter 5).

See also discussion of hierarchical educational models in the educational literature, such as Goldstein, 2003 or Raudenbush & Willms, 1991.

A study of the complex dynamics which exist between a single school and the wider educational context in which it resides, and the inner workings of the school itself (e.g. within specific classrooms) provides the data for the meta-analysis in this chapter, i.e. an impact project conducted in one language school in Italy.

The study focused on the macro-micro interface, and applied the principles and practices of case study research and mixed method designs, as discussed in Chapter 3 (methodology).
This single case study is known hereafter as the *Florence Language Learning Growth Project* (FLLGP). Building on the analysis of the IELTS and PL2000 cases, it formed part of the wider investigation of impact in the Italian context as discussed in Chapter 5. As a meta-analysis it provides the final input for the developing model of impact in this thesis. The FLLGP case study is of particular relevance to the developing model in that it focuses specifically on micro aspects of impact, including the dynamics of a single school and the participating individuals - the learners (as both students and as test takers) and their teachers. It thus examines in greater detail an aspect of impact studies dealt with only peripherally in the previous chapters.

Our analysis in this chapter allows us to explore in greater depth the benefits of case study methodologies as part of the repertoire of the impact study research tools discussed in Chapter 3. It also allows us:

i. to focus in more detail on the individual differences of the participants, (the individual learners in both learning and test taking contexts and their teachers), and

ii. to consider how an investigation of these differences might inform the way that tests and testing systems actually influence the test takers as language learners.

This second point is often neglected in the washback studies discussed in Chapter 2, yet it is one of the questions which most interests educators. Although the test taker is considered to be the main participant in a test event, and the primary user of the results of the assessment made of his/her abilities (as set out in the *ALTE Code of Practice* discussed in Chapter 5), little is known about the way the testing experience influences him/her, both as a language learner, and as an individual in other ways, e.g. in developing personal views, attitudes, motivations etc.. As Wall (2000: 02) has pointed out with regard to washback research: “missing ….are analyses of test results which indicate whether students have learnt more or learned better because they have studied for a particular test.”

As well as the individual learner differences, the micro-level case study also allows for a focus on the school context (atmosphere, culture, resources etc) and on the individual teachers. It is clearly the case that teachers exert influences on the language learning and on other behaviours of their students and the individual differences between teachers can therefore be seen as important moderator variables in relation to the learners in their classes. The following questions may be posed:
• Which teacher variables play a role in the how and the what of test preparation?
• What effect do they have on the learners in the classroom and ultimately on how the learners behave in the examination itself as test takers?
• Are there systematic effects which can inform future impact studies and the development of the revised impact model?

This chapter, therefore, will examine closely the school and classroom context and the relevant literature and work which has been done to develop appropriate instruments to be used in researching individual teacher and learner differences, (e.g. the LLQs introduced in previous chapters). This area of research also takes us into the fields of “assessment for learning”, hierarchical modelling, latent growth analysis, and constructionist theories of learning and identity, such as socio-cultural approaches to second language acquisition and personal construct theory.

The data for the meta-analysis in this chapter have included: project notes dating back to Summer 2002, mainly kept by the Project Coordinator (see below for details); records of meetings and conversations in Italy and Cambridge between members of the project team, including the current author; and interim project reports. The raw data (paper copies of questionnaires and tests, recordings and videos, etc.) and the summaries and analyses in electronic format were made available from the archives of Cambridge ESOL: these data are summarised below and in Appendices 6.2 to 6.8.

6.2 Background to the FLLG Project

Set against the macro context described in Chapter 5 and the growing use of external certificated examinations for English in Italian schools, the FLLGP was a follow-up phase to the PL2000 Impact Study (as noted in Chapter 5).

One of the key features of the PL2000 study was that the investigations were carried out in state schools with the collaboration of the teachers and pupils. The seven small-scale case study schools provided rich and informative data which shed light on the way that the intended innovations were taken up in practice during the PL2000 project. The PL2000 Impact Study team agreed that, in light of the success of these studies, it would be valuable to carry out an in depth “language learning gain” project with a single institution targeted as a case study involving a larger number of students and teachers than was possible in the PL2000 studies. The intention would be to conduct a very thorough analysis (“thick” description) of the case and to include perspectives from a range of participants and observers (Miles and Huberman, 1994).
Discussions took place with a number of possible teaching centres in Italy where such a project might be feasible and as a result it was decided in July 2002 that the project should be carried out at the British Institute in Florence (BIF). As in the previous impact studies, practical concerns played a part in making this decision.

The British Institute is an important English language teaching and cultural centre in Italy. It was considered appropriate for the follow-up case study as the English language courses were already aligned to the levels of the CEFR and catered for all ages of student (beginning from age 5, in line with the trend set under PL2000). It was also the largest examination centre in the Tuscany region, providing preparation for Cambridge examinations at various CEFR levels. Although not a state school, it had been actively involved with the PL2000, and as the local Cambridge centre, it had been responsible for organising many of the Cambridge examinations which had been taken in the schools in the Florence area as part of the PL2000. For more information on the British Institute, see Appendix 6.1.

It was intended that the project should take a longitudinal perspective and should span the whole of the academic year from October 2002 until June 2003. Data were to be collected at several points in the school year, with a focus on the learning of the participating students and the effect of test preparation within the specific school and classroom contexts.

The Director of Studies of the British Institute, who was also the Local Secretary for the Cambridge examinations, was consulted early in the planning and it was an important factor in getting the project off the ground that she was enthusiastic about the project and its objectives. She was able to seek active support from teaching colleagues and to secure the necessary access to the classes and the individual participants within the school. The Director of the British Institute gave permission for the study to take place, subject to the written agreement of the teachers and learners. The support and professional interest from management, as well as the participation of the most experienced staff (including DOS and ADOS), were important prerequisites for conducting this study. In case studies and ethnographic research, the importance of having members of the research team who are also participants in the activities under observation (noted in previous chapters) was confirmed again in this case, i.e. observers as participants and participants as observers.

The FLLGP was partly a response to wider discussions which had taken place between Purpura, Saville and Hawkey in Summer 2001 in order to advance existing research using the LLQs. A research project was proposed on the relationship between test takers’ strategic and socio-psychological characteristics and their language test performance (Purpura, 2001) and a longitudinal study was suggested in order to investigate the
relationships between learners’ background factors and language test performance, (i.e. with repeated measures taken at the beginning, during and at the end of the school year). Such a study would provide the possibility of conducting latent growth research using structural equation modelling (SEM) to estimate the growth trajectory (see Meredith & Tisak, 1990 – see Appendix 6.9). These discussions influenced the initial design of the FLLGP, but given the scope and number of potential participants, it was not possible to implement the statistical dimension of the proposed project using the SEM analysis. The project therefore developed along the lines of the PL2000 studies with a mixed methods approach involving the collection of quantitative data (test scores, questionnaire responses etc.) and observations and interviews conducted using qualitative methodologies.

6.3 Delimiting the case: the unit of analysis

The unit of analysis in this chapter is the specific impact study known as the FLLGP. In contrast to the cases in Chapters 4 and 5, this study is a single case study, and as such, easier to delimit as a unit for the meta-analysis; the timeline, the context and participants are also more compact, facilitating precision and more detailed description. The detailed focus on people and events is particularly relevant in this case, given the micro level focus.

6.3.1 Context and Timeline

In early discussions with the then Director of Studies of the British Institute it was agreed that a feasible scope for the project might be 8 to 10 classes comprising about 100 students at B1 (FCE) level. The aim would be to replicate what had been attempted in the PL2000 state schools, but this time under more controlled conditions, using the structured data collection modes and instrumentation employed in the PL2000 case study schools. As it was a private language school, BIF classes were voluntary (and fee paying) and the class sizes smaller than most state schools (although comparable with the extra-curricular groups which were set up in many schools within the PL2000).

In subsequent planning meetings (see notes of planning meetings between Saville and Hawkey, August 2002), it was agreed that the scope could be extended to include both young learner and adult classes at two CEFR levels (B1 and B2). For each group, some of learners would be intending to take the relevant Cambridge examination at the end of their courses (i.e. PET or FCE) whereas others would not be planning to be externally assessed. Those not intending to take the examinations could, therefore, be considered to be small “control groups” for comparative purposes in the quasi-experimental design.
The context and timeline represented in Figure 6.1 (in a similar way to the previous case studies in Chapters 4 and 5) show twelve selected class-based courses leading to outcome measures at the end of the course aligned to two CEFR levels, B1 and B2. These twelve classes were chosen from an existing range of courses on offer. The timeline represents a single academic year (about 8-9 months). It was important for the project to start promptly in October 2002 as an immediate follow up to the PL2000 precluding a long planning phase. The opportunity to gain a better understanding of the interactive processes between the micro and macro contexts was an important consideration and the information obtained from the PL2000 impact study was still relevant and provided a well-described macro context (as a kind of baseline for the study) without a requirement for collection of additional data.

As the school year progressed, the examinations became imminent with the potential for test preparation effects to become more marked (the possible “seasonal washback” suggested by some researchers – Bailey, 1999, Watanabe, 1997).

This case study is not an example of an educational innovation involving new examinations, as in the PL2000 schools. The Cambridge examinations had been available for many years in the British Institute and were an established part of the school’s provision. The antecedents included the wider use of externally certificated examinations within the educational milieu as a result of the PL2000.

6.3.2 Participants
At the start of the project (October 2002), the potential cohort of learner participants was a target group up to 139 students - adults and younger learners - enrolled in the twelve classes at CEFR levels B1 (PET) to B2 (FCE).
Younger learners are defined by the BIF as children (aged 11 to 12) and teenagers (from 13 to 15). Adult courses at the BIF include older teenagers (16+) and adults of all ages (although predominantly between 20 and 50 years old).

Within the BIF’s own course structure (level, age, examination type), each class group is identified by a letter and a number: J4, J5, M4, M5, E3, E4 etc. The younger learners were in courses of 90 hours and the older learners in courses of 120 hours. (See tables in Appendix 6.2 for British Institute class structures).

As these were real classes (not experimental or control groups), changes and fluctuations were not under the control of the project team; such fluctuations and changes are typical of real world research and a notable feature of impact research in school contexts as we have seen in earlier chapters (see also comments in teachers logs – Appendix 6.6). During the academic year, the number of learners remained quite stable, there were various minor changes in the classes and therefore in the total number of students involved. Participants also changed their original intentions to take part in activities and also to enrol for particular examinations.

Each of the twelve FLLG project classes which were identified as suitable and for which the teachers agreed to take part, was composed of 10 to 12 different learners, enrolled in the standard BIF courses at B1 to B2 levels. Summaries of the “learner profiles” were completed by the Project Coordinator from the information given by the learners themselves and data held by the school.

In addition to the learners, the other main participants in the school included the Director of Studies (DOS) and the participating class teachers, one of whom was also the Assistant Director of Studies (ADOS).

Appendix 6.2 summarises the teachers, the number of learners, the BIF class names, the target levels with reference to the CEFR, and the intended Cambridge examinations where applicable. Nine classroom teachers (referred to hereafter as Teachers A to I) agreed to take part in the project. Five of the teachers were female and four were male. Three teachers made a more substantial contribution than the others in that they participated in the project twice, i.e. they taught two different classes each: Teacher B (PD – male), Teacher C (JG – female), and Teacher E (AL – male). This provided an opportunity for comparison across different learner groups with the same teacher.
Eight of the nine teachers completed a teacher profile questionnaire and agreed to have their classes observed and video-recorded. One teacher (Teacher G, a YL specialist and teacher trainer) did not agree to complete the questionnaire or to be filmed (see summary of the 11 FLLGP/BIF class groups in Appendix 6.3).

Beyond the participating classes, several other individuals involved in the administration of the Cambridge ESOL examinations at the British Institute were important in facilitating the data collection and the practical arrangements for the visits of the Project Coordinator and others.

6.3.3 FLLGP scope and research design
The project’s scope and design can be summarised as follows: it was a small-scale, longitudinal case study to investigate relationships between background factors and learner behaviour (including language performance), at the beginning, during and at the end of the school year where some students were preparing to take a Cambridge ESOL examination and others not.

A common feature across all the classes in the school was the use of the CEFR to establish proficiency levels and to guide the course content and methodology (i.e. broadly speaking a communicative approach to language teaching).

There was an explicit focus on “growth” (see below for discussion) and, while the project was influenced by quasi-experimental research designs with some attempts to control, or at least account for, the different groups and the intervening variables, it employed data collection techniques consistent with case study designs and a mixed method approach, i.e. questionnaires, structured and in-depth interviews, classroom observations and document analyses (e.g. teacher logs, etc.). Attempts were made to address problems associated with case study research noted in earlier chapters, including: a potential lack of clear aims and implicit rather than explicit theoretical input; difficulty in establishing causal relationships; inadequate piloting of data collection techniques and instruments which may thus contain invalid items; bias and other uncontrolled interviewer/researcher effects.

The project design and the role adopted by the Project Coordinator attempted to address these problems and set out to adhere to the requirements for conducting successful case studies – as discussed in the literature (e.g. Yin, 2003); and noted in Chapter 3.

The FFLGP was therefore characterised by the following features:
• a clear and approved research plan with explicitly stated working hypotheses;
• the use of validated data collection instruments and instruments (in this case adapted from the previous related projects);
• triangulation of data through the use of information collected from a range of sources (including teacher questionnaires and interviews, student and teacher questionnaires and classroom observations);
• the referral of data collected by the main researcher to other experts in the team for checking.

The project records show that the starting point was the research design and working impact hypotheses which then led to the action plan. We should remind ourselves at this point that in formulating hypotheses about impact in a school/classroom setting, including traditional washback hypotheses, an understanding of the classroom variables and the relationships between them is essential. As we have noted in previous chapters, direct or causal relationships between test and teaching/learning are unlikely to be observed but the relative effect of the numerous, complex and inter-related intervening variables needs to be better understood (for recent discussions of washback variables see Spratt, 2005: Green, 2007; for a wider discussion of moderator and mediator variables in social research see Baron and Kenny, 1986).

The characteristics of the different people (participants/stakeholders) and the specific language classrooms where test preparation takes place are important intervening variables. The three main categories of intervening variables from earlier washback models (e.g. Bailey, Green) which we have focused on are: contextual variables, learner variables, and teacher variables. These were a main focus in the FLLGP.

The meta-analysis is structured around these three variables, represented schematically in Figure 6.2 derived from the ELT literature and recent research in general education, discussed below.
Figure 6.2 Nested levels/contexts

The “nested” boxes in Figure 6.2 represent a typical hierarchical model of educational contexts where an individual school is embedded within the wider milieu (macro context). Within the school there are classes with teachers, and within the classes there are learner groups comprised of individual learners. Classes can be compared on the basis of the class type (curriculum, level, age group etc.) and also on the basis of specific features within each class (the teachers, the learners, the materials used etc). Classes and/or learner groups intending to take a particular examination can be compared with those who intend to take a different examination or no examination at all. The red arrows show the direction of the hypothesised effects of an external examination on the teacher and on the individual learners within test preparation groups. The figure also shows the results of tests or examinations of this kind are typically used outside of the classroom and the school by “secondary users” in the wider milieu.

The FLLGP design allowed for comparisons to be made between learners in different groups. Comparisons could be made between different classes and between individuals and groups within the same class, e.g. learners not intending to take the external examination compared with those intending to take it (see Appendix 6.2 for class/group comparisons). As we have seen in earlier chapters, access to data for comparative purposes is an important aspect of impact research.
In the BIF case, the assessments included the relatively high-stakes, external examinations from Cambridge ESOL (FCE/PET), as well as lower stakes, institutional tests used in the school to assess the learners less formally. The adoption of the Cambridge examinations as an alternative to BIF’s own internal tests formed part of the school tradition. The use of the Cambridge examinations was not an innovation.

The relevant learning gains achieved by the learners over the course of the year were considered to be of two kinds: improvement in language proficiency based on the CEFR level for each class groups, and other changes in their language learning behaviour, including their learning strategies, attitudes and motivations. The learning outcomes were assessed by the BIF examinations and relevant Cambridge examinations. The individual teachers provided important intervening variables in mediating the requirements of the chosen examinations, thus impacting on the learners as test takers, as well as on their behaviour in other ways.

Consideration of all these variables provided the basis for the specific impact hypotheses and research questions in the FLLGP (Hawkey, 2003):

- how do the learners/teachers feel about the different examinations and tests being used and how does this influence their behaviour?
- what happens in the exam/test preparation classroom compared with other classrooms?
- what materials are used in test preparation classes and how do they relate to the curriculum/syllabus?
- how are these materials used and why are they used in those particular ways?
- how do the learners go about their learning in test preparation groups? For example how much do they learn of what is taught and at what pace and how does this differ from other groups?
- how does the specific classroom experience affect the learner’s exam/test behaviour and ultimately the outcome scores?

6.3.4 The action plan and data collection schedule

As in the IELTS and PL2000 studies, the importance of an explicit action plan accompanied by appropriate project management techniques had been recognised for the FLLGP. As in the PL2000 case, the UK coordinator was Roger Hawkey (RH), thus building on his past experience and providing continuity with the previous studies. He was again supported by staff in Cambridge.
As well as organising the project in the UK, provision was made in the action plan for the time series design involving three site visits to the BIF by the Coordinator (Figure 6.3).

Figure 6.3  Visits of Project Coordinator

The plan clearly set out the activities for collecting data at these three moments of the school year – i.e. at the start before much new teaching/learning had taken place; during the school year; and at the end of the school year. The plan also set out clearly the roles of the other protagonists, including individual participants within the British Institute itself. On the first two site visits the Coordinator was accompanied by the Cambridge ESOL Country Manager (Liam Vint - who had previously participated in the PL2000 case study) who facilitated the interaction between the British Institute and Cambridge. The purposes of the three visits (24 to 26 November, 9 to 11 March, and 11 to 14 May 2002) were as follows:

- to liaise on project progress and to address any problems with Director of Studies who was acting as the “on-site project manager”;
- to interview participating British Institute teachers on student progress, language learning, teaching and testing approaches, and Project objectives, procedures and instrumentation, as required;
- to observe and video-record lessons with the project classes;
- to collect as much performance data as possible from the participants, including samples of the students’ work;
- to review and agree the future actions of the project as it progressed.

The data collection instruments that were used were adapted from the “impact toolkit” which had been assembled for use in the PL2000 case study schools. These are summarised in Table 6.1 below. The data collection focused on the three main variable sets indicated above: contextual, learner and teacher. The language proficiency variable was measured at the beginning and the end of the year with tests and examinations anchored to each other and aligned to the CEFR levels on which the courses were based. It was intended that this aspect of language learning gain, a main focus of the study, would be assessed in this way.
The table also shows the three phases of data collection, the number of participants from whom data was collected and the format of the data as it was collected and stored for further analysis.

Table 6.1 Summary of data collected: before, during, after

<table>
<thead>
<tr>
<th>Before</th>
<th>Number</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-course data collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CommuniCat (1) - computer-based English language level test</td>
<td>119</td>
<td>Logfiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scores in Excel format</td>
</tr>
<tr>
<td>LLQ (1)</td>
<td>118</td>
<td>Logfiles</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scores in Excel format</td>
</tr>
<tr>
<td>Can do - self-assessments (1)</td>
<td>110</td>
<td>Word files</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scores in Excel format</td>
</tr>
<tr>
<td>Student Questionnaires (1)</td>
<td>134</td>
<td>paper copies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analysed by project coordinator</td>
</tr>
<tr>
<td>Teacher Questionnaires (1)</td>
<td>8</td>
<td>Word file</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Analysed by project coordinator</td>
</tr>
<tr>
<td>Student speaking samples</td>
<td>75</td>
<td>Audio cassettes</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scored by project coordinator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scores recorder in Excel format</td>
</tr>
<tr>
<td>BIF placement test</td>
<td>25</td>
<td>Paper copies</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scores in Excel format</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>During</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>During course data collection</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Teacher logs and/or progress reports to be analysed qualitatively to</td>
<td>142</td>
<td>Wordfiles</td>
</tr>
<tr>
<td>produce summaries</td>
<td></td>
<td>Analysed and summarised by</td>
</tr>
<tr>
<td></td>
<td></td>
<td>project coordinator</td>
</tr>
<tr>
<td>BIF student registration data</td>
<td>All</td>
<td>Paper copies</td>
</tr>
<tr>
<td>Target group lessons observed and analysed over the three site visits</td>
<td>26</td>
<td>Video tapes</td>
</tr>
<tr>
<td></td>
<td>classes</td>
<td>Analyses with Word commentary</td>
</tr>
<tr>
<td></td>
<td></td>
<td>by project coordinator</td>
</tr>
<tr>
<td>Class registers and mark sheets</td>
<td>All</td>
<td>paper copies with Project</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Coordinator</td>
</tr>
<tr>
<td>Students recorded and/or videoed speaking in class; scoring by Project</td>
<td>80</td>
<td>cassettes, audio and video with</td>
</tr>
<tr>
<td>Coordinator, to be entered on his Excel</td>
<td></td>
<td>Project Coordinator</td>
</tr>
<tr>
<td>Student writing samples (1)and (2)</td>
<td>Paper copies for 90 students</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scored by project coordinator</td>
</tr>
<tr>
<td></td>
<td></td>
<td>75 scores in Excel format</td>
</tr>
<tr>
<td>Student mock exam results</td>
<td>121</td>
<td>121 exams - scored by Teachers</td>
</tr>
<tr>
<td></td>
<td>exams</td>
<td>121 paper copies of result</td>
</tr>
<tr>
<td></td>
<td></td>
<td>entered into Excel by</td>
</tr>
<tr>
<td></td>
<td></td>
<td>project coordinator</td>
</tr>
</tbody>
</table>

| After                                                                 |        |                                 |
| Post course data collection                                          |        |                                 |
| CommuniCat (2)                                                       | 29     |                                 |
| LLQ (2)                                                              |        |                                 |
| Can Do - self-assessments (2)                                        |        |                                 |
The instruments used, mainly adapted from the earlier impact projects, were consistent with the mixed methods approach; they included questionnaires, logs and other forms. An “oral analysis form” for use by Project Coordinator was specially prepared for the analysis of the samples of spoken language which were recorded.

6.3.5 Data collection

The data collection schedule was ambitious and depended heavily on the co-ordination skills of the Project Coordinator and on the goodwill and cooperation of the participants within the school. As we will see below, some parts of this plan worked better than others.

a). Learning gain and level of proficiency

Proficiency gain in English was one of the main dependent variables; i.e. changes in the learners’ knowledge and use of English which could be attributed to their participation in the BIF courses and their candidature for the various examinations. The repeated measures in the project included an independent proficiency measure – taken before and after the courses – using Communicat (Cambridge ESOL, 1999), a test which claims alignment to the CEFR levels. This was intended to function as a common “anchor” to facilitate comparisons of the learning gains across the different classes but only a limited number of learners agreed to retake the test at the end of the course.

The assessment of the initial proficiency level of the learners was also made using the BIF placement test, used to assign learners to suitable groups. Scores on both these tests were supplemented by the collection of a spoken English samples on audio tape for comparison at a later stage (as neither of the other tests focused on speaking ability).
During the courses, evidence of the students’ writing ability was also collected from the learners on two occasions and rated by the Project Coordinator. Where mock examinations were used to monitor progress and/or readiness for the final assessments, the teachers who made the assessments provided the scores for the project coordinator.

The assessment of the proficiency after the course was either made using the BIF end of course examinations or one of the relevant Cambridge examinations – PET or FCE. Eighteen students took both. Thirty five candidates were recorded during their PET or FCE Speaking tests to provide samples of spoken language at the end of the course.

Finally, the learners had an opportunity to self-assess using the Cambridge Can Do Self Assessment instrument: the majority of learners completed this at the start of the project, but again it proved difficult to get their teachers to have them repeat this measure at the end of the-courses.

b). The characteristics of the participants – learners and teachers
Data was collected from all the participants using questionnaires and interviews. Appropriate ethical guidelines were followed in order to ensure that the aims of the research were clearly understood and written permission was obtained to collect and use the data.

Basic demographic data about the learners was made available to the project team from the BIF student registration database. The learners also completed a questionnaire at the start of the course and at the end (Student questionnaire 1 and Student questionnaire 2) and there was a specific Young Learner (YL) questionnaire. Those who took a Cambridge examination at the end of the course also completed the standard Candidate Information Sheet (CIS), which all Cambridge ESOL candidates complete (i.e. comprising standard demographic data). A version of the LLQ was to be completed by learners at the start of the project. However, as for the Communicat and the Can Do measures, it proved difficult to persuade administer it at the end of the course (a summary of LLQ responses obtained at the beginning of the year is provided in Appendix 6.8).

The teachers completed a Teacher Questionnaire and Teacher Profile and during the course, they completed logs and/or progress reports (some more fully than others). The teachers’ logs were based on a pro forma provided for their use; these were
analysed qualitatively by the Project Coordinator to produce summaries. Class registers and mark sheets were also made available to the Project Coordinator.

In some respects the teachers could be considered a relatively homogeneous group; they were all native speakers of English with relevant ELT qualifications and between 3 and 23 years of teaching experience. Four of them carried out other duties in the school, including roles such as Head of English, Senior Teacher and Resource Coordinator. Their responses to the questionnaire (Appendix 6.5) revealed features in common with respect to their approach to teaching. They tended to favour learner activities in class consistent with an orthodox communicative approach. In general oral/aural activities were preferred in class to writing, which tended to be given for homework. All the teachers “frequently” used pair-work discussions and most also used group discussions “frequently” or “quite often”. Opportunities for learning outside the classroom were encouraged, e.g. “reading story books” and “watching videos”. None allowed considerations of final examinations to become too dominant and “practice for” or “discussing examinations” tended to be done “occasionally” rather than frequently. Their views on how well the school achieved “key objectives” (e.g. in line with the PL2000 and the use of the CEFR) showed that they felt that these were generally met “well” or “very well”.

The relatively small number of teachers involved in the study (compared with the relatively large number of learners), also meant that it was possible to profile them in some depth, observing some significant differences in their approaches to their classes and in their participation in the study. Three teachers were very enthusiastic participants and were willing to include two of their classes in the study; one was unwilling to be observed or recorded. Some of the teachers completed very full logs and commentaries whereas others provided short notes without much commentary. All expressed important personal beliefs about their own ELT practices, and some reflected on how their teaching and attitudes had changed in the course of the year (although two noted “no significant change”).

c). The classroom context: classroom observations and classroom data
What happened in the classrooms (the atmosphere, relationships and events that took place) and the potential effects that the individual teachers had on the classroom and on how the learners behaved in class (and possibly in the external examinations) was captured through the teacher logs and through the observations made of the classes on video. As we shall see in the discussion below, these data provided the basis for some of the most interesting findings from the study.
Twenty-six of the target group lessons were directly observed and were analysed over the three site visits by the Project Coordinator; the data comprised extensive commentary in written form compiled from the video recordings by the Project Coordinator and stored in Word format. Evidence of the learners’ speaking ability in class was also captured on the videos and was scored by Project Coordinator (scores stored in Excel format).

d). Data summary

The FLLGP Project data have been summarised in six appendices:

- Appendix 6.3: overall data summary.
- Appendix 6.4: summary of the 11 FLLGP class groups and the quantitative and qualitative analyses collected from each.
- Appendix 6.5: summary of responses from the Teacher questionnaire and Teacher Profiles.
- Appendix 6.6: learner profile summaries.
- Appendix 6.7: comments on learners from the teacher logs.
- Appendix 6.8: responses to LLQ

6.4 Outcomes of the FLLGP

6.4.1 Project reports and views of the coordinator

The Project Coordinator produced interim project reports and the project as a whole led to a range of outcomes and follow up actions. Interestingly no final report was produced and the richness of the data was not fully exploited during the timeframe of the project. The most important interim report was produced in November 2003 (Hawkey 2003) after all data had been collated and summarised. In this report Hawkey noted that the status of the project (the data and progress in describing it) was “encouraging”, and that although the early analyses were proving interesting, it was “also very challenging”. He believed that, at that point, the project was well on the way to meeting its original aim as: “… a longitudinal study of relationships between background factors and language performance, at beginning and end of the school year, with a latent growth analysis focus, and an experimental research study (pre-tests > treatment (+/- test preparation) > post-tests) of the effect of test preparation on test performance”.

However, despite the excellent cooperation received from the BIF participants (especially Sarah Ellis – DOS - and her teachers), the project had put “quite a heavy burden” on a “very busy Institute”. It had also proved to be a challenge to convince some teachers and learners (who had all paid for their courses) that the time needed
to complete certain the data collection instruments was justified from their point of view (see above). There were particular objections to completion of the “anchor measures” represented by Communicat (only 29 responses were received for post-treatment administration) and the LLQ.

The design of the FLLGP had claimed an influence from quasi-experimental research (as noted above) and there was an attempt to account for important variables (e.g. target level, intention to take an external examination, young vs. adult groups, the same teacher for more than one group or level, etc.). Variation in the dependent variable (i.e. target language performance) was observed and measured using standardised tests and examinations anchored within the same external frame of reference (CEFR).

Figure 6.4  “Experimental” and control groups in pre/post design

| “Experimental” Groups 1, 2, 3 (randomised) and control groups | Treatments (TL programme A, B or C +/-) | Nil treatment for control groups |
| Measures of target language performances of Groups 1, 2, 3 and control groups |

However, the problems encountered in controlling the intervening variables in the real world context meant that, as the project progressed, it became less feasible to maintain the experimental aspects of the design and the most valuable elements were related to the qualitative data.

Hawkey was confident that the basic learner variables were captured adequately in the project but information about the ways in which these interacted with other factors had been more difficult to capture effectively. The context of the BIF (a private and well-resourced institution, as summarised above) should have provided better opportunities than most schools in Italy to collect these kinds of data effectively, especially with the help and support of the staff. However, the variability and choices open to the participants and the caution needed in the implementation of the study because of their paying status, meant that the emerging picture was extremely dynamic and complex. Unlike those in the schools described in the PL2000 case, the BIF learners were even more varied in their age profiles, their backgrounds and their motivations for learning English.
In their responses the FLLGP the teachers showed they were aware of the individual characteristics of the learners in their classes and recorded observations in their logs with respect to a range of personality traits and other features of behaviour. They were also aware of the ways in which the behaviour of individuals affected their own learning and how they also influenced others in the class. Examples of the teachers’ records are cited in some detail here to indicate the ways in which such qualitative data can inform our understanding, in an institutional context, of some of the more individualistic aspects of test impact study.

Teacher D’s log is notable for the way in which he sought to capture the essential individual differences between the ten teenage learners in his class (BIF Class: M5, an FCE group). All members of this group were secondary school students and according to the teacher showed normal teenage behaviour and interests (“sex, drugs and rock and roll!”). He used short, colourful comments of the following kind in his general characterisations of his 10 young students:

- Best worker – a little cocky
- Quiet and very accurate
- Always asks the right questions
- “Flower child”
- Quiet and conscientious
- Habitual absentee
- Distracted, wildly inaccurate, but has fun
- Quiet and mysterious; very shy
- Hard worker
- Rampant playboy, oldest but most immature in class

These comments were, of course, “short hand” summaries, not intended to be flippant or disparaging but made, in a way natural to the particular teacher (whom we also know from the FLLGP teacher questionnaire and interview data), with a view to understanding the particular needs of the learners in his class and how each fitted within the group as a whole. The summary comments were expanded with additional comments by Teacher D, whose characterisations illustrated revealingly the wide range of different features he perceived in a group which was relatively homogeneous in other ways.

Teacher F taught a young learners group aged 13-15 years (BIF Class J4A) aiming to take PET. This teacher also made comments about the individual characteristics of the teenagers in the group and how these affected their learning or impacted on the group. For example, she commented on one of the younger learners (aged 13) who
was “very laid-back” in her approach to English, rarely doing her homework and finding most activities easy without exploiting her full potential. She speculated in her log: “I wonder if the fact that her mother is also an English teacher affects her attitude towards learning?” Teacher F noted that her group had been studying together at BIF for a couple of years and “already got on very well together”. For some, older siblings or parents were also taking classes in the BIF. There was a positive atmosphere in the class and the students were “all very motivated, particularly for young learners”; they seemed “to enjoy studying English” helped by the fact that “their parents give them a lot of support, but without applying excess pressure at home”. This kind of influence was not a major factor for most older learners; they tended to be self-motivated to learn but experienced greater difficulty in fitting their language learning into their busy work and home lives (e.g. Teacher I’s B1 adult group).

From the learner questionnaires, it was shown that decisions to enrol at the BIF, and for some to take examinations, varied from very instrumental reasons, including current/future work and study purposes, to personal interest or the influence of other people in their lives (e.g. “stakeholders” such as teachers and parents). For some of the younger learners in particular, parental pressure to improve their English and gain useful qualifications was a key factor. In Teacher D’s FCE Group (noted above) the predominant reasons for studying at the BIF was the influence of parents, including ambitions for their children to get ahead in English and in some cases, the “snob value” of attending the BIF.

The use of the LLQ provided information about the attitudes and motivation of the learners as well as other socio-psychological traits. However, the usefulness of the LLQ was limited with only 60 questionnaires completed by the learners at the beginning of their courses, thus making statistical analyses and generalisations problematical. Some usable analyses were, however, derived from those who did complete it and the brief discussion in Appendix 6.8 illustrates how the instrument was intended to be used. These data shed some light on the nature of the BIF learners, although care should naturally be taken in interpreting the analysis given the small sample.

It was the teachers who were reluctant to administer the LLQ and other questionnaires, perhaps being mindful that their students (as paying customers) should not be asked to spend too much time on non-learning activities in an impact study. Conversely, the learners themselves enjoyed (and probably benefited from) some impact study activities (e.g. the little oral “tests” and having a guest in the classroom whom the teacher might include in communicative activities) and there
was almost certainly a Hawthorne effect for some (i.e. a change of behaviour in response to a change in the environmental setting).

Hawkey concluded his interim report stating that the FLLGP had encountered “many of the difficulties that classroom-based language learning studies typically have to deal with”, and that such studies are “at most quasi-experimental”. He points out that this is because the researchers are rarely in a position, for practical and ethical reasons, to randomise groups, controlling for variables such as learning background, motivation, attitudes etc. or to assign subjects to different treatments systematically through different teaching programmes, materials etc. Even though the FLLGP had a definite “before, during, and after” design, and there was active participation and support of the management and teachers of BIF, the possibility of controlling the variables was limited.

While these issues proved difficult in some respects, the insights gained were also very informative and useful in understanding the nature of classroom-based research within the impact model. Uncontrolled variability is problematical in experimental or quasi-experimental designs but in qualitative research complex data of this kind is often anticipated and may actually be required in order to understand the phenomenon under investigation.

The video recordings and interviews were particularly successful and the limited analysis which was conducted by the Project Coordinator invited further research using the data. However, the reason why no final report was produced appears to have been the practical difficulty experienced in working with such large amounts of data within the intended timeframe and resources which were available to the FLLGP (i.e. an operational impact project, still part of a developing corporate validation policy, which needed to fit the institutional priorities at the time). Perhaps it should have been anticipated that more resources would have been required during the analysis and reporting phases. It is known from the literature on qualitative data analysis (e.g. Miles and Huberman, 1994) that “thick descriptions” generate large quantities of primary data of this kind and that making use of such data provides a challenge for researchers. Our developing model takes account of this factor.

One of the problems in dealing with classroom video is that it needs to be stored and indexed effectively so that the retrieval of critical incidents can be systematic and matched up with other data (notes, interviews etc.), or cross referenced to quantitative data scores on tests or responses to questionnaires. Hawkey and
others in the project team recognised that further analysis of the FLLGP video data would be difficult unless through the up-to-date IT solutions of a searchable video database, a positive outcome of the FLLGP, briefly discussed below. It may also be noted that a PhD project supported by Cambridge ESOL is at the time of writing being carried out by Michael Perrone at Columbia Teachers College, New York (under the supervision of Purpura), in part using video data from this database.

After having reviewed and summarised the FLLGP as described above, reflections and insights were collected through conversations with Hawkey, the project coordinator. These conversational data shed light “on the narrative” and provided a unique perspective on how and why the project developed in the way it did. They also inform our developing model.

Two aspects of the project were focused on in conversations and e-mail exchanges with the current author during the meta-analysis of the FLLGP (2007-8) reported in this chapter:

i) reasons why some data were not collected as anticipated in the original design and action plan;

ii) the final outcomes and the lessons to be learned from what happened.

With regard to point (i), Hawkey confirms that, in his view, all three impact case studies could be said to have had a “quasi-experimental design”, in the sense that they were designed to have “a pre-, mid- and post-data collection structure”, with the potential for test scores to be collected to assess the target language and other developmental progress. But he was clear in the case of the FLLGP that the difficulties in collecting score and questionnaire data were not signs of failure: “I don’t see FLLGP as a quasi-experimental design that went wrong”. Given that FLLGP was an impact-oriented project sponsored by Cambridge ESOL, Hawkey had been aware from the start that “it was not primarily an academic research study”; rather it was part of “a systematic approach to getting broader and deeper data” in order to be better informed about preparation courses for Cambridge exams and other tests, and the related effects on the stakeholders.

“Ultimately the aim was to draw some better informed inferences based on qualitative and quantitative data about the washback/impact of the examinations as a key element of test validation. In order to achieve this, we knew we needed to probe learning, teaching, individual differences, motivations and attitudes – in other words as many as possible of the “intervening” variables which we knew enriched the
increasingly comprehensive and sensitive impact model which was replacing the simpler washback model.” (Hawkey, 2008, personal communication).

Hawkey expected that this approach would require “a delicate professional relationship” in working with Cambridge ESOL’s stakeholders in their own context, and that: “We were lucky to have the privilege of getting inside working institutions, classroom videoing and interviewing, administering instruments outside the curriculum. ... I was aware that we were there on trust and could be asked to leave at any moment”.

This meant that adaptability to the setting was going to be a key feature of the approach: “I always expected that we would need to modify the FLLGP ‘shoe’ to fit the BIF ‘foot’”. Hawkey emphasises the crucial importance of the interactions with key stakeholders in achieving this fit; “without their active collaboration the project would not have been possible”. The BIF DOS was the most important local “champion”, having committed to the aims and objective during the early planning phase. Hawkey notes that the DOS was “an absolutely key stakeholder figure in our working research model”; she engaged in a constant two-way exchange of information with him as the Coordinator, picking up and communicating “the vibes” about the development of the project within the school, e.g. in relation to an outsider observing processes in classrooms, and keeping the Director of the Institute informed and supportive after the DOS and Hawkey had briefed her to seek approval at the start of the project.

In discussing the aims, plans, and processes the DOS played a role in their development and modification and provided practical help for Hawkey during (e.g. in videoing) and after his school visits. The FLLGP thus provided an important lesson for Cambridge ESOL, feeding into the developing model for impact-related research sponsored by a public examining body.

Hawkey concludes that the study would probably have been impossible to follow through if he and the DOS had not had such a good relationship. In particular their collaboration led to an appropriate level of care and sensitivity in the data collection. This was supported by other stakeholders in the local setting; for example, Hawkey singles out the local support from ESOL’s in-country Manager, who played a role in planning, negotiating and participating in the research (e.g. as scene-setter, video cameraman, and local problem-solver).

Because this was a “working project” there was also “an essential PR aspect which Cambridge ESOL’s local staff could not ignore”. On the one hand, they were able to
exploit the project in their communications about Cambridge interests within the
Italian context. On the other, there was a need to ensure that the research method
itself did not lead to negative consequences (bad public relations for the
examinations board). Certain ground rules had already been established in order to
avoid this for the collection of data in school visits during the PLIS. For example, the
following guidelines were followed:

- the principle that the “customer is always right” (you cannot insist on
doing things in the school/classroom which are not acceptable to the
local participants).
- classes would only be observed and videoed if everyone concerned was
happy with it and consented (one teacher was never videoed, but she
was observed, and interviewed; another teacher was neither observed
nor videoed, but she was interviewed).
- The teacher logs were pre-structured but remained very open-ended and
were never demanded or required in particular formats (in Hawkey’s
view these “were truly revealing” and the resulting portraits often covered
aspects of beliefs, attitudes, and motivations that would otherwise have
gone undetected though they are vital to ESOL’s need for rich impact
data).
- data collection was entirely at the discretion of the teachers and DOS
(the Coordinator never said “you must administer this”).

With regard to this last point, Hawkey notes that the Language Learning
Questionnaires were not wholly accepted by some of the teachers (especially when
they compared notes with colleagues) and it was because of this that these data
were not collected as planned. It seems that some teachers questioned what they
believed were “outdated” constructs, (e.g. in the attitude and motivation items) in the
questionnaires. Thus they did not want to administer the LLQ (which was their
prerogative), which led to missing data. Hawkey observes that this was an
“interesting finding” rather than error on the part of the research team. There was
similar reluctance among some teachers to administer the FLLGP post-test (as
discussed above).

In relation to the final outcomes and the lessons to be learned from what happened
(point (ii), Hawkey comments on the fact that no final FLLGP report was produced
and considers it “an issue worth exploring”. He suggests two possible reasons. First
there was a change in focus in his own contribution at the end of the data collection
phase; he felt that through the video database project “the essential subtler side of
the intervening variables could be pursued”. That “new” project led to a full write up and a publication which had not been envisaged at the start. Secondly, he confirms (see just above) that his action plan had not predicted the amount of support and additional resources that would be needed to exploit the data fully.

6.4.2 Summary and conclusions

In summary, the FLLGP generally went ahead as anticipated in the action plan but with mixed results. The BIF provided a case study comprising 139 learners and nine teachers following a curriculum based on CEFR target levels, and with an external examination for about half the learners (i.e. PET and FCE targeting the B1 and B2 levels). The use of the CEFR for establishing the class levels and the possibility of external certification were reinforced by the prevailing milieu in the macro context, including PL2000 initiatives and the wider recognition of the Cambridge examinations to certificate outcomes related to the CEFR levels. These contextual factors, and the prior experience with the PL2000 research, determined the kinds of impact that the research team expected to observe in developing the impact hypotheses.

The expectation that the action-oriented approach adopted by the CEFR would lead to teachers introducing communicative activities into the classroom was well founded. Four skills were addressed with an emphasis on speaking and listening practice in class, and homework and activities outside of the classroom being recommended for practicing other skills. The prevalence of this approach was confirmed in the questionnaire data, teacher logs and observations made in class (e.g. see relevant appendices to this chapter).

For the examination-oriented classes, greater attention to the requirements of the examination was predicted, at least in the latter part of the year leading up to the examination period (May/June). In the examination preparation classes, the potential for higher motivation and more targeted learning was to be considered against the possibility of a narrowing of the curriculum (perhaps with time spent on unproductive test-taking techniques) and a possible increase in stress and anxiety.

The teachers were very aware of the requirements of the different examinations and were experienced in using course materials based on the Cambridge ESOL examinations. Their comments also revealed ways in which they adapted their teaching to meet the particular test-taking needs of their learners. All learners in Teacher E’s young learner group shared a main priority in “wanting to pass FCE”; the teacher commented that his aim was “to give this group of learners the
opportunity to practise and improve their language within the context of exam preparation”. The course syllabus for the class was based on the FCE course book which he thought to be “rather dry and structured in approach”. To compensate he “moved around” the book “in order to introduce an element of variety to a somewhat monotonous diet”. He was happy with the results of this: he noted that the learners “were used to working in pairs and small groups and seemed to accept this type of interaction pattern”; in his view this led to an atmosphere that was positive and conducive to learning.

Teacher B in his first group (BIF Class FC5B - an adult FCE group) referred explicitly to the “ruthless nature of following the course book and correcting homework and dealing with the variety of exam skills”. This had meant “a reduced amount of time” being spent on speaking. However, having observed this, he tried to “redress the balance”, and having been very satisfied with mock results in other skills, he determined “to work in a systematic way on Paper 5 speaking skills in class”. He also noted that listening posed particular problems for some learners, especially “in dealing with the rate of quick delivery, colloquial listening texts tested in FCE past papers” and determined to work on this area.

Teacher B’s second group was BIF Group E3G (an adult B1/PET group). Only some of this group were intending to take PET and there was no pressure to do so on the others. He noted that learners in this group did not have immediate needs for a Cambridge qualification (and that the “teacher doesn’t push PET”). He referred to this class as being “like a social club” …. “non-academic, not exam-motivated - they decide whether they take the exam or not”. There was also the teacher’s belief, shared by some of the learners that the qualification of value was FCE (B2). PET (B1) did not offer “tangible benefits” i.e. it was worth waiting to take FCE in future.

The example of Teacher B’s two classes shows that the presence of particular examinations (especially FCE which had greater currency and recognition as a qualification) did affect the behaviour and attitudes of both the learners and the teacher. However, FLLGP data did not clarify how systematically this occurred in the school as both the examination classes and the non-examination classes were using materials and methods which were similar, suggesting that variables other than examinations provided motivations and impacted on attitudes.

Sub-groups within the classes (e.g. of 2-3 learners working together) and the overall group dynamics were noted as being significant by several teachers. These aspects
of “social context” in the classroom influenced opportunities to practice particular language skills or elements and affected group ethos.

In Teacher B’s first group (the adult FCE group), he noted “good intra group relations” and that the class was a “homogeneous group”. Some of the learners had been together in Teacher B’s PET class from the previous year, and the newcomers mainly conformed to the “group mould”, characterised as “a studious and book bound approach to learning”. Despite the homogeneity of the group there were significant differences in sub-group activities; for example the “huge success” of his strategy to get the learners to practise speaking by introducing them to a native speaker of English who was studying Italian at the BIF. A sub-group of five learners took part in “out of class conversations exchanges” (1 hour in Italian and 1 hour in English). Teacher B observed that these learners got a “real psychological lift” and that “their ability and skills and confidence to cope with a conversation with a person they don’t know” had increased. One of the other members of the group was “too shy” to accept this offer and her lack of confidence was a “cause of concern” for the teacher.

The in-depth qualitative approach adopted in the FLLGP proved revealing on classroom group dynamics, a key, but under-researched issue in the era of communicative and task-based language teaching. Following type of evidence from the teacher profiles, amenable to further pursuit in the video database, of course, has the potential to inform, to a depth often missed in washback research (see Spratt, 2005).

Teacher C and Teacher E both taught two groups. Teacher C commented that in her second group, (an adult group planning to take a BIF examination), there was a “very good class dynamic” in particular, she noted that “they enjoy the communicative approach to language teaching”. Teacher E noted that in his adult group taking the BIF examination (E4D) certain sub-groups formed, and he was particularly aware of, for example, a sub-group of four learners regularly borrowed graded readers from the BIF library, and that two other students tended to dominate the class at times and especially to “monopolize student-to-teacher interaction”.

This second dynamic also manifested itself in antagonism between the two learners themselves which “occasionally raised the temperature in the class”. Managing these group dynamics, Teacher E allowed the students to choose where to sit and to determine who they worked with during pair and group work. He noted that this was “conducive to a more relaxed atmosphere” an “more likely to generate language”. In
general, Teacher E believed that learning was taking place in a relaxed and enjoyable environment in his classroom and he was very happy with the progress his students were making. He also commented positively on how this also affected him and his own motivation as a teacher: “This helps to build my confidence as a teacher which in turn-motivates me”. Revealing indeed.

Overall the FLLGP did not reach a conclusion on the important question of whether the learners learned more or better because of the specific examination course they followed, perhaps partly due to the ending of the project before all data had been fully exploited, (as noted above). However, the project did highlight a culture of “examination preparation” within the school with the use of examinations aligned to the CEFR evident. Yet this was not allowed to become overly dominant and the teachers were at pains to ensure that their primary focus was on the learning content. If they perceived the examination becoming too dominant they sought ways to compensate (as we saw for Teacher B).

From the perspective of the impact research design and methodology, although the quasi-experimental design was not, as noted above, pursued, the questionnaire data proved difficult to capture, the qualitative approach proved to be popular with many participants and very revealing. Certainly, some of the features of the previous impact projects were successfully reapplied in the FLLGP and provide insights for the impact model, especially the focus on teacher and learner variables emerging from the qualitative research. There was notable success in setting up the study in “real time” within the limited window of opportunity which presented itself after the main PL2000 innovations; the need to act rapidly or to miss an opportunity to gain useful insights may turn out to be an important feature of the impact model. Yet, rapid response presents a dilemma for the educational researcher; time constraints can mean that extensive planning and piloting are not feasible and this can limit the possibilities for controlling variables. This leaves the researcher with the choice of using “opportunity samples” or not proceeding with the study (see Hawkey, 1982, chapter 2 for discussion of this point). If the researcher does proceed with a project under these circumstances, the research itself can be undermined if “standard procedures” and validated instrument templates are not available to be used (or to be adapted) within the resources which are available. The availability of the instruments and procedures from the earlier studies enabled the FLLGP project to proceed within the required timeframe.

A further positive outcome was the decision taken by Cambridge ESOL to organise the video data into an interactive database (as noted above). It was recognised that
an interactive database would facilitate the retrieval and analysis of data in response to particular research questions. Project Coordinator working with other staff from Cambridge ESOL documented this development as an outcome of the three Cambridge ESOL impact studies and submitted it for publication; the aims, principles and approaches involved in the development of the impact research video database were described, including the software selection, the main design features with sample screenshots, and a discussion of envisaged uses of the database in the future (Hawkey et al, 2007).

6.5 Analysis of the FLLGP case

In this section we look in more detail at the three main variables: the context, the learners and the teachers. Recent constructivist approaches to classroom-based learning will be considered and the implications for the impact model discussed.

6.5.1 Modelling the classroom context

The BIF context was summarised above as an example of a hierarchical structure with the micro contexts nested within the wider educational milieu. This way of thinking about ELT practice dates back to the 1970s with models proposed by Campbell, Mackey, Spolsky, Strevens and others (see Stern, 1983; chapter 3). Mackey’s interaction model, for example, shows the curriculum as deriving from governmental policy on education and language, which then interacts with other variables within schools and within society to determine “what the learner gets” (i.e. instruction). Such models are also found in more recent educational research; for example, Cohen, Raudenbush and Ball (2003: 124) use a figure of concentric circles to represent the instructional dynamic. For them, this constitutes the core of educational processes and the dynamic interactions among teachers, students (learners), content and environments (contexts) are at the heart of educational research (see Ball and Forzani, 2007: 538).

With this in mind, the classroom context can be modelled to enable the impact of assessment to be explored in more depth. Figure 6.2 is used as a starting point for the model, which will be built up with reference to the FLLGP case and to the analysis discussed above.

The BIF classes differed from state school classrooms because of class size/composition and other features (high student motivation, fee payment and, perhaps, the L1 and training of the teachers). However, aspects of the classes and the management of the lessons observed in the BIF still represent typical classroom learning environments with archetypical features. These features include:
• a classroom with desks and various audio visual teaching aids;
• a class teacher present in the classroom (all or most of the time);
• students in a (fairly) fixed group, varying from small groups of 10-12, as in the BiF, to large classes with 20-30+ students as in state schools;
• fixed length lessons (e.g. 1 hour at a time) taking place several times a week for a predetermined period of time (e.g. a course over an academic year from October to May);
• a curriculum plan and teaching syllabus based on external reference points (national guidelines, the CEFR etc);
• a range of course materials (e.g. text books, tapes, videos etc.);
• opportunities to learn outside of the class including spontaneous activities and unplanned practice as well as homework set by the teacher;
• assessments of various kinds, including tests/quizzes carried out by the teacher and external examinations for some or all of the learners, e.g. the PET and FCE examinations for the BiF learners who opted for external certification;
• progress reports and final evaluations supplied by the school to the learners and their sponsors (parents, employers etc.).

[Note: Recent developments in on-line, distance and blended learning have begun to change these features. The role of the teacher and the physical contexts are different in “virtual learning environments” (VLEs): although not be covered in this discussion, such developments may interact with assessment practices to lead to alternative methods of testing language learning. This will be touched on below.]

a). The milieu – the wider educational context
The milieu in the FLLGP case was provided by the post-PL2000 educational context in Italy (see Appendices 5.1 to 5.3). This is represented by the outer box of Figure 6.2 – shown in red.

In this setting, the CEFR and its level system had become widely adopted and relatively well-understood within educational circles, largely because of the government policies. Many stakeholders, including parents and learners, had had opportunities to find out about it through the PL2000 and the “value added” features of external examinations had been recognised. This was demonstrated by the fact that the test takers or their parents were prepared to contribute to the cost of taking those examinations in state schools.
Why teachers and learners choose a particular examination, rather than an alternative, is an important impact consideration. It is also an example of the dynamic interaction between the school context and the milieu. In state schools, assessment procedures are often mandated by the educational authorities outside the school, as for national tests such as SATS in the UK. But even when there is no obligation to use particular tests or examinations, the choice may be influenced by prevailing national/regional policies (as in the PL2000), or by other considerations. Such considerations include the currency and recognition of the results, the reputation or prestige of the awarding body, or by promotional activities of examinations board and commercial providers of tests (e.g. “test publishers”). This means that the tests or examinations which are used in a school/classroom setting are the result of choices made by the participants within the school (i.e. teachers or school managers) based on influences from outside (e.g. parents of younger learners). An examination culture was evident in the BIF, but as evidenced above, not to the extent that it dominated other priorities, such as an adherence to communicative language teaching principles.

The main external examinations in the BIF case were the Cambridge FCE and PET; this type of external examination is shown in our model (Figure 6.2) as exerting an influence from the milieu and as having potential effects within the school generally and in specific classrooms. The results obtained by the test-taking groups have implications for them as learners within the school (e.g. to provide evidence of progress towards desired learning goals) and also as “qualifications” in the wider milieu (e.g. for the work, study and other purposes which the learners stated in their questionnaires).

The local geographical context was an important aspect of the milieu; in the case of the BIF, this was the tourist city of Florence and the well-to-do region of Tuscany where English is used widely with tourists and visitors. Not only did this influence stakeholders’ attitudes and choices towards English language education in their local area, for diligent learners there were numerous “opportunities to learn” outside the regular classroom, (as recorded in some teacher logs).
b). The school context
The school context was the British Institute, its prestige, historical links to the Cambridge University “brand”, library, competent and proactive management team (see above) all constituting the culture of the BIF. We also saw evidence of “a BIF community”, including family links with the school (e.g. cases of both parents and children enrolling for courses) and a tendency to re-enrol from year-to-year in order to progress through the levels.

The communicative approach to teaching/learning and the principles of the CEFR were supported by the beliefs and attitudes of the teachers, captured in a range of documents and observed in classrooms. The teachers and many learners appeared satisfied to choose Cambridge examinations as their external “exam route”.

The BIF attracts learners who can pay to learn English and who recognise that the BIF is at the quality end of the ELT market (and so is relatively expensive). The attitudes and expectations of the learners and their sponsors are influenced by BIF’s “brand values”, including historical trends in pass rates and success in getting learners to reach their intended objectives. The importance of external qualifications for motivating students and in providing independent verification of the quality of education offered by the school is recognised by the management and teachers.

c). The language classroom context
The inner box of the model (the lowest level in the hierarchy), represents the classes and groupings of individual learners within them. The BIF teachers recorded the friendships, rivalries and other relationships which developed in their classes and exploited the sub-groups for extending language practice outside of the classroom.

The BIF classes were small and learners were grouped on the basis of placement procedures which were rigorous and well-managed, taking into account a range of considerations. Language level and prior experience in the school (e.g. previous courses completed) were important factors in making up the classes at the beginning of the year.

Figure 6.2 does not represent the classroom dynamics exemplified by the BIF class groups and needs to be extended to facilitate our analysis. One aspect of an extended analysis would be focus on the use of language in the classroom and the similarities and differences between classroom uses of language and language in “real world” domains. In the SLA literature there is extensive treatment of this issue,
but our concern here is with the construct of CLA in a classroom context of learning (sometimes referred to as “instructed second language learning”, e.g. Gass and Selinker, 2008: 368-394), from the perspective of a test developer.

The level of the teacher’s own proficiency in English (as L2) can be an important variable but this was not relevant in the BIF case, as the teachers were all native speakers.

Figure 6.5 (from Jones, 2007) represents the use of language in “naturalistic settings” (outside the classroom): here the L2 learner/user is shown interacting with another speaker to complete a communicative task. Central to this is the communicative language construct, (for our purposes based on the CEFR), and as we saw in Chapter 2, the potential for washback from language assessments exists when the focal construct is shared between the curriculum goals and the examination. To be convergent with learning goals derived from the CEFR (as in the BIF case), the examination also needs to be consistent with the CEFR’s action-oriented approach. (See Hawkey, 2004).

Figure 6.5  Language activities and language learning in naturalistic settings

In Figure 6.5 the L2 learner engages in language activities to meet immediate communicative needs: a “genuine” communicative task takes place in a social context where the language user has to deal with particular topics/themes using the target language. This “naturalistic” use can be contrasted with the structured approach to teaching and learning which occurs in classrooms. The communicative approach attempts to replicate aspects of target use communication within the “scaffolded” environment of a classroom by providing action-oriented curricula. The
PL2000 study illustrated how this was attempted at a national level in the Italian context.

The “resources” activated by engaging with a communicative task include: knowledge (language and topical knowledge), communication strategies, and other physical and psychological traits of the person. Any of these resources can facilitate or hinder the endeavour. Use of language may be hindered by psychological traits, such as extreme shyness, or a physical trait, such as a lisp, leading to difficulties in communicating effectively. The focus on social and cognitive factors in combination with the personal attributes of individuals, draws attention to nature of successful performance in such communicative tasks. In other words, it reminds us that language knowledge is a necessary, but not a sufficient condition for successful communication.

A communicative task requires the learner to activate language knowledge and other cognitive processes to complete the task successfully within a social context. In the role of test taker, the learner has to engage with the tasks in the test to produce responses to be assessed. A task is considered (relatively) authentic if the testing context and learner’s response adequately reflect genuine communication in the target use domain. In other words, if the test activates the candidate’s language knowledge and social capacities in ways which are similar to those needed to communicate in the target use domain it achieves situational and interactional authenticity (see Bachman, 1990 for a discussion of this point).

This approach, which has both social and cognitive dimensions, represents an “interactionist” position on learning. Within an interactionist approach, the construct is represented by three components:

- a theory of language knowledge and related cognitive processes (including linguistic and psycholinguistics features);
- the language tasks to be performed and their related contextual features;
- the communicative behaviours of the learner in engaging with the tasks.

Weir’s (2005) socio-cognitive model proposes: theory-based, context and response validity (see Chapter 1). In the role of the test taker, the learner engages with test tasks to produce the responses to be assessed. The task is (relatively) authentic if the testing context and learner’s responses adequately reflect genuine communication in the target use domain. In other words, if the test “activates” language knowledge and social capacities similar to those needed to communicate.
in the real world it achieves a degree of *situational and interactional authenticity* (see Bachman, 1990). As part of our expanded impact model, the test provider needs to consider construct definitions which take into account the ability to use language in communicative events. Test developers who fail to build in adequate consideration of situational and interactional authenticity in their construct definitions may find it more difficult to achieve positive washback in the ways described by Green (2007, chapter 1).

To accommodate this we can extend Figure 6.5 to include the *classroom context of language use*, as opposed to the “unscaffolded” real world environment. See Figure 6.6

**Figure 6.6** The classroom context of *language use*.

---

Figure 6.5 represents the micro context as observed in BIF classrooms. Unlike naturalistic learning contexts, the classroom and the role of the teacher contribute *specifically planned effects on learning*; the interaction between teachers and learners, and the interactions between the learners themselves (in pairs or in groups) create “opportunities to learn” by using the target language within pedagogical tasks.

d). **Conditions for language learning in the classroom context**

Figure 6.6 shows the learner engaging in language activities with other learners in the classroom environment. The teacher here is responsible for structuring the teaching programme according to the syllabus, for creating the learning environment, and for setting up specific tasks and activities to focus attention of the learners on...
particular points to be learnt. In other words the classroom should create appropriate *conditions for language learning*.

The physical conditions, constraints, texts, themes, domains and tasks are parameters which are selected and manipulated to provide a supportive context. School managers and class teachers play an important role in implementing curriculum guidelines based on external rules and regulations, such as those provided by Ministries or contained in the specifications and support material for external examinations, e.g. FCE or PET (cf. the dynamic interaction with the external environments noted above). It seems that the ways in which these processes work in practice are both complex and variable, and there tends to be a combination of external influences and internal freedoms/preferences within the school and individual classrooms. This was confirmed in the FLLGP. It had also been a feature of the PL2000 where the CEFR provided the basis for syllabus design and the Ministry guidelines set the general parameters for implementing the lessons and the external assessments: within the schools themselves there was freedom to devise the courses, choose the materials and the tests and to implement the teaching methodology. Individual teachers also brought their own understandings and preferences into their classes.

Another feature of the classroom as a social setting is the importance of relationships which are established *between* the participants - between the teacher and the learners, and between the learners themselves. There is a growing body of literature on group dynamics which suggests that, not only the teacher, but also classmates and peers affect the learning of others and can be influential in numerous ways. Generally positive relationships in the classroom, as well as “specific dynamics” within a learner group, (e.g. characterised by relationships of interdependence) have been shown to be important in engaging learners, (Dörnyei, 1997). Apparently similar classes manifest different group dynamics and the ways in which assessment is carried out in particular groups can be influenced by views, attitudes, motivations and willingness to put in effort. We saw examples of this in the BIF.

e). Language proficiency in the classroom context

A major determinant of learning progression in classrooms is the extent to which the elements of knowledge in the syllabus are mastered successfully by the learners; in the BIF case we referred to this as the *proficiency dimension*, referenced to levels of the CEFR.
In a communicative approach, teachers place more or less emphasis on certain features of the language and language use. In general, however, the language tasks which the teacher chooses have a basis in the specified course content and the typical learning unit link communicative activities to language forms, lexis and so on. But, as the FLLGP classroom recordings remind us, other important things occur in the classrooms, both planned and spontaneous, and teachers help learners to develop language learning strategies and provide assistance in the affective domain. Teachers are also influential in persuading their learners to put in more effort, with reference to achievement goals, such as success in public examinations. The FLLGP research design recognised these points and collected data to shed light on how these processes work.

f). The role of assessment within the classroom

In addition to structuring the input, the teacher plays an important role in monitoring the output and evaluating progress against the objectives; this might involve carrying out informal assessment and in providing feedback to the learners to help them in the next stage of learning (see recent literature on SLA and input: e.g. Swain and Suzuki, 2008). The type and amount of feedback the learner receives is an important impact of assessment on learners. The extent to which assessment is designed to be formative in its uses and to be integrated into the teaching programme needs to be considered as an aspect of the impact model.

In a useful volume on language assessment for teachers, Stoynoff and Chapelle (2005: 11) present a continuum of test uses: at one extreme they place low-stakes uses including diagnostic and achievement tests in class; in the middle they put placement uses; and at the other extreme they place high-stakes uses, including tests for admission and certification. But for some commentators (e.g. Oats, 2006), the stakes associated with tests cannot be determined by the intended uses because the importance of tests for individuals cannot be determined in advance. Apparently low-stakes may have important personal impacts for individual learners (see also Hughes 1993). Similarly, Madaus (1988) has argued that it is the perception of the stakes which influence test taker’s behaviour.

However, if we accept Stoynoff and Chapelle’s analysis, low-stakes tests tend to be more informal and internal to the classroom, whereas the high-stakes tests tend to be more formal, originating outside of the classroom and with currency in the environment outside the school. This remains a helpful distinction; in the BIF case, a range of assessment procedures of both kinds were available, and to some extent,
the teachers and the learners themselves could choose whether to opt for the higher stakes (external) option.

Jones’ classroom model (Figure 6.6) allows us to consider the ways in which the different types of assessment used in the classroom can impact on the participants. The model can be extended again to represent informal and formal assessment within the classroom context; Jones does not refer to the “stakes” but points out two ways in which language assessment can be carried out with different focuses. One is a predominantly formative focus, and the other is a predominantly summative focus, (the latter typically represented by the formal, external examinations of the higher stakes kind noted by Stoyanoff and Chapelle). See Figure 6.7

For Jones, both approaches share a central concern with the learner’s capacity to engage in language activities, and it is conceivable that the same test could be used in both ways, depending on how the test results are presented and interpreted. In general, however, some test designs are more suited to formative uses than others because of their link to curriculum, and the way in which test performance is reported and used by teachers and learner. Educational assessments which are formative in focus are fundamentally “connected to learning” (Goldstein, 1989), and can be contrasted to assessment which is deliberately unconnected to particular learning environments. (Assessments which are “separated from learning” include psychometric measures such as. IQ or personality tests.)

In Britain, the Assessment Reform Group, (authors of the pamphlet entitled “Assessment for Learning: Beyond the Black Box”, 2002), have emphasised the need to connect assessment to learning more effectively. They suggest that a clear
distinction should be made between assessment of learning, for the purposes of grading and reporting, and assessment for learning “which calls for different priorities, new procedures and a new commitment” (2002:2). This echoes the summative versus formative focus of assessment noted above.

The Assessment for Learning (AfL) movement has taken on a political dimension, reflecting a reaction against national testing for accountability since the 1980s, and the essentially summative focus of national school examinations in England. Proponents believe that too much attention has been given to comparing children, teachers and schools using examination results (e.g. using league tables) and not enough to assessing children’s individual achievements.

The discussion of formative assessment (particularly when understood as AfL) is relevant here because of the focus on the dynamics of interaction between teacher and learners in the classroom, on teaching processes such as goal setting and feedback, and above all on building and maintaining learner motivation. In a language classroom, formatively-orientated assessment might focus on specific language points and the development of strategies for learning or for dealing with the communicative language tasks (shown on the left-hand side of Figure 6.6). When the assessment is designed to target the achievement of goals and learning objectives within a specific pedagogic context, the focus can be considered formative. The results of the assessment will help to structure and focus the immediate next steps for the learner in improving his/her language ability. Traditionally, if this kind of assessment has been carried out at all, it has been teacher-based and has employed less formal methods. In the USA, the concept of “dynamic assessment” (similar to assessment for learning in the UK) has been proposed as a way of integrating assessment and instruction with the goal of enhancing language learning development (Lantolf and Poehner, 2004, 2005).

When the assessment focuses on the overall outcomes of learning, i.e. the knowledge and abilities acquired and available for use outside of the classroom, the focus tends to be a summative one. In this case; the results are often intended for secondary users rather than the learner and his/her teacher, and may be used for making decisions outside the classroom or school (e.g. for admission to higher education or employment). External examinations, such as those in the PL2000 offered by Cambridge ESOL, are regularly used for these purposes and as they are more rigorously standardised than local tests, the results are deemed to be more dependable and generalisable outside of the school context. The stakes attached to them by society also tend to be higher. This was one of the explicit arguments
employed by the Italian Ministry in proposing the use of the “enti certificatori” within the PL2000 (i.e. for standardisation and accountability). We also noted above the importance of the image and prestige associated with the Cambridge examinations in the BIF context.

In our discussion of the impact model, we have asserted that there is an expectation that any test or examination will have the potential to affect the teaching/learning context – whether intended or not. In light of the discussion above, we can also conclude that there is likely to be a tension between the role of tests or examinations as independent and generalisable measures, and the need for them to be helpful and sensitive to individual learning contexts with which they inevitably interact (or in some cases are integrated within). In recent years, there has indeed been increased recognition in the assessment literature of this tension; on the one hand there has been a call for tests to provide more specific and individualised information about learners and on the other there has been a growing reliance on standardised examinations for high-stakes decision making purposes.

The dynamic interaction between schools and the wider socio-political environment has seen examination results being used increasingly for political purposes in many countries; for example, this is the case when learners’ results are used to compare, not just the learners themselves, but also the performance of their schools and teachers. This tendency as we have noted above has attracted criticism in recent years from the Assessment Reform Group and others who claim that the use of league tables and other forms of “system evaluation” based on high-stakes examinations can lead to negative impacts on teaching/learning and can distort the relationship between the testing system and what goes on in schools. This is a case of Goodhart’s law (described in Chapter 2) whereby the accountability targets lead to an over-emphasis on test preparation to boost results and a reduced focus on individual learning needs.

The model of impact needs to take account of these societal processes, particularly in the development and use of language examinations which have the potential for accountability uses.

g). Individualising assessment - learner’s needs in context
How and when the assessment takes place can facilitate future learning and/or provide evidence of the current level of achievement for societal purposes. How far the same assessment procedure can fulfil both roles remains an interesting area of
research and is certainly relevant to the conceptualisation of impact. Understanding and meeting the assessment needs of individual learners is crucial in this endeavour.

Research in language teaching focusing on learner needs (Mackey, 1970; Strevens, 1976, 1977) is reflected in the language assessment literature dating back to the work of Carroll (1963). Spolsky, in his 1990 paper on the social aspects of individual assessment, makes the point that most proficiency tests tend to ignore the “dynamic dimension” of learning. Examinations take a “snap shot view” of the learner and there is no way of knowing “whether the subject is in a process of rapid language learning or has long since reached a plateau.” (1990:12). This is because most proficiency tests are designed to be used summatively and to be applied to large numbers of people at the same time.

The assessment for learning approach places greater emphasis on the formative dimensions of assessment (e.g. Black and Wiliam, 1998) and information deriving from classroom procedures and teacher-administered observations provides a richer profiling of the learner’s current level and helps to structure the next stage of learning. However, as we have seen in the BIF case, the ways in which assessments are actually used in classrooms and at the school level depends on both institutional and personal preferences and are influenced by factors which impinge from the wider macro context, including laws and regulations and the currency and recognition attributed to particular tests and examinations, (e.g. for work, study, international mobility, etc.). The tension we have noted between the need for specific, individualised, formatively-oriented assessment and the demand for general qualifications carrying wider recognition and currency within society is likely to persist.

The volume on “inference and generalisability” edited by Chalhoub-Deville, Chapelle and Duff (2006) provides cross-disciplinary perspectives on how researchers conceptualise inferences about learner acquisition and performances, as well as the dependability and generalisability of their research findings. Researchers regularly find themselves grappling with tensions and dilemmas which arise from the importance of taking an interest in individuals and individual differences, and the pressures coming from academia and other areas of society for summation, dependability and generalisation (see also the debate between positivist and hermeneutic approaches discussed in Chapter 3). For Cambridge ESOL, the individualisation of assessment presents a particular challenge: how can the currency and recognition of well-known, international assessments be maintained and at the same time the information about individual candidates’ knowledge and
ability be increased in a practical and affordable way? The more specific the test is to an individual language learner, the less generalisable the score is likely to be, and the more difficult and expensive the testing procedure must be to provide the accuracy of measurement required.

In the language testing literature authoritative writers such as Spolsky and Davies who worry about the appropriate uses and the social consequences of language assessment (of which test impact is an aspect) have been concerned with this dilemma for some time. Spolsky (1990) concludes on the social aspects of individual assessment with the following statement, “from the social and ethical point of view, individualized assessment is particularly important when it makes clear the complexity of second language knowledge and the difficulty of summing it up in any simple measure.” (1990: 13).

As usual, Spolsky reminds us of the importance of interpreting test results with care and of the limitations in our understanding of language ability and how to measure it. Davies (1990, 179) has also pointed out in similar vein that, “the twin features of variability and error are endemic to all attempts to study and measure language learning”. Variability, he adds, is the result of our inability to define the construct precisely enough (“linguistic imprecision”) and the inevitable error in the assessment systems we develop (“measurement failure”).

Hughes (1990: 16-19), in responding to Spolsky in the same volume agrees that “fuller and more individual information ought to increase the fairness to the test taker”. However, he voices a word of caution by suggesting that two conditions should be met if the benefits of individualisation are to be realised: the information provided about the individual should be accurate; the information provided should be made use of by the intended test users in order to justify the additional effort/cost involved. He points out that even in cases where score profiles are provided (e.g. for tests like IELTS), test users often rely on the overall score for decision-making purposes, rather than taking account of the additional information provided in the profile. Hughes also points out that observational, teacher-administered types of testing suggested by Spolsky, require extensive training and continued monitoring if adequate validity and reliability are to be achieved.

One way in which examinations board can address this dilemma is through the adoption of test development and validation processes which include appropriate consideration of impact at the design phase. In this way, anticipated impacts,
including the potential for supporting individuals’ learning needs, can be built into the system.

h). Integration of assessment within the classroom context

In our analysis so far, we have attempted to model the classroom variables to consider ways in which assessment impacts on language use and language learning processes. The relationship between the form of assessment and uses of the results have been shown to be important, and the potential for the integrating assessment within the teaching/learning context depends on these considerations.

The degree of integration can be placed on a continuum (see Jones, 2004); different types of assessment range from “no integration” to “total integration” of learning and assessment, as shown in Table 6.2. In this table all types of assessment are connected to learning to some extent, so that even a language test/examination which is self-standing can have an explicit relationship with the learning (even if it is not integrated within the learning context).

Table 6.2 Degrees of integration of learning and assessment

<table>
<thead>
<tr>
<th>Least integrated</th>
<th>Test or exam type</th>
<th>Features of integration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-standing test</td>
<td>No intended integration between the test and the learning. Assessment take place before or after the course, for example, an entry or placement test or a one off “benchmark” test for comparison with other learner groups, as in a survey.</td>
<td></td>
</tr>
<tr>
<td>Self-standing test with possibility of support materials Not linked to any one learning context or course structure</td>
<td>The test is intended to be used in a wide range of possible contexts in conjunction with a local determined syllabus/methodology. Materials may be available to support teaching/learning e.g. Cambridge ESOL examinations – FCE, etc.- supported by commercially available text books and resources</td>
<td></td>
</tr>
<tr>
<td>External test or examination linked to a specific learning context</td>
<td>The test is specifically linked to particular learning materials as in achievement-oriented tests or examinations. The assessment typically takes place at the end of the course</td>
<td></td>
</tr>
<tr>
<td>Most integrated</td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test plus integrated practice/learning materials</td>
<td>The test and learning content are cross-referenced (i.e. there is extensive construct overlap) and the testing is integrated within the course not just at the end.</td>
<td></td>
</tr>
<tr>
<td>Assessment-driven learning</td>
<td>A learner’s path through learning material is determined on the basis of performance on “formative” tests, possibly with diagnostic features. Learning objectives are adjusted on the basis of the testing.</td>
<td></td>
</tr>
<tr>
<td>Totally integrated assessment</td>
<td>No separate tests. A learner works through calibrated learning materials and is guided and evaluated by a “hidden assessment engine” and the learning progresses. This requires a technological solution such as a “learning management system” and calibrated assessment tasks built into the structure of the course of learning.</td>
<td></td>
</tr>
</tbody>
</table>

Cambridge ESOL examinations are typically used summatively at the end of courses, and so are relatively *unintegrated* into teaching. However, the analysis of the cases discussed in this thesis suggests that the influence of these examinations *does* extend into the classroom in important ways. For example, provision of examination-oriented course books and “wrap around systems” of support for teachers and learners can facilitate formative functions even when the curriculum leads to an external examination.

The extent to which the examinations share a *common frame of reference* with the schools and society allows for a shared understanding of the language constructs and pedagogic principles and for an *indirect link* to be established between the classroom and the examinations, as we have seen in Italian schools where the CEFR is used widely. As technology develops and learners opt for “blended” learning solutions, totally integrated assessment becomes a possibility and the distinction between teaching and assessment may become increasingly blurred. This has significant implications for the ways in which test impact can be conceived and investigated in future.
6.5.2 Modelling individual differences in the language classroom

In light of the discussion in the previous section where we modelled the classroom context in which assessments plays a part, we now turn our attention again to the learners and teachers as individual stakeholders and participants in that context.

a). Learners/test takers as individuals

The FLLGP case showed that apparently similar class groups create very different dynamics as a result of the interactions between the teachers and individual learners in the groups which can partly be attributed to the characteristics of the learners themselves. In our discussion of the learner profiles in the BIF we noted that the teachers were aware of the individual differences of their learners and sought to characterise these in various ways. While an individual learner’s needs can be taken into account and accommodated by teachers, it is more difficult for test designers to do so for the reasons discussed above. This is especially true for international examinations which are not designed to be integrated into highly specified learning contexts, but are used in many different contexts. This leads to the potential tension between specificity and generalisability noted above..

In Jones’ classroom model (Figure 6.6 above), the language learner-cum-test taker is represented as an individual who has certain features in common with other people in the same learning and test taking context: these features include the age range, the gender balance, the profile of prior learning, etc.. Green discusses these characteristics as presage variables (see also Naiman et al, 1978; Hawkey, 1982), in his research into IELTS washback (Green, 2007: 102). Intended test takers for a particular examination often share common goals related to uses of the L2 in the target language domains and may have similar goals related to the use of their test results, e.g. for societal goals in higher education, work, travel, etc. In state school contexts, achievement of the minimum grade for “promotion” to the next level may be the goal shared by many students and so for some learners, “passing the test” will be a more important goal than learning the language for future purposes.

The cohort of individuals sharing common features is considered to be the “target candidature” for an examination. However, the same group of individuals will be differentiated in important ways, as we saw in the BIF classes. Even when learners/candidates are relatively homogeneous as in the PL2000 case, individual learner characteristics will vary considerably. Personal characteristics are important intervening variables in the way learners perform within the language classroom or during a test (see Figure 6.8).
In applied linguistics it has been recognised for many years that individual learner characteristics contribute differentially to a learner’s ability to acquire a second language (e.g. Altman, 1980; Ellis, 1994; Larsen-Freeman and Long 1991; Skehan, 1989, 1998; Spolsky 1989). It has also been recognised that the same characteristics may also influence how they perform in language assessment contexts as test takers (Bachman, 1990; Hawkey, 1982; Kunnan, 1991, 1994, 1995; Purpura, 1999, 2004).

In language testing, the relationship between demographic characteristics and L2 performance on tests is a fundamental consideration in test validation and the availability of demographic data is essential for routine validation studies (e.g. to investigate differential item functioning (DIF) and possible bias). Research studies have also examined test performance with relation to: age (Farhady, 1982; Spurling & Illyin, 1985; Zeidner, 1987); gender (Farhady, 1982; Kunnan, 1990; Ryan & Bachman, 1990; Sunderland, 1995; Zeidner, 1987); cultural background (Briere, 1968; Zeidner, 1986, 1987); and language background (Alderson & Holland, 1981; Elder, 1995; Farhady, 1982; Ginther & Grant, 1997; Kunnan, 1990, 1995; Ryan & Bachman, 1990; Swinton & Powers, 1980).

The relationships between test takers’ background knowledge and their performance on language tests has also been investigated (Alderson & Urquhart, 1985; Clapham,
Hawkey (1982) was one of the first applied linguists to research the socio-psychological and strategic characteristics of test takers. He took his impetus from early work on “learner-centred” pedagogy (Altman and James, eds, 1980; Naiman, et al, 1978; Strevens, 1980) and the observation that the “extra linguistic factors” play an important role in language learning. He found that personality, cognitive style and attitude were all significant factors (1982: 24). His framework (1982: 12) comprised “five factors with relevance for data collection and investigation: programme, presage, process, product and community factors”. More recent studies have examined test performance in relation to: styles (Chapelle, 1988; Hansen & Stansfield, 1984, Stansfield & Hansen, 1983); attitudes toward language learning (Clément & Kruidenier, 1985; Gardner, 1985, 1988; Zeidner & Bensoussan, 1988); motivation Clément & Kruidenier, 1985; Dörnyei, 1990; Gardner, 1985, 1988; Gardner & Lambert, 1972; Kunnan, 1995); anxiety (Bensoussan & Zeidner, 1989); and use of cognitive and metacognitive strategies (Kunnan, 1995; Purpura, 1999). Kunnan (1995) was one of the first language testers to investigate the relationship between test taker characteristics and test performance using structural equation modelling (an alternative to multiple regression for analysing multivariate data sets and especially suitable to handling latent variables of the kind discussed here).

The ways in which learner cope with tasks in communicative situations in real life, in the classroom or during a test can be attributed to these variables and so research and validation in language assessment need to account for these variables more systematically. With this in mind we can return again to Jones’ model of the language classroom.

Routine test validation procedures should enable collection of the demographic data and information about other knowledge factors related to prior learning and experience. However, as we have seen in the FLLGP, it is more difficult to collect data on strategic, cognitive and socio-psychological characteristics. The use of the LLQs within the impact studies sought to address this issue, but the difficulties encountered in the FLLGP demonstrated that it is not easy to operationalise such instruments in real classroom contexts. It was noteworthy that the quantitative aspects of the project were less successfully implemented and less useful than the qualitative aspects and also proved to be particularly unpopular with the teachers. We can conclude from our FLLGP analysis, that there is a need:
i. to develop better instruments for the collection of learner data in classrooms as part of test impact research;

ii. to incorporate recent insights into language learning into our understanding of classroom activities and the ways we use to capture classroom data.

With regard to the point (i), it is clear that more systematic data are required on how complex, learner factors interact with other contextual variables, especially the interplay between learners and teachers, including the extent to which learners’ attitudes and motivations are influenced by the teacher as a result of the assessment practices in the classroom (especially the provision of feedback). In order for questionnaires to work more successfully, they need to be integrated more effectively into the learning-teaching process to reduce the burden on both learners and teachers. While this has proved difficult in the past, technology may offer ways forward, especially in overcoming practical difficulties. Language learning is enjoying increasing support from electronic media (on line or computer-based); opportunities will arise in future for embedding data collection into the learning processes in less obtrusive ways (e.g. as part of an on-line "blended learning" course).

With regard to point (ii) it was observed that the theories underpinning the LLQ need to be updated, taking account of recent research in second language learning and the movement away from a reliance on experimental methods towards the interpretive approaches in educational research (e.g. to combine questionnaires with interviews and other qualitative techniques). Attention has recently been paid to constructivist concepts, such as learner identity and personal constructs, and the learner considered to be an active participant in the process of organising and structuring knowledge. In impact-related research, ways of capturing data to investigate these concepts need to be developed bearing in mind that environmental and cognitive processes both play a role in classroom-based language learning.

The work of Bourdieu (1977) has been extended to education – as noted in Chapter 2. His “participants” having socially acquired predispositions, tendencies, propensities or inclinations chimes with the FLLGP focus on individuals in classrooms. Another important concept in constructivist thinking is that of identity, not a static feature, but shaped by where we live, who we meet, our income, our temperament, age, etc. (see Norton, 1995). Language learning and the context within which it takes place can impact on the identity of the learner (see for example, Young, 2007; 258 on self-identities of participants in social interaction). Personal Construct Theory (PCT) also offers useful ways to consider the changing nature of individuals (Kelly, 1955: Fransella, 1996). As a constructivist theory, it has points in
common with other ‘post-modern’ movements, (e.g. the social constructionists) which draw attention to the ways in which personal identity is “constructed” and “transformed” in a social context. PCT’s use of repertory grid techniques has proven particularly useful in applied research settings by providing visual and semantic ‘maps’ of an individual’s construct system and how it applies to important facets of an individual’s life (e.g., relationships with other group members).

These constructivist approaches offer new directions for investigating impact, especially learner differences at the micro level. As we shall see below, some of these considerations can also be applied to teacher differences.

b). Teachers as individuals
The next section deals with the teacher as the other essential participant in the social setting of the language classroom exemplified by the BIF case. The teacher profiles and logs in the FLLGP revealed how teachers from the same school approached their learners in significantly different ways. The teacher’s ability to perform her role effectively is the result of many individual characteristics in combination, and the understanding of teacher identity is now a major consideration in teacher education (Kennedy, 1996 and 1997).

The previous studies of washback summarised in Chapter 2 identified a number of teacher-related factors which influence the extent to which a test or examination impacts on their classrooms, including teacher beliefs and attitudes. Teachers may hold beliefs about examinations in general or specific beliefs about the assessments which are used in their own school. Examples of where teacher attitudes and motivations towards examinations were found to be important, as well as personality factors, and willingness to put in effort and to innovate in their teaching were noted in the FLLGP data (see above).

Figure 6.9 summarises features which can be used to characterise individual language teachers (based on Figure 6.8).
Teaching knowledge covers both the teaching content and social aspects of the classroom environment (pedagogic and social priorities), and it has been shown that "opportunities for learner engagement" depend on what the teacher does and says in the classroom, e.g. how classroom discourse is used, see Walsh, 2002.

Bailey, Burkett and Freeman (2008: 618) view the classroom as a context in which "social processes, the content of lessons and the dynamics of power and relationships are framed by and indeed depend on language". A key feature is the teacher's ability to use the target L2 effectively to facilitate classroom processes and to exemplify the target language. Limitations in L2 comprehension may also lead the teacher to misunderstand learning objectives and examination requirements, or to narrow coverage of language skills to fit within his/her own "comfort zone".

If examinations are to be used as intended by the examination provider, a determining factor will be the teacher's ability to understand and deliver the intended content. Communication between examination providers and teachers is, therefore, an important factor in achieving the intended impact. Failure to communicate the curricular goals and the constructs of the assessment or failure on the part of the teacher to understand the information provided, may lead to unsatisfactory or unintended outcomes in the teaching/learning. The implication is that washback and impact are jointly constructed between teachers and examination providers, and if assessment is to be used as a “lever of change”, the complexities of educational innovation need to be taken into account. Examination providers also have a
responsibility to monitor whether communication is effective and if necessary to seek repair.

Teacher knowledge and socio-psychological factors are revealing on the role of teachers in preparing learners for examinations. The teacher’s effect on learners’ motivation and willingness to devote time/effort to their learning is particularly important (as motivation is a strong predictor of outcome success – see Skehan, 1989). Recent work in educational psychology (e.g. Ormrod, 2003), also sees motivation as having a range of effects on how students learn and on their behaviour in other ways.

However, as students are not always internally motivated they may need to be supported by the environmental conditions (i.e. “situated motivation”). Teachers can encourage learners “to engage” in learning, to make more effort, to initiate tasks and to persist in relevant activities. They can also provide opportunities for cognitive development by promoting autonomy and by providing informative feedback (as in assessment for learning), see Dörnyei, 2002.

FLLGP evidence suggests that teachers also need to create a caring and supportive classroom environment and to provide opportunities for cognitive development by promoting autonomy, setting goals, and providing informative feedback. Dörnyei (1994) outlines three key components involved in his “situated framework of L2 motivation” claiming that learner motivation can be seen as a response to any of the three: the affiliative motive (i.e. learners’ desire to please the teacher); the “authority type” (i.e. an authoritarian, democratic or laissez-faire teaching style); and direct socialization of student motivation. Teacher behaviours that might enhance learner motivation are also set out in several other theoretical frameworks, including classroom environment research, self-determination theory and goal orientation theory.

6.6 Conclusions and implications for the expanded model of impact

In this chapter we have described and analysed a single impact study carried out to explore the micro level of language test impact within a single school context. We have used the analysis of this case to extend our understanding of impact processes at the level of the school and classroom and have examined in detail the roles that individual learners and teachers play. In this section we sum up the key points and draw conclusions.
6.6.1 Context and the impact of assessment

The discussion of the FLLGP case and the modelling of the language classroom context has shifted attention onto classroom-based language learning and the dynamics of teacher and learner interaction in that socio-cultural context. Language assessment is one of many factors within that context which influence the language acquired by the learners and on their behaviour. We summarise the main points from the discussion above under 4 headings:

a). The communicative language construct.
b). Approaches to language learning.
c). Linking formative and summative assessment.
d). New models of assessment to support learning.

a). The communicative language construct

The importance of the communicative language construct and how this interacts with assessment policies and practices emerged in the discussion of classroom-based language learning. The need to incorporate the social dimension into the model of language assessment is, therefore, as important as paying appropriate attention to the cognitive side. We bring the two perspectives together within an interactionist approach in a socio-cognitive model of language assessment such as the model presented by Weir in his 2005 volume and operationalised by Cambridge ESOL in their approach to validation (see Saville, 2004, quoted in Chapter 1).

b). Approaches to language learning

We have noted that most contemporary research stresses the need for the active involvement of learners in structuring their learning to meet their own requirements, based on their individual personalities, strengths, weaknesses and preferences. The teacher has a vital role in supporting the learner by helping to ensure that motivation remains high and by providing effective feedback. We have also seen that research focused on the language learner/test taker as an individual has tended to target the psychological dimensions (affective, cognitive, meta-cognitive characteristics). The review of the FLLGP case suggests that these are important in designing and carrying out impact studies. Central to the socio-cultural view is a belief that language learning and teaching are social acts and that the roles and relationships of the learner and the teacher are socially constructed, with social identities being formed and changed by the processes involved in learning and teaching, as pointed out by Zhu Hua et al (2007:1).
A description of the micro context of the language school and classroom needs to capture the societal structures and the cultural values embodied within the institution, the motivations of individuals and their identities, the various policies and ideologies represented and the nature of the relationships between the teachers and the learners. The concern for washback arose from recognition that tests really do affect what goes on in classrooms meaning that testers needs to take an interest in processes as well as products. But, as we have seen in the FLLGP, the current concern for impact in this thesis takes us beyond washback and the model of impact needs to recognise that the learning/teaching processes which can be observed in classrooms are not just affected by the main factors within the context, they actually help bring them about and “transform” them in ways which are both complex and dynamic.

A logical conclusion from this view is that testing systems which are designed to have positive impact need to recognise this dynamic relationship. By being situated in learning/teaching contexts, the test and testing system is modified (“transformed”) by the participants and by their own roles in the processes as active participants or stakeholders (not as passive “victims”). Assessment systems designed to achieve positive impact should facilitate this social interaction and lead to processes and changes, including language learning and growth, which are deemed to be desirable.

The impact on the teacher (his/her identity, knowledge, skills, attitudes etc) is as important as the impact on the learner given that much of what occurs in the classroom context is mediated by the teacher and is created by the interactions which the teacher facilitates. In her thesis “Searching for washback”, Burrows (1998) reminds us that ‘the teacher is a “medium” through which any change must pass”. Thus washback research should take account of teachers’ belief systems and consequent responses to change, especially where the introduction of new assessment procedures constitute an innovation. The FLLGP research endorses this view, suggesting that this kind of mediation is likely to occur in all classrooms where examinations are relevant, whether or not there is a particular assessment or curriculum innovation taking place. Effective communication and appropriate support for teachers is a prerequisite for achieving intended impacts, whether within a process of curriculum renewal or as part of the operational cycles of existing examination systems.

c). Linking formative and summative assessment

In our discussion we have asked to what extent formal assessment can really be considered formative, and whether washback and other impacts will always be
limited by the generally summative nature of the assessment process when external, high-stakes examinations are used (the dilemma which we discussed above). We have seen the Assessment Reform Group’s claim that assessment which is specifically designed to improve learning is the single most important way to raise educational standards. They would also claim a serious danger of negative results from “test-driven teaching”.

Pellegrino and Chudowsky’s “assessment triangle”: cognition, observation and interpretation (2003: 112) is clearly compatible with the Weir 2005 socio-cognitive model introduced above. Pellegrino and Chudowsky argue relatedly that our growing understanding of cognition, the greater availability and use of sophisticated technology and the application of advanced statistical models in assessment, create new opportunities to devise assessments that can impact directly and positively on learning, even perhaps when the assessments themselves are not directly integrated into learning (unlike formative and dynamic models). This would be the case for international language examinations like IELTS, PET, FCE, on which we have focused in this thesis. The technology already available enables us to observe the processes of dealing with the tasks set as assessment objectives, and the latest statistical models can help us to interpret the responses given by the learners in carrying out the tasks. Again this suggests that, as assessment systems become more technologically and psychometrically sophisticated, opportunities to bridge the formative/summative divide more effectively will emerge.

d). New models of assessment to support learning

In the classroom model, we have discussed the role of examinations in providing evidence of achievement linked to the CEFR proficiency framework. Existing proficiency examinations such as those provided by Cambridge ESOL can legitimately be considered achievement-oriented in contexts of use where there is construct overlap of the kind discussed by Green (2007a). However, these examinations have limited potential for formative uses (assessment for learning), as they cannot be fully integrated into the learning process. New models of assessment are needed to successfully implement both the formative and summative functions.

Jones (2006) has proposed a Cambridge ESOL computer-based model of formative assessment based on a simple two-dimensional model of learning. The two dimensions are difficulty (also seen as growth) and support sometimes referred to as scaffolding. He argues that the right combination of difficulty and support will favour learning, with growth measured by mastery of increasingly difficult learning points. This notion, he suggests, relates to what Vygotsky (1978) called the “zone of
proximal development” (ZPD) and he discusses how this might work in an integrated learning and testing environment. Jones’ approach is illustrated diagrammatically in Figure 6.10: the increasingly difficult language learning objectives appear on the vertical axis (for example, the semantico-grammatical concept of modality using conditionals in English), and along the horizontal axis, provision of more or less support for the learner in order to assist the learner in mastering these points while engaging with the pedagogic tasks.

The two dimensions – growth and support – define the “space” in which learning takes place, and the learner can be “located” on a diagonal indicating the range of tasks he/she should be engaging with, from easier tasks with less support to harder tasks with more support. In Vygotskian terms, this defines the ZPD and it is in this zone that learning will take place best. Learning tasks which are below this zone will present too little challenge and those above it will be too difficult (see also Poehner, 2007, above). It is not clear, however, how the DA model could be operationalised in language classrooms such as those described in the FLLGP. But Jones concludes that appropriately described and calibrated test tasks using a measurement model (e.g. based on IRT) can be used within a common frame of reference, such as the CEFR, for both formative and summative purposes with a predicted degree of impact on learning as a result. Further research is clearly invited here.
6.6.2 Implications for the revised model of impact

In conclusion, the following five points have emerged from the meta-analysis in this chapter and are of importance for the developing model of impact:

a) There is a need to understand better how language examinations interact with specific contextual features (local contexts and the hierarchical nature of educational systems).

b) It is important to understand how the individual differences of participants interact with other contextual variables.

c) Language learning needs to be understood as change.

d) How the language construct is defined and communicated to stakeholders is crucial.

e) Impact research needs to make better use of case studies and mixed methods approaches to investigate assessment processes at the micro level.

a). Local contexts and the hierarchical nature of educational systems

The nature of educational systems needs to be better understood and taken into account, especially the importance of local context. This concurs with Fullan (1999:21) who states “local context (readiness to learn, local capacity, etc) is a crucial variable and no program can expect to spread successfully if it does not take into account the various contexts which it will inevitably encounter.” Educational systems are typically organised hierarchically and the overall approach to education in many countries is influenced by a range of socio-political and socio-cultural factors and especially by decisions made by government at the highest level in the hierarchy (e.g. the Ministry of Education as we saw in Chapter 5). Within hierarchical models, the lowest-level units are the smallest and most numerous, whereas the higher-level units are larger and fewer. In this chapter we have analysed the BIF in relation to its hierarchical structure and this has provided insights into the interaction between the micro context of the language school/classroom and the wider macro context of language education within Italian society.

In the impact research we have reviewed so far, we have referred to the macro level of impact as the interplay between language examinations and the overarching socio-political/socio-cultural factors. The micro level of impact, on the other hand, is concerned with interplay between the lowest levels in the educational hierarchy and the possible impacts on individuals in localised contexts (i.e. schools and classrooms). The importance of understanding the local contexts better and of using appropriate data collection methods to do so has been confirmed in our analysis in this chapter.
Educational hierarchies do not occur accidentally and so is important to understand the nature of the hierarchical relationships which are set up or which continue to evolve as innovations and changes are brought about. In this chapter we have not only attempted to investigate what happened at the local level of the hierarchy (within the BIF), but we have also considered the interactions between the school and wider socio-political contexts (local, regional and national).

In the analysis of the FLLGP, we modelled the school context with the lowest-level units (139 learners nested within twelve classes in a single school). As the BIF is a private school there were no formal relations with the higher levels in the educational hierarchy, i.e. the local education authorities or the regional ministerial departments, as was the case in the PL2000. However the temporal context and immediate antecedents to the FLLGP study were provided by the PL2000 reforms within the national system under the authority of the Ministry of Education. These reforms provided important influences from the wider milieu, including the prevailing policy on using the CEFR and government encouragement for taking international language certification. In this respect, we observed the potentially dynamic interaction between the school and the wider socio-political milieu in relation to its assessment policy. Complex reasons for choosing particular examinations existed, partly influenced by socially-oriented concepts (prestige, image, snob value, etc) relevant to the local context. The examination culture within the school also impacted on the individual classes in different ways, and certain examinations (particularly FCE) had stronger influences on both teacher and learner behaviour than others.

At the micro level the classrooms were characterised as social-cultural contexts which provided “opportunities for learning”; we noted the individual differences within learner groups (within the same classes) and that variation “emerged” as a result of in-group interactions among participants. Learners within a specific classroom context were therefore potentially influenced by all levels in the educational hierarchy (national, regional, local, school, class and group).

Finally, there are several statistical approaches which are suitable for comparing school contexts based on hierarchical models and which can be used in both experimental and quasi-experimental designs. One approach uses structural equation modelling (SEM) and the other uses multi-level modelling, sometimes also called hierarchical modelling (Goldstein, 1995). Both approaches are used extensively in educational research and offer productive possibilities for use within future test impact projects. But it is by no means certain that such approaches would
tease out the kinds of subtle individual differences and group interactions that the exploratory and mainly qualitative FLLGP achieved to inform such future impact study.

b). Individual differences of the participants and contextual variables
It is important to understand how individual differences of participants interact with other contextual variables. This had been recognised by the team that set up the FLLGP and was confirmed as a significant issue in impact research by the analysis we have carried out of the project. The nature of the difficulties in capturing and interpreting the data within impact research projects was also instructive. The LLQ did not deliver as much usable data as anticipated, partly because the questionnaires themselves proved unpopular with participants, especially some of the teachers, and partly because the constructs underpinning the questionnaires did not fully address key issues related to the students as language learners and test takers. We may conclude that constructivist concepts such as “learner identity” and “personal constructs” need greater attention and we need to capture more data on the ways that learning goals can be supported by tests and examinations. This needs to be pursued in light of recent models of motivation and the role that teachers and other contextual variables play in this.

The FLLGP supports further investigation of constructivist concepts in relation to teacher differences and the importance of the teacher as a “mediator” within classroom processes and between the examination and the learners, a crucial factor in achieving intended impact.

c). Language learning/teaching needs to be understood as change
In studying impact, it is important to identify changes which might be attributable to assessment systems as they occur over time (the time frame being an important variable). We also know from earlier chapters that understanding and accounting for change processes are important steps in setting up impact-related research.

The FLLGP revealed limitations in the setting up of experimental studies within real classrooms, and confirmed the difficulty of establishing causal relationships in impact research. Non-experimental or quasi-experimental approaches may, however, be more feasible and these methods need to be developed within impact research (see below).

If, however, the required constructs can be operationalised, observations can be treated as measures and statistical procedures can be used to model the change
(e.g. SEM and multilevel models, noted above). Such models may allow us to investigate whether there is systematic change within individuals, and also whether the pattern of change varies from individual to individual. In other words, we can reach some conclusions about individuals as they change over time, and also how this change compares with the other individuals in the same group. The potential for using these techniques within the “impact toolkit” should be pursued.

In our discussion of language learning we focused on changes in language proficiency, as well as changes in other language learning behaviours. We noted the following influences on learners:

- The dynamics of interaction between teacher and learners in the classroom.
- Teaching processes, including goal setting and feedback, building and maintaining learner motivation, monitoring output and evaluating progress.
- How and when the assessment takes place and when feedback is given.
- How test results are presented and interpreted.

Learning within the psychometric tradition had been associated with behavioural psychology and the concept of the “trait”; growth over time was typically defined as a vertical hierarchy of increasingly generalized and abstract knowledge and skills. From the socio-cultural perspective, learning takes place within a social context and depends of the relationships between the human participants and the resources that are available in that context. Learning, therefore, not only involves acquiring new knowledge, but also involves taking on a new identity and social position within a particular community of practice. This view is consistent with recent approaches to learning within cognitive psychology which encompasses a developmental as well as an information-processing perspective.

Constructivist views of learning also suggest that the assessment for learning needs to be integrated and individualised in order to support future learning (e.g. involving classroom-based assessment). However, we highlighted the dilemma in bridging external and internal uses of assessment and noted that current approaches to formal assessment have difficulty in increasing positive impact on individual learners (through integrating and individualising assessment) while at the same time maintaining generalisability and accountability within the wider community.

Engaging with learning while providing valid interpretation of outcomes is both a major challenge and an important goal for examination providers. In future it may be
possible to develop language assessment systems which combine both formative and summative focuses of assessment within the same unifying frame of reference.

d) How the language construct is communicated to stakeholders is crucial
Jones and Saville (2008: 507) claim that language proficiency frameworks, such as the CEFR, make important contributions to our understanding of learning and assessment in educational contexts by providing the basis for communication between stakeholders. The CEFR itself is “context neutral” and “under-specified” in many areas and this generality makes it suitable for relating specific contexts to each other without imposing unified solutions. Fullan (1999: ix) suggests that “organisational learning” takes place best when a balance is maintained between “too much and too little structure”. Frameworks such as the CEFR may be able to provide a suitable balance of structure. The CEFR provides a sound basis for effective communication between stakeholders, both inside and outside the school. Central is the construct of communicative language ability which we have characterised as being compatible with a socio-cognitive model. Using the CEFR as the basis for communication between stakeholders, it is possible to relate learning to assessment, and the results of assessment to the “real world” in meaningful ways (cf. Green’s washback model).

Jones and Saville (2008: 498) outline a five-step process of test validation, from test construction to “real world” interpretation, shown in Figure 6.11 (see also Jones, 2007). The steps make up a chain of inference through which interpretations, or in other words the various meanings of test outcomes (scores etc.) are supported. This is consistent with the work of Kane et al (1999), Almond et al (2002) and Mislevy et al (2002) in educational assessment, and Weir’s application of a similar approach in language testing (2005).

Figure 6.11 A five-step process of test validation
e). **Impact research needs to make use of case studies and mixed methods approaches to better understand micro level processes**

The FLLGP employed the research principles and procedures which had been developed in the earlier impact projects (described in Chapters 4 and 5). Many of these were endorsed, including the following:

- The importance of the action plan for designing and conducting impact research, especially the need to apply relevant project management techniques to establish the scope, timeline and deployment of resources. In this respect this role of the Project Coordinator was critical.
- Feasibility and practical considerations need to be prioritised in planning and conducting impact research in schools and classrooms.
- It is important to involve stakeholders as participants and to seek their interest, commitment and active support.
- Impact research is concerned with the relative effects of the numerous, complex and inter-related intervening variables which need to be better understood.
- Access to data for comparative purposes is an important aspect of impact research.
- The importance of developing working impact hypotheses was confirmed and the importance of identifying the dependent variables and seeking ways of handling independent (intervening) variables.
- Researchers need to take advantage of opportunity samples and to act rapidly. The value of having an “impact toolkit” available for rapid deployment was confirmed, with the need to continually develop and adapt relevant data collection and storage techniques.

By focusing on the micro level of impact, the FLLGP provided an opportunity to explore the use of hermeneutic approaches, as noted above. In particular, the FLLGP confirmed the potential for using case study methods for gaining insights into the complex issues surrounding the use of language tests and examinations within school/classroom contexts. The “thick” description of the school, the classrooms and the individual participants was one of the most successful aspects of the project and was provided useful insights for refining the impact model. It nevertheless posed significant problems related to data handling and follow up analysis.

The FLLGP also raised a number of issues which need to be addressed in order to collect and analysis impact data from micro contexts more effectively. From the point
of view of an examination board, operational impact projects need to fit within a range of other validation priorities taking place at any one time. It is also necessary to ensure that insights gained can be followed up appropriately, leading to action.

We can conclude that:

- Instruments and procedures need to be better adapted for data collection in learning contexts such as the BIF.
- Individual learner data needs to be collected more effectively based on current socio-cultural and cognitive models of learner characteristics.
- Technological solutions are needed to deal with the difficulty of capturing and storing micro level data, especially within an operational context.
- Procedures are needed to analyse rich data collected during case study research in specific contexts: accrued data needs to be aggregated and interpreted effectively.
- Insights obtained from micro level impact research should lead to follow up action, including recommendations for improved practices in teaching and assessment.

Finally, in order to extend Green’s concept of “washback by design”, it is necessary to go beyond a concern for test constructs (however important that may be) and to encompass concerns for the socio-political contexts of assessment, and the dynamic interaction between learning contexts and the wider milieu. In so doing, we broaden the conceptualisation of language test impact and begin to suggest ways in which examinations board can integrate impact considerations into their working practices – what might be termed “impact by design”.

In the next chapter, we bring together the insights from the three case studies in proposing the expanded model of impact.
Chapter 7

The revised model of impact and future applications

7.1 Introduction

In this chapter the outcomes of the three case studies analysed in Chapters 4, 5, and 6 are summarised in terms of the specific contribution which each has made in revising the model of impact. The revised model is then proposed and discussed in relation to Cambridge ESOL examinations.

While the case studies covered developments up to 2004, references to relevant work in the literature since 2004 were added as the thesis developed. As a contribution to knowledge more generally, ways in which the model might then be applied in other contexts are discussed. The limitations of the research are discussed before final conclusions are drawn.

7.2 Overview of the three case studies

The three case studies conducted within the context of Cambridge ESOL examinations described in Chapter 1 provided the basis for the meta-analyses. In this section the contributions to the revised model from each Case Study, IELTS, PL2000, and FLLGP, are summarised.

Case 1 described the original impact model which was developed within Cambridge ESOL in the early 1990s, informed by work being carried out on washback elsewhere (Alderson and Wall, Bailey, Hughes, etc., see above). This model conceptualised impact as one of the “VRIP features” within test development and validation and was also notable for the development of suitable instruments for the collection of impact data. Four impact projects related to IELTS as part of the ongoing programme of validation following the major revision - IELTS 1995 - provided the meta-data for the analysis. This case was seen as an opportunity to problematise the nature of impact data to be targeted and the development of the necessary instrumentation to collect those data.

The IELTS case, therefore, was a first iteration of the project to develop the expanded impact model over the period 1993 to 2003; the lessons learned were summarised and informed Case 2.
The importance of key points of advice from Alderson (1995) were confirmed: to identify and anticipate potential areas of impact using “washback hypotheses”; to collect antecedent data; to collect attitudinal data; to recognise the importance of systematising impact research. The experiences of the four IELTS impact sub-projects also suggested ways in which these points need to be modified and extended when applied to the work of the examinations board.

Case 2 was an impact study within a single “macro” educational context. This context was the Italian State system of education and a specific government reform project, i.e. the Progetto Lingue 2000, intended to improve standards of language education at the turn of the 21st century at a time of political and educational change more generally. In other words, the PL2000 case focused on a single “milieu” in ways which had not been possible with the international IELTS projects. The impact of the reforms in general, as well as the specific role of the high-stakes “external” examinations (e.g. provided by Cambridge ESOL) formed the basis of this case.

Against the backdrop of the system as a whole, a small number of “micro” contexts at the local level were also examined in detail. All cycles of state school education - primary, middle, and upper schools – were included, as well as a geographical spread of settings covering regions from the North, Centre and South of the country. These case studies provided unique insights into what was actually occurring in language classrooms and confirmed the importance of investigating possible mismatches between stated intentions (anticipated impacts) and what actually occurs in classroom contexts.

Several outcomes from the IELTS projects were endorsed by the analysis of Case 2, but in addition, the analysis of the PL2000 case provided new insights; importantly it showed that anticipated washback based on broad construct definitions can be achieved in practice where there is adequate communication and support for those involved in the implementation. The use of the CEFR as the basis for construct definition and as a communication tool was identified as a key factor here, as it enabled major participants to refer to a common system, encapsulating a shared understanding of proficiency levels and a conceptualisation of a communicative approach which were essential elements of the government reforms.

In Case Study 3 a single impact study was described and analysed. Building directly on the PL2000 case, it focused on a single micro context in Italy and involved teachers and learners preparing for a range of English language examinations in a small number of classes at a prestigious language school in Florence. This impact-related project, known as the FLLGP - Florence Language Learning Gain Project – was itself an example of a case
study, and as such it provided an opportunity to review case study methods and the wider applicability of this kind of research for the impact model.

Again, many of the insights from the other two cases were reconfirmed, while new points emerged from the analysis to contribute to the expanded model. In particular the relationships and interactions between the language assessments and the learning/teaching taking place in a small number of language classrooms were examined against the backdrop of the wider educational and societal milieu in Italy described in Case 2.

In the design of the FLLGP attempts were made to isolate and evaluate the influence of the assessment procedures, and the Cambridge examinations in particular, on the “growth” of the main participants – i.e. the learners and their teachers. The detailed description of contextual features at the micro level, as well as the longitudinal nature of the project conducted over an academic year, were particularly relevant features of this case. The in-depth coverage of individual teachers, their views and approaches to teaching and assessment, was a significant feature of this case. The importance of case study methodology and of qualitative data collection to impact research was confirmed, backing up the conclusions of the other two Cases.

7.3 Lessons learnt from case studies

The analysis and discussion in each case study was broadly structured around the features of the washback model which had emerged by end of the 1990s as discussed in Chapter 2. These seven features were, it will be recalled: the test itself; the context; the participants; the outcomes; the researcher; the research methods; and the timeline.

7.3.1 Test features

The IELTS case confirmed that Cambridge ESOL’s construct-based approach to test design and development should combine test taker and contextual models within a socio-cognitive approach. Validation procedures which anticipate impact at both individual and systemic level should be planned. Building on Messick’s unitary concept of validity, evidence should be accrued over time in developing a validity argument, as suggested by Mislevy and others (noted in Chapter 2 above).

The IELTS case in particular showed a need for the examination provider to:

- relate test impact research explicitly to the underlying test construct;
- accrue data over time as evidence of validity in specific contexts of test use;
• monitor test centres and test sessions routinely with a view to identifying instances of unplanned outcomes;
• identify the differential functioning of test materials/procedures in certain contexts and to determine the extent to which specific features of the system influence candidates’ behaviour (i.e. in taking the test or in using the language in target situations).

The nature of the language construct within the targeted teaching/learning contexts was also a particular focus in the FLLGP case. Apart from confirming the importance of construct definition generally, as in the washback literature of the 1990s, the ways in which it is defined and how it is communicated between participants was noted as especially important. Being able to know about and to understand the implications of the construct (as defined by the test provider) is critically important for key stakeholders in teaching and test taking contexts. Failure to know or understand enough can lead to constructs being wrongly interpreted and to unintended consequences. The analysis of the FLLGP case confirmed the potential usefulness of communicative devices such as frameworks and related metaphors (e.g. “learning ladders”) in providing the basis for successful communication between the stakeholders. The CEFR proved to be a significant feature in both the PL2000 and FLLGP projects; as a generalisable framework the CEFR was helpful in fostering the Italian government’s twin objectives of greater autonomy for schools on the one hand, and greater accountability of learning outcomes on the other.

7.3.2 Context

The cultural and geo-political settings in which an international test is administered and in which the results are used, provide particular challenges for an international examinations board. The IELTS case was based on the premise that examinations board operating internationally need to develop an in-depth understanding of all the relevant test-related contexts wherever the tests are used. In particular, the international dimension of IELTS suggested that the examination provider should seek to understand ways in which different local contexts can influence the uses and consequences of their tests. Changes in these contexts which might be attributable to the use of the test need to be carefully monitored to ensure that such changes are not detrimental.

The IELTS case showed a need for the examination provider to:

• take into account more effectively the ways in which test development and test use contexts relate to specific teaching/learning contexts;
investigate how multiple individual contexts relate to each other, and to compare similarities and differences between the test taking contexts within a single country and between countries;

find ways to aggregate information on the basis of data obtained in many different contexts to establish general trends and tendencies.

In Case 2 the importance of understanding the macro context (or milieu) in terms of socio-political features was discussed and exemplified. An important feature of the Italian milieu was the political landscape and the changes in government which occurred during the timeline of the project. Individual stakeholders in their different roles were also an important focus (see discussion of participants below).

Figure 4.4 (introduced in Chapter 4) showed within the highlighted box some of the components of the organisational networks which create the milieu within which the educational processes of interest take place.

Figure 4.4 – Organisational networks

The Cambridge examinations, as part of the wider PL2000 project, offered the Cambridge-based research team the opportunity to carry out another impact project building on the IELTS work, which was still going on at the time. The project provided an opportunity to focus on the processes operating between the participants and the products at the school level (i.e. the “empty box” in Bailey’s model – see Chapters 2 and 4 above), as well as introducing several key factors influencing these processes from the wider milieu, (e.g. the use of external certification and the adoption of the
As a frame of reference for developing curriculum and materials in local contexts.

This was shown in Figure 5.4 introduced in Chapter 5 by the red and blues boxes overlaid on Bailey’s original figure.

It was noted that the socio-political context of reform projects can create difficulties for impact research because the timescales and other variables are controlled by Ministries of Education and the political agenda of individual participants (politicians, civil servants, administrators etc). In such cases, the examination provider may find it difficult to control the variables sufficiently well to capture adequate baseline data or to secure representative samples of data to conduct quasi-experimental research. This may be because others in powerful roles set the agenda or because it is difficult for the examination provider to enter into local contexts without being seen to interfere.

While focusing in particular on the influence of the wider context, a key part of the study and a major contribution to the model of impact was the opportunity to focus on the interaction between factors in the macro and micro contexts. The role of the central Ministry officials in the project and other local protagonists in this impact project was particularly noteworthy, seen by Cambridge ESOL as a positive engagement with key stakeholders, as well as a major contribution to the success of the study. The case study schools provided a more detailed focus on the micro systems than the IELTS projects.
This case confirmed the complex and dynamic nature of national educational systems; the review of Italy’s educational system and the political dimensions of the reform programme which led to PL2000, demonstrated the importance of understanding the inter-relating sub-systems in which educational change takes place at different levels of organisation. This study therefore confirmed the importance of levels and scales of organisation within complex systems and the need to take the factors into account in conducting impact-related research.

The FLLGP was set up specifically to explore the micro level of language test impact within a single school context, and the individual characteristics of the participants, including their attitudes and behaviours, were a particular focus. The data collected as a result produced a “thick description” of the school context and the classroom settings within it. This provided valuable information about the process of impact at the micro level but also posed practical difficulties in collecting and analysing the large amounts of data.

The study, contextualised as it was within the socio-political milieu of Italy, confirmed the importance of understanding the ways in which variables which are determined locally within schools and classrooms interact with contextual parameters determined by influences in society more broadly.

7.3.3 Participants

Given the importance of context and setting, a wide range of participants and stakeholders were identified in the IELTS studies.

In particular the IELTS case confirmed the need for the examination provider to:

- consider who the key stakeholders are and to identify what roles they perform within the various relevant contexts;
- see the participants not only as group members but also as individuals;
- monitor a wider range of individual differences using appropriate instruments;
- see the test developer as a key participant in both conceiving and conducting impact research as an integral part of the test development and validation process;
- consider the appropriate roles for staff within the examinations board in conducting impact research, (e.g. as “outsiders” participating in local settings).
The PL2000 case identified specific individuals who were instrumental in implementing the project objectives at various levels. The importance of communication and dissemination of information to individual or particular participant groups was also noted, and especially the need to adjust messages to communicate successfully with the various target audiences.

The FLLGP building on the PL2000 focused on individual participants at the micro level. The project allowed for the perceptions, views and attitudes of these participants to be examined based on the collection of rich, qualitative data and for the actions taken and the outcomes achieved to be observed fairly intensively. There was a particular focus on the learners/candidates in their school and test taking settings; their voices and the voices of their teachers were heard in the qualitative data based on observations, interviews, recordings and documents kept by the principal researcher (whose own “narrative voice” also provided important data).

A conclusion was that individual participants and stakeholders can be considered as “promoters” or “inhibitors”, and thus can either facilitate or hinder the intended objectives of an innovative programme such as the PL2000. They can influence the general direction, the speed of take up and the “intensity” of the intended changes, (as described by Green in his washback model – see Chapter 2). The participation of Ministry staff in the Lombardy region provided a good example of how “early adopters” can promote positive attitudes and influence the take up of tests to positive effect in schools.

7.3.4 Outcomes

The revision process which led to IELTS 1995 enabled the development team in Cambridge to highlight a number of desired outcomes “by design” including features of impact. The IELTS case showed a need for the examination provider to:

- anticipate desirable outcomes in wide range of relevant contexts – in this case internationally in many countries;
- recognise that unanticipated outcomes become apparent as systems develop over time;
- seek insights into what actually happens in typical settings based on multiple sources of data;
- feed the insights from the impact research back into the testing system to bring about on-going improvements to the assessment system.
In the PL2000 case, the intended outcomes of the reforms were set by the government and the Cambridge team came into PL2000 as “outsiders”, both within the Italian milieu generally and within the selected case study schools. This “enforced” role influenced the formulation of the impact hypotheses and the type of involvement of the researchers in designing and carrying out the study.

The Cambridge staff (unlike in the IELTS case) had not been involved in planning the PL2000 project and had no official mandate to make or directly influence policy decisions prior to entering into the agreement with the Ministry. This meant that they were not able to redesign or modify the examinations to fit into the context prior to entering into agreement with the government at the start of the project. (i.e. there was no opportunity to “seek validity by design as the basis for intended washback” through the development of specifically designed tests – Messick, 1996; 252). However, the existing examination systems which were deemed suitable by the Ministry (under recommendation by the British Council) were amended in important ways, for example by introducing extra administration dates, by providing additional information for stakeholders and by supporting teachers with extra seminars and training events. The evidence from the study also suggested that the Cambridge examinations did fit into the overall aims of the project established by the Ministry, and despite requirement to match existing examinations to the required test contexts and purposes, the observed outcomes were generally positive and supportive of the government’s initial intentions.

As a result of the focus on classroom settings in the FLLGP case, a main conclusion was that examinations board involved in impact research need to reconsider nature of language learning and teaching in formal contexts when formulating their impact hypotheses. Classroom-based impact research needs to be informed by contemporary approaches to second language acquisition and teaching and the latest literature in this field based on constructivist theories of learning. The social dimension of the language proficiency construct, and importance of providing opportunities to use language interactively in learning contexts, needs to be reflected in assessment practices and should be an important consideration in impact research.

7.3.5 The Researcher

The role of the test developers within an examination board as co-researchers with other participants was an important feature of the IELTS projects. While it was important to involve “external” academics, it was equally important to involve those directly concerned with operational systems and in the teaching/learning contexts.
where candidates were preparing to take the test. There was recognition of the difficulties that examinations board encounter associated with obtaining valid data from “other people’s contexts” (which are often geographically distant from where the tests are developed and validated).

In particular the IELTS case showed a need for the examination provider to:

- reconsider insider/outsider perspectives in impact research and determine how valid data can be obtained more efficiently;
- find better ways to involve local participants as researchers within their own contexts.

In the PL2000 case, the role of the principal researcher (R. Hawkey) was very important and the development of relationships between the “outsiders” (researchers from Cambridge team) and the “insider” participants in the various local settings was a crucial aspect of the study. This was also the case in the FLLGP. An important conclusion is that the role of the test developer/researcher as “outsider” is a critical issue that needs to be addressed more systematically. Research methods need to be developed taking this into account and impact researchers from examinations board should be prepared to redevelop or adapt their instruments, procedures and project staffing to operate within new contexts in which they start off as outsiders.

7.3.6 Research methods

The research methods used in the IELTS case originated in the relevant language testing literature and the prevalent (generally positivistic) influences. However, the analysis of the IELTS case suggested a need to refer more widely to the methodologies employed in the social sciences and broader educational research community. The value of using case studies as a way of obtaining rich data and unique insights into local contexts was noted. This has had a significant influence on recent IELTS research priorities agenda (see IELTS Research Reports Vol. 8, 2008) and on the developing impact model. The importance of problematising impact in terms of hypotheses linked to other aspects of the VRIP features (especially the language construct) was a starting point, but this was strongly endorsed by the IELTS case and is now thought to be a fundamental feature of the expanded model. Above all, the IELTS Impact projects provided important practical and procedural insights into suitable research methodologies, especially related to the design and validation of instruments and procedures. The storage and access of the complex, interrelated data sets which are generated by large-scale, mixed method designs was identified as a major practical difficulty which needs to be addressed more
effectively. The IELTS case also showed a need for an examination provider to develop practical and ethical guidelines and to be aware of legal considerations in setting up and managing operational projects in other people’s settings.

Building on the IELTS experience, *mixed methods of data collection* were successfully employed in the other two cases, including qualitative techniques and case studies, as well as more traditional survey data and the collection and analysis of demographic and score data.

From both the PL2000 and FLLGP cases we can conclude that the research methods which are suitable for micro-level impact research designs share a number of methodological principles with other forms of social research. For example, in order to be *ecologically valid* it is important to recognise that the local context is a dynamic part of the system under investigation; the people and processes should not be considered as “variables to be controlled” but the varying degrees of stability or variability are important aspects of context which need to be accounted for in other ways. This conclusion points impact researchers in the direction of approaches which draw upon action research, ethnographic studies and “formative experiments”.

The FLLGP confirmed the usefulness of case studies and mixed methods in conducting impact research at the micro level; however, the project also raised a number of issues which need to be addressed in order facilitate the collection of data from micro contexts. While experimental and quasi-experimental designs can be used in some cases, greater effort needs to be put into the collection of data using non-experimental techniques. The richness and complexity of the qualitative data in the FLLGP confirmed the need to develop non-experimental techniques which can be deployed more effectively. Moreover, the ability of researchers in examinations board to handle large quantities of qualitative data while operating under operational conditions was identified as a major challenge. Inevitably there are implications for resources and for the deployment of staff when required.

### 7.3.7 The timeline

An explicit focus on the timeline was not an aspect of the washback models which had influenced the early phases of the IELTS case (i.e. Hughes, 1993 and Bailey, 1996). However, the analysis of the IELTS case confirmed the importance of making comparisons in impact research and the need for comparative data in order to do so. In particular, the need to be able to *monitor changes over time* meant that a longitudinal dimension with an explicit timeline needs to be accounted for in the expanded impact model.
In particular the IELTS case showed a need for the examination provider to:

- monitor changes over time implementing longitudinal approaches to data collection (e.g. in stages or phases of activity and employing appropriate research designs);
- adopt a project-based approach to test development and validation which takes into account the timeline and uses appropriate project management techniques to secure resources when required;
- address the difficulty of collecting useful baseline data and develop ways to handle real world problems of collecting data over time.

The IELTS case focused on the historical development of the examination itself and the ongoing evolution of the IELTS system during the phases of the impact projects; this led to a focus on change processes and innovation and flagged up the importance of the timescale in impact-related research (e.g. to monitor changes over time).

The PL2000 case provided a socio-political context with its own historical dimension. The specific goals of the educational reforms and the associated timescale were imposed by the Italian Ministry and were constrained by the features of the educational system itself. The analysis of this case reinforced the importance in impact-related research of specifying an overall timeline with time-bound phases of activity within it. In this respect, the case confirmed the usefulness of the literature on innovation and the applicability of process-based models of change; such models allow the impact researcher to categorise the many factors at work in a complex socio-cultural and socio-political milieu and can help to illustrate how these factors interact with one another within educational processes (where assessment policies and specific tests or examinations play a role). The concept of timescales and the need to define a timeline is an important feature of these approaches; Henrichsen’s Hybrid Model, for example, specifies a time-bound implementation process, with the before, during and after features. The Hybrid model, adapted by Wall (2006) in her Sri Lankan study, also proved useful in the analysis of the IELTS case, and while the Hybrid model is not a predictive model, such an approach can lead to more effective planning of impact research based on a clear understanding of pre-existing contextual features and of the potential changes which might occur.

The need to consider learning as “growth” was highlighted in the FLLGP case. Language learning itself and the developmental processes which classrooms are
designed to support, need to be understood as processes of change which affect all participants involved in schools and classrooms – especially those “doing the learning” (the learners themselves) but also those supporting the learning (teachers and others with pedagogical roles). The desirability for language test developers to conceptualise impact-related research within assessment systems which are intended to support learning (to a greater or lesser extent) was identified. In studying the impact of the system, rather than just the test or examination, it is important to be in a position to monitor changes as they occur over time and to consider how they might be attributable to specific features within the assessment system as a whole (and not just to the focal construct as in earlier washback research).

This insight also has implications (see Chapter 6 above) for methodological principles in researching second language development in the context of test impact research (e.g. the use of time-series modelling, latent growth and latent curve analysis, etc.).

7.3.8 Summary
In summary, the conclusions in Chapter 4 suggested that the impact model could be extended based on the insights from the IELTS case and by considering in more detail some or all of the seven features (noted above) in relation to other cases. A particularly important dimension which emerged was that the examinations board must take a long-term view. The need is for iterative processes and a longitudinal approach to impact research.

The IELTS case demonstrated how various kinds of impact “emerge” over time along with the evolution of the testing system itself; the growth of candidate numbers, the increase in stakeholders of various kinds and the extension of test uses meant that impact (consequence and effects of all kinds) increased proportionately. This suggested that, as the context variables increase and become more complex, so do the stakeholder categories which are relevant to those contexts. It also means that prediction is not possible and the implications of this need to be taken into account by the examinations board; for example in designing impact research studies and in anticipating the need for future action to mitigate undesirable consequences of tests. With this in mind, “anticipatory impact research questions” should be developed more effectively with the research and validation procedures.

The PL2000 provided a clear example of the complex nature of educational systems and the dynamics of change processes which operate within such systems. This confirmed that impact research, which as we have seen is concerned with
anticipating changes and in making comparisons over time, cannot be conducted effectively without developing better understandings of dynamic systems and how they operate.

The FLLGP case confirmed from the point of view of the examinations board, (and as also suggested by the other case studies), the need to design operational impact projects to fit in alongside a range of other validation priorities. Unless the necessary procedures can be streamlined, impact projects will tend to be de-prioritised in favour of other validation activities which are perceived to be easier to carry out, or to be more urgent.

To begin to address this dilemma, appropriate validation procedures need to be developed for the collection of impact data, including the operationalisation of non-experimental research designs, such as case studies. The implications of this need to be taken on board by those who manage the test development and validation activities within the board.

7.4 The revised and expanded model – a meta-framework

The research question from Chapter 2 (page 59) asked:

*What are the essential components of an action-oriented model of impact that would enable the providers of high-stakes language examinations to investigate the impact of their examinations within the educational contexts in which they are used?*

This question has been addressed based on the insights from the three case studies and the essential components have been assembled into the expanded model of impact, as shown in Figure 7.1. This model provides a *meta-framework* which builds on Milanovic and Saville’s maxims (1996), and as such constitutes an action-oriented approach. It has four inter-related dimensions:

Dimension 1: re-conceptualise the place and role of impact study within the assessment enterprise, vis-à-vis societal systems generally and language education specifically.

Dimension 2: introduce the concept of “impact by design” into the planning and operationalisation of language assessments by examination providers.

Dimension 3: re-organise validation procedures to incorporate impact research into operational activities to provide the basis for knowing about and understanding how well an assessment system works in practice with regard to its impact (as defined in point 1 above).
Dimension 4: develop an appropriate theory of action which enables examination providers to work with stakeholders to achieve the intended objectives, to avoid negative consequences and to take remedial action when necessary.

Figure 7.1 The revised and expanded model of impact

These four dimensions are outlined in more detail below.

### 7.4.1 Dimension 1: re-conceptualisation of impact

A re-conceptualisation of test impact is needed as the basis of this revised model drawing on theories in the social sciences which go beyond the work in applied linguistics and measurement which had been used in developing the earlier concept of washback. In other words a wider theoretical basis is needed.

The re-conceptualisation must also be based on a world view or “stance” which is appropriate to 21st century societal processes and must take into account recent ontological and epistemological developments. This stance will provide the basis for the appropriate research to be conducted and for the necessary actions to be taken in the other dimensions of the model summarised below.
Given the range of references and approaches which emerged in the analysis of the three case studies, the ontological approach which is suggested as the basis for this re-conceptualisation draws on “critical realism” in the social sciences (for example associated with Andrew Sayer in the UK, e.g. Sayer, 1984 referred to in Chapter 3) and contemporary views on pragmatism derived from the philosophy which originated with Charles Sanders Peirce in the late nineteenth century.

This stance will allow for recent advances in a number of theoretical areas to be drawn upon to inform the re-conceptualisation and the other three dimensions of the meta-framework described below.

From the discussion of the three case studies, the following areas are suggested as candidates for further investigation by impact researchers:

a). theories of innovation and change within complex, dynamic systems;

b). socio-cognitive theory and its application within educational contexts;

c). constructivist theory and constructivist approaches within social systems, including implications for research methods;

d). theories of knowledge and of learning, especially language learning and teaching within formal educational contexts.

Assuming a realist or pragmatic stance, the relevance of these four areas to the re-conceptualisation is now briefly discussed.

a). Anticipating and managing change over time is a key aspect of impact research, noting the importance of timescales and the timeline (change over time, planned and unpredicted) with recurrent cycles (before/during/after).

With regard to theories of change, we can usefully refer to the recent educational literature on management of innovation and the mechanisms which can be put in place to anticipate and achieve desirable outcomes through change processes. Fullan (1991, 1993, 1999) has emphasised the complex nature of change in educational processes and appeals to a combination of new theories, including complexity theory. He suggests that the solution to achieving productive educational change “lies in developing better ways of thinking about, and dealing with, inherently unpredictable processes” (1993:19). Fullan’s work also points up the social dimension of education and the relevance of theories of social systems and practices to assessment which have also been a focus of attention in language testing circles in recent years (e.g. McNamara and Roever, 2006).
b). Socio-cognitive theories which place importance on both social and cognitive considerations are particularly relevant to the conceptualisation of language constructs and to the development and validation of assessment systems (e.g. as discussed in Chapter 1: see Saville, 2004; Weir, 2005). A socio-cognitive approach is important from the language assessment point of view as we have seen in discussing construct definition; an adequate construct definition is a necessary requirement in achieving valid language testing systems, and is therefore of particular relevance to Dimension 2.

However, the modified research methodologies needed to investigate the impact of examinations within their socio-cultural contexts (as shown in the case studies) indicate that insights from socio-cognitive theory might also be helpful in conceptualising the research, e.g. in understanding how language learning and preparation for examinations takes place in formalised learning contexts.

The literature on social psychology may be relevant here as social psychologists typically seek to explain human behaviour in terms of the interaction between mental state and social context; this is an important aspect of impact at the micro level as we saw in the case studies.

c). Constructivism is important for the re-conceptualisation of impact for two reasons; first because contemporary approaches to teaching and learning in formal contexts now appeal to constructivist theories (noted in Chapter 6); secondly because it underpins the research paradigm which is most appropriate to finding out what goes on in contexts of test use, as we have also seen in the case studies.

d). Contemporary theories of knowledge and of learning, especially language learning and acquisition, need to underpin the re-conceptualisation of impact and to play a more prominent role in the study of impact.

From the learner’s perspective, we have seen how affective factors are vital for motivation to learn; feedback from tests that highlights strengths positively tend to lead to better learning. Also we have seen that in a socio-cognitive approach, social interaction is essential for language earning to take place effectively. The constructivist concept of the zone of proximal development (ZPD) from Vygotsky suggests that learning takes place at the level which is just above what a learner has mastered, but where he/she can manage with support (echoing Krashen’s proposed “i+1” level in language learning). These considerations are relevant in designing
language assessment systems with formative (learning-oriented) objectives and to the investigation of whether these objectives have been met (a concern in impact research).

In summary, the stance taken in re-conceptualising impact extends the epistemological influences which guided Messick and his predecessors in the development of validity theory in the second half of the 20th century. Messick explicitly referred to the philosophical perspectives of Leibniz, Locke, Kant, Hegel and Singer, and to the influences of their rationalism and logical positivism on the nature of scientific enquiry in the 20th century (Messick, 1989: 30). In moving beyond Messick into the 21st century, the influence of post-modernism cannot be ignored, but for examinations board and language test providers an epistemology which can provide the basis for action is required (e.g. Sayer's philosophy of social research cited above). A realist stance therefore underpins the suggested re-conceptualisation of impact and the other dimensions of the meta-framework.

7.4.2 Dimension 2: impact by design

The concept of “impact by design” has emerged as a key feature of the expanded impact model and is the second dimension of the meta-framework (see Chapter 6). In the first place, this means designing tests which have the potential for positive impacts, including well-defined focal constructs supported by contemporary theories of communicative language ability, language acquisition and assessment (cf. the socio-cognitive model). It also takes an ex ante approach to anticipating the possible consequences of a given policy “before the event”. While it requires an approach to the design of assessment systems which deliver tests which are technically valid it also requires that the test is supported by systems (processes and procedures) which enable impact to be monitored and assessed and for changes to be made at a later date if necessary.

The notion of “impact by design” provides the basis for Dimension 2 and builds on Messick’s idea (1996) of achieving “validity by design as a basis for washback” in language assessment. The importance of the rational model of test development and validation with iterative cycles referred to in Chapter 1, is a necessary condition for creating construct valid tests and for the development of successful systems to support them.

At the heart of this process, the adequate specification of the focal construct is crucial for ensuring that the test is appropriate for the purpose of the assessment and its contexts of use (cf. Messick’s twin threats to validity - construct under representation
and construct irrelevant variance – Messick 1996; 252). While the technical quality of a test can be considered a necessary condition for validity and for achieving anticipated outcomes, it is not sufficient and can only provide the latent potential for validity in use; it cannot account for all aspects of an assessment system as an operational procedure. The technical features alone, therefore, do not enable the examination provider to predict what will actually happen in practice when stakeholders begin to make use the assessment in their own social and educational settings. For providers of public examinations like Cambridge ESOL, the notion of impact by design highlights the importance of designing and implementing assessment systems, which extend the design features beyond the technical validities related to the construct, and incorporate considerations explicitly related to the social and educational contexts of test use.

Figure 7.2 models the process diagrammatically; it shows that as time passes following the introduction of the new examination (on the x axis), new contexts of use continue to occur (on the y axis). These new contexts of use mean that many more users (perhaps more than was originally planned or anticipated) have acquired a stake in the examination. While this extension of “ownership” may be unavoidable (and need not necessarily be problematical), it is an example of a change process occurring within a complex dynamic systems. It can therefore be anticipated and appropriate responses to it can be managed (cf. Fullan’s 1991 notion of “change forces”).

Figure 7.2  Change over time

As this extension of “ownership” happens, there is a risk of “drift” away from the original intentions of the test developers. For example, scores may be influenced by
differences in the candidates which are not attributable to the focal construct. In other words, construct irrelevant variance creeps in as a threat to validity. The intended relationship between use of test results and the test construct may also begin to change over time due to influences in the wider educational context; the relationship between the test construct and classroom practices (at the micro level) may shift due to local preferences and approaches. The potential for negative impact is likely to increase when the original construct is no longer suitable for making the decisions which the new users require to be made. In other words, when the examination is no longer “fit for purpose” corrective action of some kind needs to be taken.

Educational systems are complex and dynamic so that linear or causal relationships between planned changes and actual outcomes cannot be predicted with any certainty. The consequences – intended and unintended - emerge after the test has been “installed” into its real-life contexts of use. As these contexts are not uniform and are constantly changing as a result of localised socio-political and other factors, consequences necessarily vary. From this perspective the overall validity of an assessment system is an emergent property resulting from a test interacting with contexts over time. It is important to recognise this in order to implement the concept of “impact by design”.

However, the properties of a complex system are critically shaped in interaction with its environment (context), and so accurate prediction is not possible. In drawing on this view, “impact by design” is not therefore about prediction. A more appropriate term might be anticipation. In working with stakeholders’ groups and key individuals, possible impacts on both micro and macro levels can be anticipated as part of the design and development process. Where negative consequences are anticipated, potential remedial actions or mitigations can be planned in advance. So for example if “construct drift” is considered to be a risk, it can be anticipated and appropriate tolerances or limits set before test revisions are required. This approach to impact by design is congruent with the concept of social impact assessment, a form of policy-oriented social research which has been developed over recent decades (see Barrow, 2000).

In summary, the key feature of the impact by design concept is that it requires the examination provider to develop a capacity to manage the desirable and the unintended consequences which result from the use of the test in its contexts. This is consistent with the cyclical and iterative approach to test development for which we have argued, and in which on-going monitoring is over time an important feature. Evidence accrued provides the basis for appropriate action to be taken (e.g. when
enough is known about what is actually going on) and this is why Dimension 3 is an essential aspect of the meta-framework.

7.4.3 Dimension 3: finding out and understanding what is going on

It is essential to know what happens when a test is introduced into its intended educational and wider social contexts of use; we have argued that this should constitute a long-term validation plan required by the impact by design concept. For the examinations board, finding out and understanding needs to be:

a) a routine preoccupation within the operational procedures;

b) problematised within impact research agenda which allows for impact-related research studies to be conducted where appropriate.

The emergentist approach noted above within Dimension 2, encourages impact researchers to develop an “impact toolkit” of methods and approaches to “finding out”; to carry out macro-analyses of large-scale aggregated data, as well as micro-analyses of views, attitudes and behaviours in local settings. The quantitative analysis of macro-level group data can capture the overall patterns, trends and growth, while the qualitative analysis of multiple single cases enables the impact researchers to monitor variability in local settings and to work with the “ecological” features of context.

Impact research methodology must re-conceptualise impact research designs and data collection methods which are suitable for investigating with the socio-cultural and socio-political issues noted above.

Major examinations board have an important role to play in conducting this kind of research. Small scale test providers and academic researchers may find it difficult to set up and manage the systems to collect and analyse the data over extended periods (e.g. due to project-based research funding which typically covers 1 to 3 years only). Most academic researchers can only take a short term view in their research, as reflected in the kind of studies reported in the literature.

Given the life cycles of the Cambridge ESOL examinations, the examinations board has the opportunity to adopt the impact by design approach and the necessary resources to develop the systems to monitor what happens over a long period (often in phases related to innovations and revision projects).
While not rejecting experimental methods entirely, the expanded model of impact looks to “real world” research paradigms to provide the “tools” which can shed light on what happens in testing contexts. Constructivist approaches to social research which include mixed methods and quasi-experimental designs have been shown to have potential in impact research (e.g. in the three cases reviewed in this thesis). Case studies are thought to be especially useful for investigating impact at the micro level and for understanding the complexities of interaction between macro-level policies and implementation in local settings. Without such methods it is difficult for examination providers to find out about and understand how the interaction of differing beliefs and attitudes can lead to consensus or to divergence and diversity. As discussed above, examinations board need to modify their routine validation procedures in order to collect, store and access the necessary data. The “toolkit” should include standard procedures for going about this and must make use of up-to-date technology to address many practical difficulties (see Chapter 6). In adopting the expanded model, greater attention should be given to the planning and resourcing for this area of validation.

7.5.2 Dimension 4: a theory of action

For an examination board the ability to change systems to improve educational outcomes or mitigate negative consequences associated with the examinations is ultimately the most important dimension of the model. Anticipating impacts and finding out what happens in practice are not enough if improvements do not occur as a result; a theory of action is therefore required to guide practice.

Examples of theory of action can be found in the literature on reform in education and school improvements, especially in the USA. Such examples provide support for the ways in which the four dimensions of the expanded model are intended to fit together in practice (see for example, Resnick and Glennan, 2002). A theory of action provides planners and practitioners with the capacity to act in social contexts, i.e. to determine what needs to be done and when/how to do it. Being prepared to change, and being able to manage change, where necessary is critical to a theory of action, and this is an essential feature of the expanded model of impact as we have already noted. The challenge for the examination provider is to “harness the forces of change” in order to get the relevant stakeholders working together to achieve better assessment outcomes.

An ability to develop realistic action plans, to engage in constant and well-managed interaction with many stakeholder groups are important features of a theory of action for the examinations board. Some of the dilemmas which arise in assessment
contexts (as discussed in Chapter 6) can only be dealt with if a wide range of stakeholders agrees to manage them in ways which they find acceptable. As Fullan (1999: xx) puts it: “Top down mandates and bottom-up energies need each other.”

The challenge in designing successful testing systems is to bridge the requirements of the local contexts (schools/classrooms) and the wider communities; for example, to be able to handle developments in formative assessment alongside the “league table” cultures and the resulting dilemma between the individualisation of assessment vs. generalisability of results.

As we have already emphasised, communication needs to be an explicit and essential part of the plan, and because successful communication is difficult, significant effort needs to be put into this activity (i.e. it needs to be worked at over the long term). Finding the appropriate metaphor or communication channel can be very important; we have shown that framework concepts (such as the CEFR incorporating its “ladder metaphor”) can be helpful in creating shared spaces for interactive communication and dialogue and clarification of ambiguity (see Jones and Saville, 2008).

### 7.5 Implications for Cambridge ESOL

In Chapter 1 we outlined the work of Cambridge ESOL and the ways in which impact was addressed in the 1990s. In this section we will look at the implications of the expanded model within the examinations board.

Milanovic and Saville’s *Maxims of Test Impact* (1996) were a first attempt to introduce the action-oriented approach to impact into the working practices of Cambridge ESOL. They were expressed as follows (Chapter 1):

Maxim 1 **PLAN**
Use a rational and explicit approach to test development

Maxim 2 **SUPPORT**
Support stakeholders in the testing process

Maxim 3 **COMMUNICATE**
Provide comprehensive, useful and transparent information

Maxim 4 **MONITOR and EVALUATE**
Collect all relevant data and analyse as required

The revised model, with the four dimensions of the meta-framework outlined in this chapter, provides an updated platform for extending this approach, building on and incorporating
where appropriate the activities implied by the original maxims. For example, Maxims 1, 2 and 3 are relevant to Dimension 2, “impact by design”, and Maxim 4 is relevant to Dimensions 3 and 4 of the revised model (monitoring and taking action).

### 7.5.1 Revising the maxims

Maxim 1 required that appropriate planning should take place using a rational and explicit approach to test development. For Cambridge examinations, this was defined as a *process model* in which Maxims 2, 3 and 4 were essential features. Test validation was seen as an integral part of the *process model* involving the ongoing collection of data to validate the examination once operational (a view consistent with Messick’s (1989) assertion that validation is the “accumulation of validity evidence” and in line with developments in validity theory during the 1990s).

Dimension 2 of the impact model now requires that the development cycle should begin with considerations of the function of outcome (i.e. the purpose of the test), including anticipating how the test should (or might) be used, how relevant and useful it is likely to be in terms of social consequences and value implications, and the possible effects it might create, including unplanned side-effects. The adoption of the *process model* of test development should ensure the provision of the *necessary conditions* for appropriate testing systems to be developed and validated.

However, it is critical to an action-oriented approach that the process should be both cyclical and iterative; knowledge and experience gained at different stages of the process must be fed back into a continuous re-evaluation of the test following each administration of it. Figure 7.3 – now the orthodox approach to test development for Cambridge ESOL - attempts to capture this process in diagrammatic form (see Saville, 2003).
The requirement for examinations boards to develop and maintain assessment systems rather than merely single tests makes the focus on the systemic issues and long-term sustainability particularly important.

The first stage of planning should, wherever possible, involve a thorough situational analysis - similar to a baseline study - which reviews the immediate and the wider educational contexts in which the test will be developed, administered and where the results will be used. This means looking at the perceived need for the test and the principal considerations and constraints which might impinge on its development and use. One of the features of international examinations is that the contexts of test use are not easily defined or easy to predict in advance. However, in principle when conducting the initial situational analysis it is necessary to identify and describe the main stakeholders and if possible their views and attitudes need to be considered in advance. This is not always possible where sponsors (e.g. government departments) set challenging timescales for delivery.

Depending on the type of test being developed those involved in the overall assessments system include students, teachers, school managers, parents, school inspectors, government agencies, commercial enterprises, and others (as we have seen in the case studies). The external expectations of what the system should be like and its acceptability to the stakeholders need to be handled effectively. Expectations of the commercial testing market and availability of other tests of a similar kind, should be taken in to account, as well as socio-economic conditions, the
prevailing educational policies of the day, and the economic/political conditions in the anticipated contexts of use. How the test might fit into current systems and the practical limitations placed on an assessment system need to be clarified (the availability of resources for test development, test administration and the reporting of results).

The original Maxim 2 required that stakeholders are appropriately supported in the testing process by the examination provider. The involvement of the stakeholders is also essential for the concept of impact by design (Dimension 2). This involvement is crucial when designing new tests, but is also a major factor in on-going validation processes and an essential element for achieving the intended impact.

As part of the Cambridge approach informed by the “impact by design” principle, renewed attempts need to be made to involve and support a wider range of stakeholders and participants who carry out important roles within the overall assessment system:

- For item writers, improved training and guidance materials should be developed to support their work, supplemented by face-to-face training sessions as necessary, to ensure adequate awareness of current theories and applications to their work (e.g. the use of a socio-cognitive model for construct definition).
- For consultants in the field increased support can be provided for conducting research or further study e.g. through provision of data, financial assistance or help in publishing findings of relevant research.
- For examiners of both speaking and writing improved support can be provided to assist them in their recruitment, induction, training, co-ordination, monitoring and evaluation. This should include enhanced online facilities to streamline these functions and improve efficiency.
- For local administrators ways of disseminating information and of collecting feedback need to be constantly improved and revised on a regular basis.
- For candidates communication needs to be constantly updated and improved, e.g. through provision of web-pages, handbooks, annual reports, book lists, “top tip” publications, etc., to ensure that relevant information can be accessed more effectively and explained if necessary.
- For parents support may be provided indirectly through the teachers of the candidates making use of the information provided to the teachers themselves. Additional information, appropriately tailored to this audience can
also be developed and made available to explain key concepts; this can help parents to support their children as language learners more effectively.

- **For teachers** communication of key concepts (e.g. construct related features of the tests) needs to be addressed as an ongoing feature of the system. Improved handbooks, DVDs, seminars etc should be provided, not only to pass on information about the tests themselves, but also to help the teachers develop their students' language skills.

- **For commercial publishers** involvement with the tests should be managed so that up-to-date information is always provided about the tests in a timely way so that publishers are aware of changes and new developments. It is important for stakeholders to have high quality published materials which adequately reflect the focal construct and lead to appropriate test preparation activities.

- **For employers and receiving institutions** improved information about test results can be provided to help them as non-specialists (as admissions tutors, personnel managers etc.) interpret and use the results of the test most appropriately.

A key element here is, not only the requirement to provide comprehensive, useful and transparent information” (essential in fulfilling aspects of Maxim 2), but also the way in which communication is carried out to ensure that it is successful. Clarification of key issues related to language assessment such as the nature of the language constructs being assessed is particularly important.

The question of transparency, and of how to communicate the meaning of language tests, test results and issues related to language assessment to test users and other stakeholders is a major challenge for all test providers. As we have noted, the CEFR has taken on the role of the unifying frame of reference for establish the meaning of levels and a shared understanding of constructs by a range of stakeholders (as discussed in Chapter 6).

Dimension 3 of the meta-framework requires that standard procedures are developed specifically designed to investigate impact-related matters. Once a test is fully operational feedback from the candidates and their teachers at schools where the test is used needs to be captured along with candidates' performance on the test; ultimately this information should be used to evaluate the test’s performance and to assess the need for revision at some point in the future. If necessary, major revisions may entail a new planning phase and a return to the beginning of the cycle.
The iterative model is a prerequisite for the test provider to keep in touch with the consequences which result from a test being used within the educational and socio-cultural systems within which it plays an important part. This approach must be a long-term endeavour involving the development or adaptation of suitable instruments to sustain the collection of the data, including appropriate research instruments to gather impact-related data of the kind described in the three case studies.

7.5.2 Implementing the revised model within Cambridge ESOL

In moving beyond the approach to impact developed in the 1990s the expanded model based on the four dimensions of the meta-framework should permit Cambridge ESOL to enhance its test development and validation activities and will involve the following:

- A renewed commitment to assessment as a potentially positive component within dynamic educational and societal processes, including assessment in support of learning;
- A focus on understanding context within educational systems better, and in particular, a multi-level approach which seeks to monitor and adapt to unpredictable interactions between the wider milieu and local contexts/individuals;
- A socio-cognitive approach to learning and assessment, including an explicit and evidence-based approach to construct definition to achieve the potential for positive impact;
- A continuing focus on the test development and validation model which allows for cyclical and iterative processes and for changes/innovations to be implemented when necessary;
- A commitment to change management which is viewed as continual improvement, e.g. represented in quality management systems such as ISO 9001:2000;
- A recognition of the varied and continually changing constituency of stakeholders for the examinations and how to involve them more effectively in the development and validation of the exams;
- A management model which anticipates potential impacts of the examinations (positive/negative) at various levels (micro/macro) and seeks to develop appropriate impact hypotheses as the basis for operational impact research (cf. impact by design);
- A commitment to developing enhanced methods for "finding out" what is going on and for communicating more effectively with stakeholders, including the use of emerging technologies;
• A continuing effect to collection of data and to conduct appropriate analyses – both routine and as part of specific instrumental projects to find out what is happening;
• A focus on mixed method approaches (quasi experimental and constructivist models of research) as the most promising paradigm for new insights into the ways in which examinations work in various educational contexts;
• A continuing effort to develop a suitable “toolkit” for making data collection more routine and easier to manage within the operational priorities of an examinations board – including improved methods for the storage and analysis of large amounts of qualitative data, such as electronic recordings (audio or video);
• A “theory of action” within the management of test development and revision processes which includes the ability to deal more effectively with change processes and the need for innovation; the theory must based on the evidence collected and on gaining well-informed understandings of what can realistically be done to make things better.

7.6 Applying the model in other contexts
As a contribution to knowledge it is hoped that the expanded model of impact and meta-framework can be applied in the future work of Cambridge Assessment outside of Cambridge ESOL and by other examinations board.

So far some aspects of the meta-framework have been successfully applied as part of the validation plan for the Asset Languages assessment system introduced in England from 2004 onwards as part of the UK government’s National Languages Strategy. In particular, the features of the English educational milieu at the time of the development, (and especially a succession of educational reforms since 1988) make Asset Languages an interesting case in which to apply the expanded model.

The English setting has been characterised by “assessment-led change” (see Chapter 2) where policy developers have exploited the effects of assessment on schools. The school curriculum became strongly centralised for the first time in England after the 1987 Education Reform Act and control of curriculum and assessment moved from LEAs to the National Curriculum Council initially, and now QCA. Independent examinations board (including Cambridge Assessment) have continued to provide public exams at 16 and 18 although under heavy regulation and the influence of the European Union (EU) and the European Qualifications Framework (EQF) has begun to grow. For a discussion of the characteristics of this educational system see Oates, 2007. He argues that it is the interaction of the factors in the system (e.g. national curriculum, framework for
qualifications, school funding and governance, inspection and accountability arrangements, etc.) which is particularly interesting and the most important feature. Against this backdrop, reform of language education was identified as a priority; amongst many reasons for poor outcomes in learning French, German and Spanish in schools low motivation of the pupils was identified as particularly important. In 2002, therefore the Department for Education and Skills (now the DCSF) announced a National Languages Strategy which was intended to reform the provision for language education and in particular to introduce an entitlement to learn a foreign language from an early age by 2010. The overarching reform objectives were to:

- improve teaching and learning of languages, including delivering an entitlement to language learning for pupils at Key Stage 2 and making the most of e-learning;
- introduce a voluntary recognition scheme to complement existing qualification frameworks and give people credit for their language skills (i.e. Asset Languages);
- increase the number of people studying languages in further and higher education and in work-based training.

To be successful, the scheme would have to engage teachers and learners successfully in a range of learning contexts and to reverse the problems associated with poor motivation.

On the basis of these objectives and the required design considerations, impact hypotheses have been formulated as part of the original design and ongoing validation plans, e.g. one of the important impact considerations is whether the scheme can have a positive motivational effect and succeed in encouraging more people to learn a wider range of languages.

In summary, the National Languages Strategy and the approach to assessment taken by the examinations board developing the Asset Languages scheme (OCR within Cambridge Assessment) have a number of features which make the expanded model developed in this thesis particularly relevant to the project:

- A macro context provided by the educational milieu and language education policies in England;
- Many “micro contexts” within schools and colleges in diverse geographical contexts (rural, urban, etc) throughout England with varying degrees of multilingualism in the local communities;
- A deliberate reform policy intended to bring about change and to introduce improvements in the teaching and learning of languages (other than English) over an extended timeframe;
• An innovative assessment system being introduced alongside the existing models of educational assessment (GCSEs and A levels) from 2003 onwards;
• A research agenda for Asset Languages based on Cambridge ESOL’s experience in test development and validation, incorporating VRIP features (including definition of focal constructs of language ability) and on-going validation plan to monitor uptake and other contextual factors (from 2006/2007 onwards).

At the time of writing some aspects of the earlier impact model had been implemented within the project, but more could now be done by the examinations board to take into account the implications of the expanded model described above; for example measure could be taken to understand the nature of the innovative processes better (Dimension 2), to monitor the contexts of test taking and test use using appropriate data collection and research methods (Dimension 3), and to develop an action-oriented approach which will enable improvements to be made as required (Dimension 4).

In adopting these measures it can be anticipated that the impact of the Asset system as a whole can be better understood within the national educational setting, and as a result, improvements can be made in future which achieve the intended outcomes of the project more effectively.

7.7 Limitations of the research

While it is hoped that this thesis makes a positive contribution to knowledge in the area of impact, and particularly in guiding examination providers in their test development and validation work, there are a number of limitations which need to be highlighted.

The first relates to the nature of the meta-analysis carried out and the potential for researcher bias. It was explicitly stated at the start of this thesis that the perspective taken would that of a UK examinations board with outcomes anticipated to improve working practices. It was suggested that this analysis could only be carried out by an “insider” with a comprehensive understanding of the examination systems. While this is seen as a strength and a main feature of this work, it also introduces limitations related to the nature of the meta-analysis and the conclusions reached. As noted in Chapter 3, hermeneutic research of this kind is prone to biases and lack of objectivity unless care is taken to address these issues in the research design; triangulation and other approaches to verification can counter threats to the validity of findings and the analysis of the three cases benefited from the prolonged involvement of the current author and from peer debriefing/checking with well-informed colleagues which helped to ensure that the analyses were plausible. However, it is in the nature of the method for grey areas to persist and for conclusions
reached to reflect the attitudes and preferences of the researcher. This has implications for the expanded model and means that the suggested applications need to be treated with caution until they can be verified in practice.

Secondly, the suggested application of the expanded model by Cambridge ESOL, while building on current practices, still needs to be implemented and the efficacy evaluated. This will inevitably take time and so has been outside the scope of the current work. The expanded model points in the direction of theories and research methodologies which may be promising in developing the impact toolkit and while there is already some evidence that this can be achieved, the benefits of the model to the examinations board have yet to be evaluated in practice. Furthermore, how well the concepts can be transposed to other organisational contexts has yet to be demonstrated. In line with the action-oriented approach advocated in the expanded model itself, further cycles are needed to address possible weaknesses and to make adjustments as appropriate.

Thirdly, the social dimension of language testing has recently become a major focus of attention and work in the field is now developing rapidly. The No Child Left Behind act in the USA has generated a vigorous debate on the relationship between educational systems and assessment practices, and in the UK, the growing dissatisfaction with the “hybrid uses” of examination results in English schools has led to a growing body of research. This work will need to be taken into account when considering changes to the expanded model in future. Fresh insights from social impact assessment – SIA - (also known as social impact analysis) are relevant to the model developed in this thesis, and in a forthcoming paper entitled Standards-based assessment in the U.S.: Social and educational impact, Chalhoub-Devile refers to SIA in advocating an approach to test impact consistent with Dimension 2 of the meta-framework outlined in this chapter. The influence of SIA on language assessment been too recent to be fully incorporated into this thesis but it is encouraging to note that the expanded model can cater for such recent developments.

7.8 Concluding remarks

In this thesis it has been argued that providers of high-stakes language examinations, have an obligation to investigate the impact of their examinations within the educational contexts in which they are used. The aim has been to propose an expanded model of impact that will enable this to happen more effectively.

The historical context of examinations within the British tradition was an important and formed part of the review in the first chapter. The starting point for the thesis was a discussion of examinations within educational processes in general, and in particular the role that examinations board, such as Cambridge ESOL, play within educational systems.
Explicitly, therefore, the perspective taken has been that of the University of Cambridge ESOL examinations board, a major provider of international language examinations; this has also reflected the author's own interests and responsibilities in developing a model of impact to guide practice within the organisation.

The thesis employed a meta-analysis to describe and review three impact projects, and based on the analysis, the main outcome and contribution to knowledge is the expanded model of impact designed to provide examination providers with a more effective "theory of action".

When applied within Cambridge ESOL it has been shown that this will allow anticipated impacts of the examinations to be monitored and to inform processes of change more effectively. It is hoped that this will lead to better motivated improvements to the examinations and the related systems in an on-going and iterative way.

Unlike research reported previously in the literature, the proposed model “builds in” impact considerations from the start, and seeks to anticipate potential effects and consequences with a commitment to monitoring and making changes as required; the suggested approach within the expanded model is referred to as impact by design and included “anticipatory impact research questions”.

It has also been shown that the model should be applicable within other examinations board and institutions which provide public examinations. How the expanded model might be applied within a major national project by a public examination provider was briefly discussed with reference to the Asset Languages project, part of the UK Government’s strategy for language learning in England.
IELTS is a **task-based testing system** which assesses the language skills candidates need to study or train in the medium of English.

All candidates must complete four Modules - Listening, Reading, Writing and Speaking to obtain an IELTS Test Report Form. Candidates are tested in Listening, Reading, Writing and Speaking. All candidates take the same Listening and Speaking Modules. There is a choice between Academic and General Training in the Reading and Writing Modules.

Total Test Time: 2 hours 45 minutes

The tests are designed to cover the full range of ability from non-user to expert user. In addition to a **band score** for overall language ability on a nine-band scale, IELTS provides a score, in the form of a **profile**, for each of the four skills (listening, reading, writing and speaking. (see IELTS website and Annual Reviews)
The Academic version tests a person’s ability to study in English at undergraduate or postgraduate level. General Training is suitable for people who are going to an English-speaking country to work or train at below undergraduate level. It is also used for immigration purposes to Canada, Australia and New Zealand.

The first component of the IELTS assesses **Listening** skills in a test lasting 30-40 minutes with 40 items in four progressively more demanding sections, the first two focusing on social needs, the second two on educational or training topics.

The academic **Reading** test (60 minutes, 40 questions) includes three non-specialist, general-interest texts, lengths totalling 1500-2500 words, taken from magazines, journals, papers, books, on issues appropriate and accessible to under- or post-graduate participants.

IELTS academic **Writing** test is a 60-minute paper requiring the production of a text of 150 words and one 250 words. Both academic writing tasks are intended for the assessment of candidates’ responses in terms of register, rhetorical organisation, style and content appropriate to topics and contexts which appear similar to those in the Academic Reading Test.

The IELTS **Speaking** test is a face-to-face oral test with a trained examiner. It assesses the candidate’s ability to communicate with other English speakers using the range of language skills necessary to study through the medium of English.

For more information see IELTS website: [www.IELTS.org](http://www.IELTS.org)
Appendix 4.2

1. The development of ELTS/IELTS

The first discussions of a new test of English for academic purposes took place in the 1970s as a replacement for the English Proficiency Test Battery (EPTB). This was a traditional, largely multiple-choice, test battery that had been used by the British Council (BC) in its overseas student recruitment since the mid 1960s for the purpose of screening international applicants to universities and colleges in the UK. As a result of this review, it was decided by BC that the EPTB should be replaced (Davies, 2008 chapter 2) and at the same time, a decision was made by the BC to invite the University of Cambridge Local Examinations Syndicate (UCLES) to form a partnership with them for the development and delivery of the new test. Cambridge therefore became involved in the testing of English for academic purposes at this time (although the Certificate of Proficiency had been in existence for many years and was generally accepted for matriculation purposes).

The partnership has continued since then and has developed in a number of ways especially with the addition of a third, Australian partner in the late 1980. Since then there have been three owners/sponsors of the testing system, entailing a division of accountabilities and responsibilities to be managed jointly. This has been a unique feature of IELTS, bringing specific benefits and challenges to be addressed. Within the partnership Cambridge ESOL is mainly responsible for the test development and validation while the other two partners, the BC and IDP IELTS Australia, are responsible for enrolling candidates and administering the test around the world.

2. The ELTS construct

The concept of communicative language teaching emerged in the ELT profession in the late 1970s and early 1980s. This reflected changes taking place in linguistics and in language education more generally (e.g. with regard to syllabus and curriculum design – see White, 1986). In language testing, especially in Britain, the communicative approach led to a focus on testing language skills in a more authentic way and with greater reference to specific contexts and purposes.

These influences fed into to the development of ELTS and to a number of other “communicative language tests” which emerged at around the same time (e.g. CUEFL - see Morrow 1977, 1979; Hawkey, 2004; Davies, 2008).
ELTS was first produced and administered in 1980 as part of the new partnership between the BC and UCLES. In addition to the new ownership arrangements, it had a range of innovative features reflecting the changes in language learning and teaching theory noted above. For example, test tasks were intended to be based on an analysis of the ways in which language is used in academic contexts (a needs-based approach) and were intended to reflect the use of language in the “real world” of academic life (in other words to have a high degree of situational authenticity and relevance). It was thought that this would achieve greater “content validity” and better prediction of success in academic study where English is the medium of instruction (i.e. improved predictive validity).

Within the BC, the Director of the Consultancies Department of the English Language Division, Munby, who had completed a PhD and a related volume on communicative syllabus design (1978), exerted a strong influence on these developments. His work emphasised the importance of designing teaching syllabuses linked to the specific needs of the learners. Munby’s communicative needs model influenced the test development committee which worked with item writing teams on the specifications and test design. B.J. Carroll, Consultant in Testing in the BC Consultancies Department, was also influential at the start and he produced a seminal report (1978) which set out the features which were envisaged for the new test. However, while specifications were produced in this process for the new test, these were never finalised and do not appear on record, as pointed out by Clapham in her book on IELTS (1996: 53-54): “no final specifications were ever produced” (see Alderson and Clapham, 1992).

This was clearly a major deficiency of ELTS which was addressed during the later developments of IELTS. Subsequent developments in the Periods 2 and 3 noted above, included a major validation study conducted in the mid-1980s, and two substantial revisions to the test itself: the first in 1989 and the second in 1995, which also led to the focus on test impact and the studies reported here.

In historical terms, the original ELTS was particularly innovative in terms of the authenticity of test content, and the fact that it offered a choice of six specific modules covering five broad areas of study of UK tertiary education, plus one non-specific area. The six modules were: Life Sciences, Social Studies, Physical Sciences, Technology, Medicine and General Academic. There was also a Non-Academic test for vocational candidates which later became the General Training Module (in period 2).

Each candidate was required to take three sections related to their own subject area and two “common tests” in the General section as follows – shown by M for subject-related modules and G for the general sections:
Another innovative feature of the test which was intended to enhance authenticity was that the three subject modules were *thematically linked* in order to create an integrated skills approach: in the Writing paper (M2) candidates were required to write on a topic connected to one of the texts in the Study Skills paper (M1). Similarly, in the Interview (M3) the candidate would be asked to discuss a topic already covered in M1.

The test was designed to be flexible in terms of *when* the test could be taken and was characterised by being “on demand” and with choices in content to meet the test taker’s needs.

According to Davies (2008), the original impact on the key stakeholders “was considerable” and Clapham (ibid) notes that ELTS was generally “well liked” by many of the users. The reporting of results on a 9 band scale with profiling by skill rapidly became well-known and the influence was considerable, especially in the UK. The 9 band system was in fact adapted by others as a way of conceptualising proficiency levels for English and used in other frameworks (e.g. it influenced the ESU Framework which was developed in the mid 1980s and used a nine-band scale – Carroll and West, 1989).

However, compared with other international tests of English (such as TOEFL), the candidate numbers remained low with 4,000 in 1981, and this only rose slowly to reach around 10,000 by 1985. From this perspective the impact was quite restricted at that time.

There are a number of possible reasons why the take up was relatively slow, including:

- the complexity of the ELTS system of modules themselves;
- the choices to be made by candidates regarding these modules;
- the complicated procedures for administering the test at centres.

From the start, test users encountered problems in making an appropriate choice of module and it became clear that there were other practical difficulties with the administration of the test relating to the number of test items and the time taken to complete the test.
3. The weaknesses of ELTS

ELTS was criticised for a number of flaws in the way it had been operationalised and for some of the technical aspects. The focal constructs certainly lacked clarity and transparency from a theoretical perspective, and this was especially apparent in the absence of fully developed and documented test specifications. It was not surprising, therefore, that soon after the test was launched, there were urgent calls for more rigorous validation studies to be carried out and suggestions for revisions to be made.

A specific proposal was made by Davies who had devised the predecessor to IELTS – the EPTB – in the 1960s. Davies’ suggestion was actively pursued by the BC and UCLES, who commissioned him to conduct a validation study starting in 1982 (see Criper and Davies, 1988; Hughes, Porter and Weir, 1988).

Clapham (op cit: 56-57) discusses the setting up of this “ELTS Validation Project” which also led to her own involvement and eventually to her own PhD research. The study involved teams in Edinburgh, Cambridge and London, as well as consultation with other academics. It lasted about four years and partly as an outcome of this study, the first major ELTS Revision Project was established in 1986. This predated the establishment of the ESOL Division in Cambridge and the project was “outsourced” under the academic direction of Charles Alderson at Lancaster University.

4. The ELTS Revision - 1986 onwards

The ELTS Revision began in August 1986 and over a period of 18 months the draft specifications for the revised test were developed by the team working with Alderson. The first drafts were circulated in January 1988 with small-scale pilots taking place that year. Wider scale trials took place in 1989 and the new test went live in the autumn of the same year when the project was handed over to Cambridge to manage as an operational test (by which time the ESOL Division had been operating for about 12 months). Some collaboration with Lancaster continued after that and in particular Caroline Clapham received support from Cambridge to conduct her research which led to her PhD (which was influential in period 3 for the 1995 revision of the test). While no longer directly involved with IELTS after 1989, Alderson and colleagues in Lancaster played an important role in Phase 1 of the impact studies reported below.

Towards the end of the 1980s, the two British partners sought to widen the international participation in the ELTS revision project, and looked for suitable partners in other English speaking countries, mainly in Australia and Canada. As a result a suitable Australian collaborator was identified in IDP - the International Development Program of Australian Universities and Colleges (IDP), now known as IDP Education Australia. This organisation
joined BC and UCLES to form the tripartite international partnership which continues to
date. One outcome of this partnership in Period 2 was the secondment of David Ingram,
Professor at Griffith University to the revision project, and this led subsequently to his
continued involvement in test development as Chief Examiner in the early 1990s. The
international collaboration was also reflected in a new name for the test: the *International*
English Language Testing System or IELTS.

The main recommendation of the revision team was to simplify and shorten ELTS and to
seek a suitable balance "between practicality and maximum predictive power" (Davies,
2008); the number of subject-specific modules was to be reduced from six to three and the
Non-Academic test was to be replaced by the General Training Module.

When the new IELTS replaced ELTS in 1989, candidates took two *non-specialised*
modules, Listening and Speaking, and only two *specialised* modules, Reading and Writing.
The non-specialised modules tested general English while the specialised modules were
intended to test skills in particular areas suited to a candidate’s chosen course of study. The
specialised reading and writing modules (still incorporating a direct link between the reading
and writing activities) were available in three discipline fields which linked together related
fields that had previously been separate modules in the ELTS battery. The reduction meant
that there were three modules, as follows:

Module A – Physical Science and Technology;
Module B – Life and Medical Sciences;
Module C – Business Studies and Social Sciences.

A noticeable impact following these changes was the *growth in candidature* which was largly attributed to the more streamlined administrative procedures. Test administrators were able to run sessions on a more regular basis and this in turn meant that greater access to test dates was provided by the centres for the candidates. This illustrates how the *practical issues* have been balanced with the other considerations within the history of IELTS – a point brought out by Davies in his historical review of IELTS.

Over the five years following the revision, the number of candidates continued to rise by
around 15% each year, and by the time of the next revision in 1995 there were over 40,000
candidates in 210 test centres around the world.

5. IELTS 1995 and the impact projects

The third period in the timeline began two years before the most recent major revision took
place in April 1995. By that time a number of important changes had taken place within the
IELTS partnership which influenced the way the project was carried out. The most important of these was the influence of the ESOL Division within UCLES, as it was known at that time. Set up in 1987, it did not play a role in the previous revision process until the final phase of the project when the new IELTS test was introduced in 1989. In that year an “Evaluation Unit” was established which meant that validation work on their English tests and examinations no longer needed to be “outsourced” to external consultants. The 1995 revision, therefore, followed the setting up of routine evaluation procedures for IELTS by Cambridge after 1989. This included the calibration of the test materials for the Reading and Listening modules through pre-testing and standards setting activities, as well as the collection and analysis of “live” test data from centres. A greater emphasis on investigating who the candidates were in relation to their demographic features and how they performed on different parts of the test formed part of this routine analysis.

A commitment had been made by the IELTS partners after 1989 to respond to developments in the field, whether in applied linguistics, measurement theory or teaching practice, and to implement further revisions as necessary. In particular, the commitment to change based on systematic collection of validation evidence was accepted within the IELTS partnership by that time and there was an intensive period of review which started in 1993. This set out to consider whether further changes needed to be made to the IELTS system in light of the evidence being collected. This review led to a second revision project and to the revised test which was launched in 1995 (see Charge and Taylor, 1997 for a discussion of the 1995 revision).

6. The IELTS research programme from 1995

Despite the 1989 simplification of ELTS, a continuing difficulty, both for the test developers and the test administrator, was the relative complexity of the subject specific modules and integrated skills approach. This was apparent during the review period and as a result the 1995 revision led to further simplification of the testing system itself. The notable changes to the test were as follows:

- The field-specific Reading and Writing Modules A, B and C were replaced with a single Academic Reading Module and a single Academic Writing Module (the research behind this change to the test design is reported in full in Clapham, 1996).
- The thematic link between the reading and writing activities was removed.
- General Training Reading and Writing Modules were brought into line with the Academic Reading and Writing Modules in terms of timing allocation, length of written responses and reporting of scores.
In addition, systems were put in place to ensure that better quality data on test performance and on the background of candidates could be collected routinely so that issues of fairness relating to test use and users could be more effectively monitored. This was a fundamental requirement when turning attention to the impact of the test as part of routine validation. The commitment to “review and revision cycles” has continued and several smaller revision projects followed the major revision in 1995. First a project to revise the IELTS Speaking Test was launched in 1998 and led to the revised IELTS Speaking Test in July 2000; secondly a development project for new assessment criteria in the Writing Test was carried out and as a result a new system for marking writing became operational in January 2005. (Details of these revisions are not reported here – see Cambridge ESOL Research Notes articles which describe the rationale and report on progress of these revisions: www.cambridgeesol.org/rs_notes).

Despite the changes over nearly 30 years since ELTS was first introduced, the current test (2008) still retains significant design features which date from the 1970s, including a task-based approach with an emphasis on the comprehension of extended texts in the Reading and Listening papers, the direct testing of spoken language performance through a face-to-face Speaking test and of writing ability through the use of the essay and report formats in the Writing test. The criterion-referenced approach and the well-known nine-band scale have also been retained, although changes have been made for rating purposes to the rating scales and there has been greater emphasis on explaining the score outcomes to users. These features are of course construct-related and therefore need to be reflected in any investigations of the intended impact of IELTS on teaching and learning (indeed this was present from the start in B. Carroll’s work).

Advocates of the “communicative approach” argued that ELTS/IELTS would have a more beneficial influence on teaching and learning when compared with traditional, discrete-point, multiple-choice tests (e.g. the TOEFL which had already been in existence for nearly 20 years). Although this is not well-documented in the literature, many professionals working in of Asia (Japan, Korea, Taiwan etc.) during that period became aware that test cramming had become a real concern in language education. Discrete-point, multiple choice tests within high-stakes examinations, such as university entrance exams, were seen by many practitioners in countries such as Japan as an obstacle to innovation and the development of more effective teaching/learning practices (Lo Castro, 1988; Saville,1990). Proponents of exams based on communicative principles suggested that the problem could be addressed through the more communicative approach because of the “better fit” between test preparation activities and desirable teaching and learning practices. It was argued that learners could develop useful language skills while at the same time preparing for the test; in other words, there would be less cramming, focusing on unproductive test
taking techniques, and a more appropriate balance of useful language learning and test preparation. From the early days, therefore, there were expectations within the ELTS/IELTS stakeholder community that there would be a potentially positive relationship between the test and teaching for it. These ideas were strongly felt by proponents of the communicative approach, such as Morrow who proposed the concept of “washback validity” (Morrow, 1979), but no explicit model of washback was developed at that time; as we saw in Chapter 2.

So while beneficial washback effects were predicted for IELTS, this remained without a precise conceptualisation of what might be involved. When the impact projects began in 1995, the construct-related features of ELTS/IELTS were, therefore, a logical basis for generating impact hypotheses. How the hypotheses emerged, and which particular ones were arrived at, provide an important lesson in conceiving and conducting impact research.

The construct-related features, which were endorsed and refined during the ELTS/IELTS revisions, were evident in the IELTS impact studies and it was significant that IELTS was the first testing system to be chosen for a major impact project by the Cambridge team. As noted, the impetus for the IELTS impact studies was the opportunity presented by 1995 revision and the commitments made at that time to continue to validate the newly revised test. By that time impact already formed part of the validation agenda for the ESOL examinations as represented by the VRIP acronym (as discussed in Chapters 1 and 2 - see also Saville, 2003). Furthermore, the rapid growth in candidature following the 1995 revision and the high-stakes associated with the test for many of those candidates (over 500,000 candidates were taking IELTS by the end of the period reported in the IELTS case) meant that the potential for unintended effects and consequences in many different contexts had risen significantly.

As part of validation programme following the 1995 revision, the Cambridge validation team decided that there would be a number of phases, each with cycles of activity, e.g. trialling, data collection, analysis, and reporting. In his book on Impact Theory and Practice, Hawkey (2006) pointed out that this type of research, by its nature will tend to be “continuous rather than monolithic and one-off”. This conclusion was certainly reinforced during the project and additional insights and clearer guidelines have emerged from the IELTS case to support this view. IELTS research activities are now co-ordinated within a framework for research and validation agreed by the three IELTS partners, and the focus on impact has gradually been incorporated more widely into the research and validation programme for IELTS since then. The Joint IELTS Research Committee (chaired by the Director of Research and Validation of Cambridge ESOL) oversees this research agenda and ensures that the three owners of the test are equally represented in deciding the research priorities. Impact has
been given increasing emphasis by the Joint Research Committee and many of the IELTS-funded research projects have focused on monitoring the effect the test has on users in different contexts and in relation to a variety of stakeholder roles. Many of the studies conducted under the IELTS-funded research programme can be categorised as washback/impact studies and cover this area of test validation. Hawkey (2006) in reporting his work as coordinator of the IELTS Impact studies from around 2000, analysed all the IELTS-funded projects with this in mind and noted that out of 44 impact-related studies “..... 15 were mainly concerned with the IELTS skill modules (reading, listening, writing, speaking), 12 with IELTS stakeholders (including candidates, examiners, receiving institutions), and 11 with IELTS preparation courses and candidates’ future target language-related needs”.

Impact research now makes an important contribution to the monitoring of IELTS and helps stakeholders to understand, use and interpret the test scores better (see Taylor, 2008 – introduction to IELTS Research Reports, Volume 8).
Appendix 4.3

The Cambridge-UCLA Language Learner Questionnaires (LLQs)

This was originally a joint study carried out between the Department of Applied Linguistics at the University of California, Los Angeles and the Cambridge ESOL in the early 1990s. It led to the LLQ bank of items being developed by Bachman, Weigle and Purpura in collaboration with the Research and Validation staff at Cambridge ESOL (see Bachman, Cushing & Purpura, 1993; Purpura, 1999; Purpura and Saville, 2001).

This project attempted to find ways of investigating the individual differences of language learners that effect performance in the second language situation and particularly in the context of studying for and taking language tests and examinations. A deeper understanding of the affective and cognitive factors influencing second language acquisition could have important implications for test construction and the interpretation of results, particularly in controlling and reducing the bias attributable to these factors, which may disadvantage certain types of learners.

The project developed a questionnaire item-bank containing items specifically designed to capture information concerning a range of affective and strategic factors. In the first stage of the project, it was important to make sure that the factors chosen for investigation were relevant to language testing and second language learning based on a clear reading and understanding of the relevant literature.

A lot of time was spent on shaping a taxonomy of factors that the study would concentrate on and several important studies which were available at that time (Anderson, 1981, 1983; Bachman, 1990; Corder, 1973; Gardner, 1979; Hunt, 1982; Rubin, 1989; Tarone, 1977) were used as starting points for the definition of the factors.

A second stage of the project was to develop the item bank for each of the areas of interest. This was done by writing questionnaire items, or adapting pre-existing ones, to capture the relevant information. The items were reviewed and edited for content and clarity to ensure they were targeted at the factor under investigation and that the language was simple enough to be understood by English language learners (although for low level learners it was necessary to translate them into specific L1s).

Pilot questionnaires were constructed and administered to a small number of EFL students at Santa Monica College, California. The students were undergraduates enrolled on ESL courses. They were asked to read through the questionnaires and mark
the items they found confusing or difficult to answer, and to explain the difficulty. In addition, they were asked to audio-tape their reactions to the attitude, motivation and anxiety questionnaire. Ten of the students were interviewed about their reaction to the cognitive and meta-cognitive questionnaires. On the cognitive strategies questionnaire, students were asked to indicate how often they used each of the strategies described and to indicate by way of highlighting a light-bulb symbol if they had never used the strategy but thought it might be a good strategy to use in the future.

The responses to the items were analysed and the qualitative information provided by the students' reactions to the items was collated. The qualitative and quantitative information this exercise provided was used to review the items a second time. A number of items were dropped or re-written and new versions of the questionnaires were produced.

After this stage there were five questionnaires containing items grouped in the following way:

- Questionnaire 1: Attitudes and Anxiety (72 items)
- Questionnaire 2: Motivation and Effort (70 items)
- Questionnaire 3: Cognitive Strategies (71 items)
- Questionnaire 4: Metacognitive Strategies (68 items)
- Questionnaire 5: Communication Strategies (50 items)

The third stage of the project was to run a final pilot of the questionnaire items to ensure they functioned as they had been designed to do and to ensure that the information they captured was usable and relevant to discerning the affective and cognitive strategies language learners bring with them to examinations and the classroom. This phase of the study involved a number of local teachers and their EFL students. The teachers were given the questionnaires to present in the classroom. The students were asked to respond to them, but the content also formed a useful departure point for discussion on study skills and learning styles.

The responses of the students were analysed in Cambridge so that the reliability of the items could be assessed a third and final editing session was required before the items were formally entered into the item bank.

This project was intended to provide a battery of questionnaires designed to measure the background characteristics of the Cambridge candidature and to examine how these characteristics might impact on the performance of the candidates taking the Cambridge examinations when linked to test performance data at the item level. The early paper and
pencil versions of the LLQ questionnaires played a role in the IELTS Impact projects Project and later the use of more fully developed and validated versions also played a small role in PL2000 (see Chapters 4 and 5). The questionnaires in these cases were effectively used as supplements to the standard Cambridge Candidate Information Sheets to collect additional background information about some groups of test taker. The data was useful although limited.

The original bank of language learning items concentrated on the selected socio-psychological factors and strategic factors, drawing on the research of the 1980s and 1990s noted above. The socio-psychological questionnaires were designed to measure attitudes, anxiety, motivation and effort, while the strategic factor questionnaires aimed to measure cognitive strategies, metacognitive strategies and communication strategies (although the items to investigate communication strategies were never fully developed and validated for use in the LLQ).

The socio-psychological factor questionnaires relied heavily on the work of Gardner (1979) and the UCLA-based researchers designed questionnaires which were similar to his Attitude/Motivation Index (AMI) but with numerous modifications to the scales and the items. The development of the socio-psychological factor questionnaires was also influenced by theories of attitudes and motivation in cognitive and educational psychology, such as Eiser (1986), Hovland & Rosenberg (1960) and Brecker (1984).

Attitudes were viewed in terms of a three-component model involving: affect, or evaluative statements of feelings and preferences; cognition, or statements of opinions and beliefs; and behaviour, or statements of overt action or intent (conative response). Each questionnaire which was developed contained questions from these three perspectives.

The design of the attitudes questionnaire was also influenced by an understanding of attributional theories of attitude which had been used to explain instances of relative success or failure on a task in terms of the degree to which learners perceived the task as difficult. Research had shown that success occurs when the task is perceived as easy, and failure can be attributed to unreasonable task demands (McCombs, 1991; McClelland et al., 1953; Weiner, 1979). Perception of task difficulty was therefore included as one of the scales in the attitudes questionnaire and the other two scales were attitudes toward speakers of the target language and attitudes towards learning a world language.

The design of the motivation questionnaire was again based on Gardner; in this case, his instrumental motivation and integrative motivation scales were adapted for use with the Cambridge candidature. The development of the motivation questionnaire was also
influenced by Weiner's (1979) *attributional theory of motivation*, which viewed "effort" as having motivational consequences resulting in success. According to Weiner (1979), success is a result of hard work, while failure is due to a lack of effort. Those who believe they have some control over their success seem to exert more effort in pursuit of their objectives.

Since *achievement motivation* and *effort* are potentially important factors in language learning these scales had been included in the questionnaires. "Achievement motivation" refers to beliefs and opinions about one's ability to achieve, while "effort" refers the concrete actions a learner is willing to achieve, such as "trying hard to learn new vocabulary."

The final socio-psychological factor questionnaire involved anxiety, an affective condition which may undermine language learning or performance on a task and particularly relevant to formal test taking contexts which are known to increase anxiety in most people.

Gardner’s AMI had defined anxiety in terms of the language classroom (i.e. class anxiety), but it was recognised that Cambridge learner/test takers may also experience anxiety associated with using the language outside the classroom (in real-world communicative situations) where a lack of adequate linguistic resources in some context may lead to anxiety). Also, of course, Cambridge candidates may experience anxiety when taking the examinations. As a result of these considerations, the anxiety questionnaire was designed to measure three types of anxiety: *language class anxiety*, *language anxiety* and *test anxiety*.

The development of the strategic factors questionnaire originated in Gagné, Yekovich and Yekovich's (1993) *model of human information processing* and the development of the cognitive and metacognitive strategy questionnaires have been particularly well-documented by Purpura (based on his PhD studies which were supported by Cambridge ESOL). The Cognitive Strategy Questionnaire was designed to measure a number of comprehending, memory and retrieval strategies, while the Metacognitive Strategy Questionnaire was designed to measure assessing the situation, monitoring, evaluating and self-testing.
### Socio-psychological and strategic constructs

<table>
<thead>
<tr>
<th>SOCIO-PSYCHOLOGICAL FACTORS</th>
<th>STRATEGIC FACTORS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A. Attitudes</strong></td>
<td><strong>A. Cognitive Factors</strong></td>
</tr>
<tr>
<td>1. Attitudes towards Learning English</td>
<td>1. Selecting Strategies</td>
</tr>
<tr>
<td>2. Attitudes towards English speakers</td>
<td>2. Comprehending Strategies</td>
</tr>
<tr>
<td>3. Interest in Foreign Languages</td>
<td>3. Storing and Memory Strategies</td>
</tr>
<tr>
<td>4. Perception of Task Difficulty</td>
<td>4. Retrieval or Using Strategies</td>
</tr>
<tr>
<td><strong>B. Motivation</strong></td>
<td><strong>B. Metacognitive Strategies</strong></td>
</tr>
<tr>
<td>1. Integrative Motivation</td>
<td>1. Goal-setting</td>
</tr>
<tr>
<td>2. Instrumental Motivation</td>
<td>2. Assessment</td>
</tr>
<tr>
<td>3. Achievement Motivation</td>
<td>3. Planning</td>
</tr>
<tr>
<td><strong>C. Effort</strong></td>
<td><strong>C. Communication Strategies</strong></td>
</tr>
<tr>
<td>1. Message Manipulation</td>
<td>1. Message Manipulation</td>
</tr>
<tr>
<td>2. Form Manipulation</td>
<td>2. Form Manipulation</td>
</tr>
<tr>
<td>1. Class Anxiety</td>
<td>1. Message Manipulation</td>
</tr>
<tr>
<td>2. Language Anxiety</td>
<td>2. Form Manipulation</td>
</tr>
<tr>
<td>3. Test Anxiety</td>
<td>2. Form Manipulation</td>
</tr>
</tbody>
</table>

In summary, the LLQ bank was theoretically based and well researched and the development of the questionnaires themselves followed best practice procedures in instrumentation validation, (as discussed and exemplified in Chapter 4 in relation to the IELTS Impact Studies). An iterative process was followed from the start: pilot questionnaires were submitted to a series of analyses including reliability analysis, which allowed the researchers to reduce the number of items and increase the homogeneity of the scales. In some cases, scales were dropped and others combined. Following this, several analyses were performed on the socio-psychological factor questionnaires and on the strategic factor questionnaires. This resulted in a number of changes to the original questionnaires. In follow up validation studies of the questionnaires the following research questions were also addressed by Purpura using both qualitative and qualitative techniques (e.g. what was the factorial structure of the questionnaires?, how did the items rank as indicators of the underlying questionnaire constructs? how could the questionnaires be shortened and streamlined for use?). As a final stage, following Byrne (1994), the questionnaires were submitted to item-level confirmatory factor analyses using structural equation modelling (SEM) to identify the items that best measured the underlying constructs. Poorer items that loaded on more than one construct and items that produced significant correlated errors were removed, and the remaining items could be ranked. Following this research and development process, the bank of validated questionnaire items was available for use by the late 1990s.
The availability of the LLQ bank and the four validated questionnaires provided a resource for use in research and validation studies. Some success was achieved in using the questionnaires in one-off research projects (such as Purpura's PhD). However, it was recognised early on during the validation phases that it would be difficult to use such instruments in routine validation, and that the data could not be collected without significant effort and inconvenience on the part of the respondents.

The practical problems of administering relatively “unfriendly” questionnaires in busy classrooms had also been recognised and in a subsequent development phase starting in 1999/2000, a project was set up to develop a shorter, computer-based (electronic) version of the LLQ. It was intended that the streamlined electronic questionnaires could be administered more effectively to larger numbers of people and a new multilingual feature was to be added allowing easier access by lower level learners, (i.e. enabling learners to choose which language they wished to answer about - the target language - and to select their own language in which to answer the questions). For example, Italian students in English classes could answer questions in Italian about their experiences of learning English (thus removing the hurdle of understanding and answering questions in the target language).

The first version of the computerised version consisted of a composite questionnaire covering the language learning factors described above. It was optimised to be as short and efficient as possible with 60 questions that could be comfortably answered in about 35 minutes. Every item in the questionnaire was selected from the best functioning items (based on the validation studies) inked to the underlying traits. This computerised product was intended to fulfil a dual purpose: on the one hand, to provide learners and their teachers with an easily accessible pedagogic tool which could be made available in classrooms and self-access centres; on the other to provide the Cambridge researchers with an instrument for wide-scale data collection on learner variables in a range of projects including impact studies.
Appendices to Chapter 5

Appendix 5.1
The Italian educational system

In order to contextualise the PL2000 project and the Impact Study within the state system in Italy, the key features are set out below to allow for comparison with other national systems. The conventions of the International Standard Classification of Education (ISCED) framework are adopted here to describe the structure of the Italian educational system (ordinamento scolastico). ISCED was designed by UNESCO in the 1970s and now provides one of the most widely used frameworks used in tender documents in the field of education (e.g. by the European Commission). It was conceived as ‘an instrument suitable for assembling, compiling and presenting statistics of education both within individual countries and internationally’ and was approved by the International Conference on Education (Geneva, 1975), and subsequently endorsed by UNESCO’s General Conference. The present classification (ISCED 1997) was approved by the UNESCO General Conference in November 1997 and therefore can be usefully applied to the Italian system at the time of the PL2000 project in the late 1990s.

The ISCED level system

<table>
<thead>
<tr>
<th>ISCED Level</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Pre-Primary Education</td>
</tr>
<tr>
<td>Level 1</td>
<td>Primary Education or First Stage of Basic Education</td>
</tr>
<tr>
<td>Level 2</td>
<td>Lower Secondary or Second Stage of Basic Education</td>
</tr>
<tr>
<td>Level 3</td>
<td>(Upper) Secondary Education</td>
</tr>
<tr>
<td>Level 4</td>
<td>Post-Secondary Non-Tertiary Education</td>
</tr>
<tr>
<td>Level 5</td>
<td>First Stage of Tertiary Education (Not leading directly to an advanced research qualification)</td>
</tr>
<tr>
<td>Level 6</td>
<td>Second Stage of Tertiary Education (Leading to an advanced research qualification)</td>
</tr>
</tbody>
</table>

The education system in Italy during the time of PL2000 can be represented as follows showing ages and levels:
It can be seen from this table that the Italian school system is divided into three broad cycles like many other countries, i.e., primary, secondary, and tertiary. Much of the controversy and focus for reform in recent years has been to do with sub-divisions of the secondary cycle. The school reforms which began in 1950s and early 1960s are still continuing to date. Traditionally the two main cycles of school-based education in Italy are known as the *scuola elementare* and the *scuola media*, the latter being divided into *media inferiore* and the *media superiore*. The *media* lasts 3 years, and involves an examination at the end of the third year; the *superiore* lasts from 2 to 5 years depending on the school type.

Recently there were attempts to rename the cycles of the *ordinamento scolastico* to bring them more in line with other international models. The proposed new names introduced by Minister Moratti in 2003 (law 53/2003) were: *scuola primaria* (ex elementare), *scuola secondaria di primo grado* (ex media) and *scuola secondaria di secondo grado* (ex superiore). At the time of the PL2000 impact study the traditional names were being used and these were used in Chapter 5.

Compulsory education starts in primary school and now continues until the second year of upper secondary school (age 16). The raising of the school leaving from 14 was one aspect
of the changes brought in two phases – first to 15 and then to 16 – starting under Minister Berlinguer (making the additional two years an obligation rather than a right for those who wanted to take up the option – see below for more details on the Berlinguer reform). The primary school can also be preceded by (up to) 3 years of non-compulsory nursery school (scuola dell’infanzia). In addition to state provision, there are also private schools at all levels of education. Funding for these schools is primarily from private organizations; however, private schools may also receive some state funds if they follow the same guidelines as state public schools in terms of curriculum, personnel, and management.

In the primary school, which lasts 5 years, the curriculum is uniform for all children, although parents can choose a private or a state-funded school. Subjects studied are the same in both, except in special schools (e.g. for the sight and hearing-impaired, and other children with learning difficulties). The types of upper secondary school (superiore) vary more than in some other European countries (e.g. England and Wales) and they are differentiated by subjects and activities. The main division is between three main types - the Liceo, the Istituto Tecnico and the Istituto Professionale.

Any kind of secondary school that lasts 5 years grants access to the final examination, called Esame di Stato conclusivo del corso di studio di Istruzione Secondaria Superiore (final state examination at the end of secondary school) or more traditionally, the Esame di Maturità. This examination takes place every year in June and July and grants access to any faculty at any University (although there may be other selection criteria applied by the universities for access to some courses such as Medicine). This examination has also been a main focus of discussion and reform in recent years.

Secondary education in the Liceo (“grammar school”) involves a broad, specifically academic curriculum and is generally considered necessary for University preparation (typically at age 19). There are several different types of liceo, as follows:

- **Liceo Classico**, which features Latin, Ancient Greek, Italian, History and Philosophy as its most important subjects.
- **Liceo Scientifico**, where the emphasis is more on scientific and mathematical topics, such as geometry, calculus, chemistry or physics - although Latin, Philosophy and literature are also taught.
- **Liceo Linguistico** puts emphasis on language learning.
- **Liceo Artistico** which is oriented toward arts teaching.

The Liceo Scientifico and the Liceo Classico are generally considered by stakeholders (parents, teachers, the press etc) to be the most prestigious in Italy due to the focus on the
traditional academic subjects and the basis for university careers in prestigious disciplines (law, medicine, engineering etc).

The technical Institutes called in Italian ITIS (Istituto Tecnico Industriale Statale) and ITC (Istituto Tecnico Commerciale), are oriented towards practical subjects, such as aeronautics, business administration, computer science, etc. The Istituto Professionale also offers a form of secondary education oriented towards vocational subjects which enable students to start searching for a job as soon as they have completed their studies. The focus is even more specific in terms of vocational course offerings than the Istituto Tecnico.

At the time of the PL2000 nearly 900,000 teachers were employed in the system (approximately 100,000 teachers in private schools in comparison to 760,000 in the public schools, although the latter figure does not include thousands of other teachers without a permanent contract (source: Ministry of Education website).
Appendix 5.2

Educational Reforms in the 1990s

The political back drop and educational reform programmes of the 1990s provide the macro context or milieu for the PL2000 and the impact study. As in most European countries, educational reform in the 20th century in Italy was strongly linked to political movements.

Historically, the most important educational reform in Italy dates back to 1925 during the Mussolini era under Minister Gentile, and in essence many elements of the system have remained in place until the present day. The teaching programme for elementary schools was modified in 1955 and again in 1985, and the middle school curriculum was also changed twice in 1963 and 1979. The upper school curricula and reform of the “cicli scolastici” remained under discussion up to the 1990s and the far reaching reforms of Minister Berlinguer from 1997 onwards provided the context for the PL2000.

The political landscape of Italy changed in the 1990s after the end of the Cold War and the educational reforms of the period need to be contextualised within the wider politics of the time, especially the emerging power struggles between the centre-left and centre-right coalitions. This was very different from the prevailing left-right politics of the post-war period until the late 1980s (see Caciagli and Zuckerman (eds), 2002: Hine and Vassallo (eds.), 1999).

The four Italian Ministers of Education - 1996 to 2008

<table>
<thead>
<tr>
<th>Minister of Education</th>
<th>Term</th>
<th>Prime Minister</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luigi Berlinguer (DS)</td>
<td>17 May 1996 - 21 October 1998</td>
<td>Prodi – first government</td>
</tr>
<tr>
<td>Luigi Berlinguer (DS)</td>
<td>21 October 1998 - 22 December 1999</td>
<td>D'Alema – first government</td>
</tr>
<tr>
<td>Luigi Berlinguer (DS)</td>
<td>22 December 1999 – 25 April 2000</td>
<td>D'Alema – second government</td>
</tr>
<tr>
<td>Tullio De Mauro (tecnico – non elected)</td>
<td>25 April 2000 - 11 June 2001</td>
<td>Amato – second government</td>
</tr>
<tr>
<td>Letizia Moratti (FI)</td>
<td>11 June 2001 - 23 April 2005</td>
<td>Berlusconi – second government</td>
</tr>
<tr>
<td>Letizia Moratti (FI)</td>
<td>23 April 2005 - 17 May 2006</td>
<td>Berlusconi – third government</td>
</tr>
<tr>
<td>Giuseppe Fioroni (Margherita)</td>
<td>17 May 2006 -2008</td>
<td>Prodi - second government</td>
</tr>
</tbody>
</table>

The Ministry of Education is known as the Ministero della Pubblica Istruzione – MPI. For a period of about 5 years it was renamed Ministero dell Istruzione, dell’Università e della Ricerca - MIUR (from 2001) but it reverted to its previous name in 2006.
In 1995 Romano Prodi became leader of the centre-left l’Ulivo (Olive Tree coalition), and in the 1996 Italian general election, he defeated Silvio Berlusconi and his centre-right coalition (Forza Italia). This led to his nomination as President of the Council of Ministers, (as the Prime Minister is called in Italy) and the formation of the first Prodi government. Although the Prodi government fell in 1998 (and Prodi went on to head the European Commission in Brussels), the formation of a new centre-left government led by Massimo D’Alema until 2000 meant that there was a five year period of continuity in the area of education. The educational reforms within which the PL2000 took place were guided by Minister Berlinguer and his successor De Mauro (until 2001), both of whom actively supported the PL2000 project within government more broadly.

In 2001 (and in the middle of the three year period of the PL2000 itself), the centre-right government of Berlusconi (Forza Italia) came back to power and Letizia Moratti took over as Minister of Education. Although this heralded a change of policy and halted some of the Berlinguer reforms (e.g. reform of the school cycles), the PL2000 project continued until the end of its planned pilot phase. Some other aspects of the innovations also continued throughout the Berlusconi period and into the second Prodi government from 2006. This element of longevity and acceptance across the political divide was an important feature of the project’s success.

Attempts to reform of the high school curriculum (the last 5 years of schooling) had been made in the late 1980s and early 1990s (e.g. the Programmi Brocca, 1988; also Progetto ’92), but it was not until Berlinguer became Minister in the Prodi government that major change seemed likely to happen. There were four main planks to Berliguer’s reform proposal:

i). Reform of the school cycles (il riordino dei cicli).

ii). Autonomy within the school system (l’autonomia scolastica).

iii). Reform of the school leaving examination (la riforma dell’esame di maturità).

iv). Raising of the school leaving age (l’innalzamento dell’obbligo scolastico).

The relevant ministerial documents from the period are listed in the bibliography. Early after coming to office, in January 1997 he published his first document on the reform of the school cycles (Documento di discussione sulla riforma dei cicli di istruzione) which itself has been inspired by a circular entitled Prospettive europee per il sistema formativo italiano which had come out the pervious September. This was presented to parliament in June 1997 (Legge Quadro in materia di Riordino dei Cicli dell’Istruzione). This area has, however continued to be controversial as was halted by Minister Moratti during her term of office.
The reform of the school leaving examination came into force in December 1997, i.e. *Disposizioni per la riforma degli esami di Stato conclusivi dei corsi di studio di istruzione secondaria superiore* (published in the Gazzetta Ufficiale n. 289 on 12/12/1997). This also has continued to be a focus of attention of the subsequent governments.

The law on “autonomia” within the school system (Law 59/97, article 21), was passed in 1997, but the relevant aspects of this policy only came into force in the scholastic year starting September 2000 (the details having been set out in ministerial decrees and circulars after the law had been passed).

This aspect has been perhaps the most far reaching reform in terms of the overall organisation and management of the school system and has left a long lasting legacy; it led to decentralization and devolution of powers from the central Ministry in Rome to regional ministries of education and it gave more power to schools (of all kinds) to plan curriculum and deploy resources (including control of school budgets).

A number of key areas of responsibility have still remained at national level and come directly under the central Ministry. These constitute a national framework for learning and teaching, including curriculum goals, as well as national targets for the number of hours to be taught in core subjects and the monitoring of educational quality (through central and regional authorities under its control).

For the curriculum within this approach, it was envisaged that each school would produce its own school-based curriculum based on the national guidelines (known as a ‘piano d’istituto’ or ‘piano dell’offerta formativa’ – abbreviated to POF). This allowed for about 20% of the curriculum to be decided at a local level. In terms of the school timetable, the only condition was that the school should open for at least 5 days a week and respect the annual number of hours for the obligatory core subjects. For the rest, all types of school were free to set the length of the class hours according to local needs (“nel modo più adeguato al tipo di studi e ai ritmi di apprendimento degli alunni”) and to schedule additional subjects outside the core depending on the availability of teachers in those subject areas.

As far as research and professional development was concerned, the schools were also free to initiate projects for staff training and for innovation in teaching practices, with ministerial encouragement to form professional networks between schools at local and regional levels. The research dimension was less well articulated and often referred to as “monitoraggio” (monitoring) rather than as scientific studies.
In summary, the aim of the Berlinguer Reforms was to create the concept of a “scholastic community” (*communità scolastica*) at a local level based around the school. This covered a number of related concepts of community – community of practice, community of dialogue, community of diversity and community of people (as set out in the law and related decrees). The PL2000 fitted well into these developments, coming into practice as it did at the same time as the overall changes related to increased autonomy (*autonomia*). While a change of government in 2001 (from centre-left to centre-right) could have derailed the project, if fact, the main aims seemed to coincide with some key priorities of Prime Minister Berlusconi in relation to the learning of English. On coming to power, he famously emphasised the need to boost skills in the three *Is* in order drive the economy forward, namely *Inglese* (English), *Impresa* (Business) and *Informatica* (IT). While some of the other elements of the Berlinguer reform programme were halted but the focus on English continued.

Documents retrieved from the Italian Ministry of Education (Ministero della Pubblica Istruzione) website: [www.istruzione.it](http://www.istruzione.it)

1992: *Progetto ’92* reorganisation of the *istruzione professionale*.
1996: *Prospettive europee per il sistema formativo italiano* (Reform of the school cycles).
1997: Autonomia – Law 59, article 21 (law on local autonomy of schools).
1997: Law 440. *Istituzione del Fondo per l'arricchimento e l'ampliamento dell'offerta formativa e per gli interventi perequativi* (Funding provision related to PL2000).
1998: Circolare Ministeriale n. 304. *Introduzione dell'insegnamento non curricolare e facoltativo di una seconda lingua comunitaria nella scuola media*. (Optional second foreign language introduced into middle schools).

Progetto Lingue: [www.progettolingue.net](http://www.progettolingue.net)
Autonomy: [www.istruzione.it/autonomia/progetti](http://www.istruzione.it/autonomia/progetti)
Ministry in Lombardia: [www.istruzione.lombardia.it](http://www.istruzione.lombardia.it)
Appendix 5.3

Foreign Language Study in Italy in the 1990s

Over the past 50 years foreign languages have been widely studied in the Italian educational system (Pontani, 1985; Freddi et al, 1986). Under the legislation in force at the time the PL2000 was conceived, foreign language study was available from age 8 and finished at the end of the high school, at the age of 19. In theory, therefore, the students could receive continuous foreign language tuition for 11 years. In fact, very often a second language was also introduced at the age of 11 and 14 (and the extension of this initiative became a major focus of reform in the 1990s).

The Italian schools typically had a school year of 34 weeks working on a six day a week basis (8.30 to 13.30 with activities in the afternoon on some days). The language course would typically take up 3 classroom hours per week (or approximately 100 hours over the whole year) within this curriculum. While this paints a potentially rosy picture (an adequate number of hours, continuity of study and availability of several languages within the curriculum), the reality was often very different. By the 1990s many problems with the implementation of the policy were being discussed in all circles of society and there were major concerns in government over learning outcomes and value for money (i.e. outcome success in language competence compared with the effort and investment being put into the system). This topic, however, was not taken in isolation but formed part of a wider debate on how the educational system should be reformed. In particular the discontinuity across cycles (e.g. primary to middle school) and lack of coherence of the provision between schools of different types had been a long standing concern (as reported by Ministry officials in the context of PL2000 – e.g. Sanzo, Barcelona: 2001).

In fact, the origins of the PL2000 can be traced back to innovative trends within the Ministry of Education in the late 1980s and early1990s, and in particular, the attempts by the politicians and officials to deal with the failure of the education system to deliver adequate learning outcomes in foreign languages (at least as perceived by the political classes and the critical Italian press).

What began to emerge in particular was a strong European agenda within the government and popular support for the EU’s policies on open borders and mobility of the work force. Associated with this was a realisation that Italy might fall behind other countries if changes were not made and radical improvements brought about in both structures and content.
These points are illustrated by the Ministerial documents and speeches which are referred to below.

This need to be competitive internationally was perhaps felt more acutely in some Italian regions than others; for example, Lombardy and the area around Milan aspired to be amongst the leading economic regions within the EU not just within Italy.

The EU’s White Paper *Teaching and Learning: Towards the Learning Society* (European Commission, 1995) had identified lack of success in foreign language learning as an obstacle to mobility for students and workers within the Union, and first introduced the ambitious “mother tongue plus two” objective for language learning throughout the Europe (see also EU documentation on the Lisbon Agenda 2000 and the Barcelona Indicators available from the website of the EU Commission).

The Italian Ministry officials also took notice of the unflattering comparison between language learning outcomes in Italy compared with other EU countries which had emerged in various European-wide studies (Eurostat data). The Minister of Education himself, in announcing the project in 2000, stated that “it is no longer acceptable to Italian society for our students to be so strongly penalised in comparison with their peers elsewhere in Europe.” (quoted in translation by Duguid, 2001: 55).

Ispettore Raffaele Sanzo was a key member of a Ministry team who had been involved in many English language projects throughout the 1980s and 1990s (i.e. the so called *progetti speciali*) and was the leading figure within the small PL2000 team at the Ministry in Rome. In his paper about the PL2000 which he presented at the first International ALTE Conference in Barcelona, Summer 2001 (see ALTE website, also SiLT 16, Weir and Milanovic, 2004), he identified many issues which were thought to be problematic and collectively responsible for the poor outcomes and lack of accountability in the Italian system. These issues covered structural, methodological and content factors, including the following:

- wide diversity in the characteristics and background of the learners themselves;
- variability in the professional qualities of the teachers;
- “traditional” approaches to language teaching with little or no focus on developing communication skills;
- lack of continuity across school years and cycles (e.g. elementary to middle school);
• choice of which foreign language to study leading to lack of continuity (e.g. across cycles and when children moved to new schools)
• large classes and random formation of groups (leading to mixed abilities studying together);
• arbitrariness of teaching programmes and syllabuses which were often set by individual teachers (in the absence of a national syllabus);
• uncertainty in setting the learning objectives and wide variation based on geographical position and/or the individual school;
• unsuitability and imprecision in assessing outcomes (e.g. the proficiency level achieved);
• poor exploitation of the new technologies which were already available.

Of particular relevance was the lack of standardisation of assessment (both content and level), which he pointed to, and an educational culture with no history of external examinations and certification of the kind provided by the examinations board in the UK for example.

The PL2000 was an explicit attempt to address these problems in a programmatic way, and most importantly for impact research which was carried out in the PL2000 impact study, the Ministry’s own stated objectives provided the basis for formulating impact hypotheses.
Appendix 5.4

Ministerial circular - 1998

Text (in the original version but without appendices) of the Ministerial circular for the introduction of a second optional, non-curricular European language in middle schools – 1998.

Many of the features of the project were taken up in an expanded form for the Progetto Lingue 2000.

Introduzione dell'insegnamento non curricolare e facoltativo di una seconda lingua comunitaria nella scuola media

Circolare Ministeriale 10 luglio 1998, n. 304. Prot. n. 9556

Premessa

La legge 18 dicembre 1997 n. 440, nell'istituire il "Fondo per l'arricchimento e l'ampliamento dell'offerta formativa e per gli interventi perequativi", ha tra l'altro previsto l'attivazione dell'insegnamento di una seconda lingua comunitaria nella scuola media.

Al perseguimento di tale obiettivo, sulla base anche del parere delle competenti commissioni della Camera e del Senato, è stata finalizzata la somma di 33 miliardi di lire del fondo complessivo previsto dalla legge, ripartita tra i Provveditorati agli Studi (vedi Allegato A). L'introduzione di una seconda lingua comunitaria offre l'opportunità di ripensare l'insegnamento delle lingue focalizzando abilità da privilegiare e definendo capacità/competenze con precisi descrittori. Si tratta di delineare e di proporre una serie di segmenti di insegnamento-apprendimento (moduli), unitari ma tra loro correlati, e di indicare strumenti e modalità per la valutazione oggettiva dei risultati.

Un'offerta di insegnamento linguistico di 240 ore, strutturato in moduli orari flessibili - articolato, di norma, su tre anni - e destinato a gruppi di alunni, anche provenienti da classi diverse, appare funzionale al raggiungimento di una competenza adeguata e sufficientemente corretta nelle abilità audio orali e nell'abilità di lettura, mentre all'abilità di scrittura viene riservato un ruolo strumentale rispetto alle altre.

In attesa dell'emanazione dei regolamenti attuativi dell'art. 21 della legge 15 marzo 1997 n. 59, che dovranno realizzare la piena autonomia organizzativa e didattica delle istituzioni scolastiche, l'introduzione di una seconda lingua comunitaria rappresenta, dunque, nel contesto dell'ampliamento dell'offerta formativa, l'opportunità di avviare un progetto di insegnamento / apprendimento con caratteristiche di forte innovazione. L'iniziativa si realizza tramite:
un insegnamento non curricolare, aggiuntivo e facoltativo rivolto a gruppi costituiti, di norma, da 15 alunni, con un minimo di 12 e un massimo di 20 in caso di particolari esigenze organizzative;

• un monte-ore complessivo di 240 ore articolato, di norma, su tre anni scolastici;

• obiettivi di tipo comunicativo con particolare riguardo alle abilità audio orali e alla comprensione scritta, funzionali alla comunicazione essenziale;

• soluzioni organizzative che permettano la scelta di percorsi differenziati e flessibili (multi-modularità);

• il ricorso alle nuove tecnologie didattiche;

• l'organizzazione di esperienze di autoapprendimento guidato;

• la possibilità di attestazione del livello di competenza raggiunto dagli alunni ai sensi della C.M. 335/97 e/o di certificazione dei crediti formativi con forme e modalità innovative definite dalle stesse scuole;

• la possibilità di certificazioni internazionali rilasciate da Enti certificatori riconosciuti dai Paesi Europei sulla base di Protocolli di intesa stipulati dall'Amministrazione.

Nel quadro dell'autonomia didattica e organizzativa e della loro progettualità complessiva, le scuole che intendono attivare l'insegnamento di una seconda lingua comunitaria per l'anno scolastico 1998/99 formulano agli Uffici scolastici provinciali la richiesta di finanziamento e predispongono le condizioni strutturali per il funzionamento dei corsi.

Linee guida per le scuole ai fini dell'attivazione del progetto saranno precise 1000 te in una fase successiva. Le scuole possono comunque presentare un proprio progetto che preveda percorsi diversi e che i Provveditori agli Studi valutano nell'ambito del piano degli interventi.

Resta salva la possibilità per le singole istituzioni scolastiche di attivare corsi facoltativi di lingua straniera, anche con il contributo di soggetti esterni.

1. Adempimenti degli Uffici Scolastici provinciali

I Provveditori agli Studi, acquisite le richieste delle istituzioni scolastiche, definiscono il piano complessivo dei corsi da finanziare, tenuto conto delle somme assegnate e di eventuali specifiche vocazioni del territorio, entro il 20 ottobre 1998.

Nell'ambito del finanziamento assegnato a ciascun Ufficio scolastico provinciale con la lettera circolare protocollo n. 27814 del 19 maggio 1998, sarà riconosciuta la precedenza alle iniziative che prevedono il concorso finanziario degli enti locali (ad es. convenzioni per potenziare centri di autoapprendimento, accordi di programma territoriali, ecc.) e a quelle deliberate da reti di scuole (ad es. scuole di un distretto che si coordinano per
offrire corsi che prevedano l'offerta di più lingue comunitarie, scuole che organizzano scambi e utilizzo di materiali didattici, ecc.).

Successivamente saranno tenute presenti prioritariamente le richieste delle scuole nelle quali:

- non sono in atto sperimentazioni di seconda lingua straniera autorizzate ai sensi dell'art. 278 del decreto legislativo n. 297/94;
- sono presenti attrezzature multimediali idonee a favorire un approccio all'insegnamento linguistico basato sull'uso integrato delle Nuove Tecnologie;
- è necessario assicurare agli alunni continuità di apprendimento della lingua straniera studiata nella scuola elementare, con particolare riguardo agli istituti comprensivi di scuola materna, elementare e media;
- non risulta attivata una percentuale di classi a tempo prolungato superiore al 30%;
- l’offerta curricolare della scuola è limitata ad una sola lingua straniera senza possibilità di scelta per gli alunni;
- sono stati attivati in precedenza corsi facoltativi di seconda lingua.

2. Adempimenti delle scuole

2.1. Modalità da seguire per la richiesta di finanziamento

Le scuole scelgono tra le principali lingue comunitarie (inglese, francese, tedesco e spagnolo) quella/e da attivare ed inoltrano la richiesta al Provveditore agli Studi per il finanziamento di uno o più corsi, tenendo conto dei seguenti parametri:

a) ogni corso sarà rivolto a un gruppo di alunni con omogeneo livello di competenza rispetto alla seconda lingua comunitaria da studiare;
b) il gruppo sarà costituito da alunni provenienti da classi anche diverse, nelle quali la lingua straniera curricolare sia altra rispetto alla lingua comunitaria aggiuntiva da apprendere;
c) il gruppo sarà formato, di norma, da 15 alunni.

I Capi d'Istituto avanzano richiesta di finanziamento entro il 30 settembre 1998.

La richiesta sarà corredata da: - una proposta, elaborata secondo lo schema di cui all'allegato B, da approvarsi con delibera del Consiglio d'Istituto e del Collegio dei Docenti; - un'informazione sull'insegnamento curricolare, sperimentale e/o facoltativo, delle lingue straniere eventualmente presente nella scuola nonché sulle dotazioni multimediali e sui materiali didattici disponibili per l'insegnamento della seconda lingua comunitaria, elaborato secondo la scheda analitica di cui all'allegato C.

2.2. Risorse
Per l’attivazione dell’insegnamento della seconda lingua la scuola dovrà predisporre le condizioni strutturali di funzionamento dei corsi. Si forniscono di seguito indicazioni essenziali e criteri-guida per la pianificazione e la messa in opera di mezzi e strumenti necessari.

a) **Risorse finanziarie**

Le richieste di finanziamento debbono essere corredate da un piano di spesa ch 1000 e specifici:

* la quantificazione dei costi per le ore di insegnamento della seconda lingua comunitaria che, rispetto alle 240 ore complessive, si intendono realizzare nell’anno scolastico 1998-99;

* gli eventuali oneri riflessi;

* le spese per materiali didattici, compreso il software relativo alla lingua da introdurre (fino a un massimo di L. 600.000);

* le eventuali spese connesse con la certificazione delle competenze raggiunte dagli alunni, nel caso si preveda di fare ricorso ad Enti certificatori convenzionati.

b) **Risorse professionali**

L’insegnamento della lingua straniera può essere assegnato a ciascun docente fino al massimo di ore previste dal contratto per prestazioni aggiuntive oltre gli obblighi di servizio; resta escluso comunque che le ore prestate, stante la loro natura non curricolare, possano essere attribuite come completamento dell’orario di cattedra.

Le scuole privilegeranno il reclutamento di docenti o esperti in possesso di titoli rilasciati da Università straniere o Istituti stranieri autorizzati (operanti anche in Italia) quali una laurea o un titolo post-laurea e/o una specializzazione per l’insegnamento della lingua comunitaria come lingua straniera e/o un attestato comprovante il livello di conoscenza della lingua comunitaria da insegnare, scegliendoli nell’ordine tra:

- docenti in servizio a tempo indeterminato titolari dell’insegnamento specifico;
- docenti con abilitazione specifica;
- esperti esterni al sistema in possesso di una laurea e/o una specializzazione per l’insegnamento della lingua comunitaria come lingua straniera conseguita all’estero o di un diploma di specializzazione, rilasciato da Università o Istituti stranieri autorizzati;
- laureati in lingue straniere con corso di studi quadriennale nella lingua da insegnare.

In via subordinata, in mancanza di docenti o esperti in possesso dei sopra indicati titoli rilasciati da Università straniere o Istituti stranieri autorizzati, il personale da utilizzare verrà scelto tra le stesse categorie.
Fermo restando il possesso dei requisiti sopra indicati, i Consigli di Istituto potranno eventualmente definire ulteriori criteri per l'individuazione del personale di cui trattasi.

Per gli esperti esterni si stipulano contratti d'opera, per i docenti in servizio si procede con l'affidamento di incarichi per prestazioni aggiuntive.

L'Amministrazione assicurerà interventi di formazione specifica per i docenti di lingua straniera in servizio che si rendano disponibili all'insegnamento non curricolare della lingua comunitaria e per quelli assunti con contratti d'opera.

c) Risorse tecnologiche e didattiche

A livello di singola scuola, al fine di favorire strategie individuali di apprendimento, le attrezzature multimediali serviranno anche alla costituzione di un Centro di Autoapprendimento con risorse quali, ad es., computer multimediali con accesso a Internet, postazioni audio-video, postazioni con televisore e antena parabolica, mediateca, corsi di lingua, materiali autentici.

Sarà particolarmente utile il ricorso a servizi telematici di supporto all'insegnamento/apprendimento delle lingue tramite la consultazione di banche dati e di archivi di materiali didattici selezionati, nonché la partecipazione a piattaforme telematiche (newsgroups, mailing lists, forum...) per condividere le esperienze più significative e per progettare a distanza, sia a livello nazionale sia a livello europeo.

Inoltre le scuole potranno avvalersi del supporto dei Centri Risorse Territoriali.

2.3. Certificazione degli esiti e crediti formativi

Gli alunni che avranno frequentato i corsi facoltativi di seconda lingua possono richiedere, al termine del percorso di 240 ore, di sostenere un esame finale che attesti il livello di preparazione raggiunto, secondo le modalità previste dalla C.M. 335 del 28 maggio 1997.

La scuola può altresì rivolgersi ad uno 1000 degli organismi certificatori convenzionati, che saranno indicati con successiva comunicazione, al fine di consentire agli alunni esami finali e/o in itinere per il conseguimento di un certificato internazionale di competenze nella lingua studiata, spendibile per il proseguimento della carriera scolastica e, successivamente, nel mondo del lavoro.

Infine, la scuola potrà anche attivare forme di auto-certificazione in itinere e finale delle competenze acquisite dagli alunni nella seconda lingua secondo un proprio progetto di ricerca-azione.

Le competenze acquisite e certificate potranno rappresentare parte di un Portfolio delle competenze linguistiche di ogni alunno.
3. Centri di supporto

3.1. Centro di Autoapprendimento a livello di singola scuola

Tra le scuole medie che attivano l'insegnamento di una seconda lingua comunitaria, i Provveditori agli Studi segnaleranno alla Direzione Generale dell'Istruzione secondaria di Primo Grado quelle già inserite nel Piano di Sviluppo delle Tecnologie Didattiche ai sensi delle CC.MM. n. 282/97 e n. 196/98 per il progetto 1b. Ad esse verrà assegnato un finanziamento aggiuntivo di L. 2.000.000, principalmente finalizzato all’acquisto di software nonché ad attività di formazione in vista della costituzione di un Centro di autoapprendimento.

3.2. Centro Risorse Territoriale

I Provveditori agli Studi potranno assegnare ad una scuola del territorio adeguatamente attrezzata (aula multimediali, software didattico, archivi di documentazione, servizi telematici, collegamenti INTERNET, ecc.) e d’intesa con i competenti organi scolastici il compito di organizzare un Centro Risorse Territoriale con le seguenti caratteristiche:

1) luogo fisico di incontro e di socializzazione delle esperienze dei docenti di varie discipline;
2) ambiente con strutture per l’innovazione e la documentazione:
   - non autosufficienti, ma collegate ai altre istituzioni, agenzie formative, luoghi della ricerca;
   - con forte valenza organizzativa;
   - capaci di far circolare linguaggi accessibili e di essere in relazione con gli altri in una logica di rete.

Nella individuazione del Centro Risorse Territoriale si raccomanda di tener conto dell’esistenza di scuole già attrezzate che hanno realizzato i progetti Rete, Multilab, Deure, Telecomunicando, ovvero di scuole che sono già state incaricate di analoghe funzioni, i cui elenchi si possono trovare nel sito web della Direzione Generale dell'Istruzione Secondaria di 1° grado http://www.bdp.it/~dgsm0001.

Ai Centri Risorse Territoriali viene assegnato un finanziamento specifico di L.4.000.000.

4. Assistenza, monitoraggio e valutazione

Le attività circa il monitoraggio, il supporto, l'assistenza e la valutazione degli interventi verranno svolte a livello nazionale da un apposito gruppo di coordinamento operante presso la Direzione Generale dell'Istruzione Secondaria di 1° grado. Saranno successivamente fornite indicazioni al riguardo.

I materiali e gli strumenti via via elaborati saranno inviati alle scuole e/o messi in rete su apposito sito.
Gli Ispettori Tecnici di lingue straniere assicureranno l'assistenza alle scuole nell'elaborazione dei progetti e nella messa a punto di modalità di certificazione dei crediti formativi e cureranno il monitoraggio e la valutazione degli interventi a livello regionale.

Ogni Ufficio Scolastico individuerà un referente provinciale per il piano d'introduzione della seconda lingua che curerà, nella fase di avvio, la diffusione delle informazioni alle scuole interessate e faciliterà la comunicazione tra le scuole e i Centri Risorse via via attivati sul territorio.

E' facoltà dei Provveditori valutare l'opportunità di costituire un gruppo di lavoro provinciale con compiti di sostegno alle scuole, d'intesa con gli Ispettori Tecnici di lingua straniera che ne c 1000 ureranno, laddove possibile, il coordinamento.

5. Formazione in servizio

I Provveditori agli Studi, sulla base di un piano di finanziamento in corso di definizione da parte di questo Ministero, predisporranno interventi di formazione per i docenti utilizzati per l'attivazione della seconda lingua straniera comunitaria.

Con successiva comunicazione saranno fornite le linee progettuali e operative dei percorsi di aggiornamento.

IL MINISTRO
APPENDIX 5.4 – Details of Progetto Lingue 2000 (original version)

Obiettivi
Il progetto Lingue 2000, finanziato con i fondi messi a disposizione dalla legge 440/97, intende innovare l'insegnamento e l'apprendimento delle lingue straniere favorendo l'acquisizione di precise competenze comunicative da parte degli allievi di ogni ordine e grado di scuola.

Metodologia
Il progetto introduce i seguenti aspetti di innovazione:

- costituzione di gruppi di apprendimento omogenei per livello di competenza e composti da non più di 15 allievi
- suddivisione del monte ore annuale in moduli di apprendimento di breve durata
- uso delle nuove tecnologie anche per favorire l'autoapprendimento e quindi lo studio individualizzato
- possibilità di ottenere certificazioni dei livelli di competenza raggiunti (misurati secondo la scala globale di riferimento del Consiglio d'Europa) rilasciate da Enti certificatori riconosciuti e accreditati a livello internazionale.

Strumenti
"Lingue 2000" (che ha preso il via dall'anno scolastico 1999-2000) è realizzato attraverso i piani operativi delineati da ciascun Ufficio centrale (dal Servizio scuola materna alle Direzioni delle scuole secondarie di secondo grado). Ogni Ufficio adotta le azioni necessarie per attuare le finalità del progetto tenendo conto delle specificità dei vari indirizzi di studio.

Selezione dei progetti
Scuole di ogni ordine e grado su tutto il territorio nazionale possono chiedere di aderire al progetto e ottenere così i finanziamenti necessari all'attuazione degli interventi previsti dai singoli Uffici Centrali.

Monitoraggio
Un monitoraggio sistematico sull'efficacia e l'efficienza del progetto viene attivato a livello provinciale, regionale e nazionale.
Appendix 5.5

Official Announcement of PL2000

(original text without appendices)

DIREZIONE GENERALE PER L'ISTRUZIONE CLASSICA, SCIENTIFICA E MAGISTRALE: UFFICIO DI COORDINAMENTO PER L'AUTONOMIA

Circolare Ministeriale n.197 Roma, 6 agosto 1999


Premessa

La Legge n. 440/97 ha istituito il fondo per l'arricchimento e l'ampliamento dell'offerta formativa e per gli interventi perequativi destinato, tra l'altro, all'attivazione dell'insegnamento di una seconda lingua comunitaria nella scuola media e ad iniziative per l'adeguamento dei programmici di studio dei diversi ordini e gradi di istruzione.

Nel quadro di tale normativa, la D.G. Istruzione Scuola Secondaria di I grado ha introdotto nell'anno scolastico 1998/1999 lo studio non curricolare e facoltativo di una seconda lingua comunitaria; la D.G. dell'Istruzione Elementare ha promosso e realizzato iniziative tese al miglioramento e al potenziamento delle lingue comunitarie.


La Direttiva n. 180 del 19 luglio, nel definire i criteri generali per la ripartizione delle somme relative all'anno finanziario 1999, destina al potenziamento dell'insegnamento delle lingue comunitarie la cifra complessiva di circa 120 miliardi di lire comprensive del piano di formazione.

La presente circolare definisce le procedure per la partecipazione alle attività progettuali.
1. Adempimenti delle scuole

Le istituzioni scolastiche che intendono partecipare alle azioni previste nel progetto sopra menzionato, previa deliberazione degli organi competenti, formulano agli Uffici scolastici provinciali la richiesta di finanziamento compilando la scheda - Allegato 1 - che consta di due sezioni:

Sezione A (Dati generali della scuola);
Sezione B, (richiesta di finanziamento per ciascun ordine e indirizzo scolastico).

La scheda deve essere inviata entro il 30/09/99.

Limitatamente alle esperienze avviate nell'anno scolastico 1998/99 nella scuola media e nella scuola elementare, in caso di programmazione degli interventi su scala pluriennale e ove sussistano ancora le condizioni, i Provveditori assicurano con priorità la prosecuzione delle azioni a garanzia di continuità per gli alunni.

Per quanto riguarda la scuola materna, al fine di assicurare la massima efficacia del progetto, potranno partecipare alle azioni previste esclusivamente:

scuole funzionanti ad orario completo (con doppio organico);
scuole funzionanti nei circoli didattici nei quali si prevede l'insegnamento della stessa lingua fin dal primo ciclo della scuola elementare.

Per quanto riguarda la scuola Media, nel caso in cui il numero degli alunni iscritti al secondo anno risulti particolarmente ridotto, la prosecuzione potrà essere assicurata anche attivando corsi in collaborazione tra scuole limitrofe, previ accordi tra i capi d'Istituto interessati.

In ogni caso le scuole sono tenute a compilare l'apposita scheda per la richiesta dei finanziamenti necessari. Nel caso di corsi organizzati su più scuole medie, la Scheda di cui all'All. 1 sarà compilata a cura del Dirigente scolastico della scuola sede del corso e destinataria del finanziamento.

Premesso che le linee operative sono esplicitate nel Progetto Lingue 2000, si richiama l'attenzione sui seguenti aspetti:

a - il gruppo di apprendimento

Il gruppo di apprendimento può non coincidere con il gruppo classe ed essere costituito da alunni provenienti da sezioni o classi diverse secondo criteri che garantiscano la massima efficacia dell'azione progettata (es.: criteri di omogeneità per competenze, motivazioni, interessi, criteri relativi a specifiche esigenze formative di alunni sotto il profilo linguistico...)

Si richiama l'attenzione sulla circostanza che, per mantenere la validità organizzativa e
didattica progettuale, ogni corso deve essere tendenzialmente costituito da 15 allievi, elevabile fino ad un massimo di 20 dove lo richiedano particolari esigenze.

Nella costituzione dei gruppi di apprendimento, tenuto conto della continuità del percorso formativo organizzato in moduli su più anni scolastici e al fine di ottimizzare le risorse impiegate, andrà acquisita espressamente la disponibilità da parte degli allievi interessati a frequentare le attività di lingua comunitaria per il monte orario complessivo dell’azione prescelta.

b – i docenti

Le scuole individueranno le risorse professionali interne e/o esterne più idonee allo svolgimento del compito di insegnamento della lingua straniera secondo i percorsi delineati nel disegno progettuale. A tal fine esse possono servirsì della consulenza del gruppo lingue insediato presso il provveditorato agli studi, menzionato più avanti.

Nella scelta del personale da utilizzare, per la scuola secondaria di 1° e 2° grado, si dovrà privilegiare il reclutamento di docenti e/o esperti in possesso di titoli di specializzazione rilasciati da università straniere o da enti a ciò autorizzati (operanti anche in Italia) quali una laurea o un diploma post-laurea e/o una specializzazione per l'insegnamento della lingua come lingua straniera e/o un attestato comprovante il livello di conoscenza della lingua comunitaria da insegnare. La scelta avverrà nell'ordine tra:

- docenti in servizio a tempo indeterminato titolari dell'insegnamento specifico;
- docenti con abilitazione specifica
- esperti esterni al sistema in possesso di una laurea e/o di una specializzazione per l'insegnamento della lingua come lingua straniera conseguita all'estero in Università o Istituti autorizzati
- laureati in lingue straniere con corso di studi quadriennale nella lingua da insegnare.

Per la scuola materna si individuano i docenti e gli esperti in possesso dei titoli indicati nel secondo capoverso del punto B, scegliendoli nell'ordine tra:

- docenti di scuola materna in servizio nella scuola
- docenti di scuola materna in servizio nel circolo
- docenti di scuola elementare in servizio nel circolo
- docenti di scuola media titolari dell'insegnamento della lingua straniera nello stesso istituto (solo per le scuole comprensive)
- esperti esterni al sistema e prioritariamente esperti di madrelingua, in possesso dei titoli già menzionati e di documentate esperienze/competenze nel campo dell'apprendimento "precoce" di una lingua straniera.
Per le ultime tre categorie un ulteriore titolo preferenziale è costituito dall’abilitazione all’insegnamento nella scuola materna.

Per la scuola elementare si rimanda alle categorie di personale docente individuate nella C.M.347 del 7.8.98.

In mancanza di docenti o esperti in possesso dei titoli di specializzazione sopra indicati, il personale da utilizzare verrà scelto tra le stesse categorie.

I Consigli di Circolo o di Istituto potranno, fermo restando il possesso dei requisiti sopra indicati, definire anche ulteriori criteri di scelta.

I docenti a tempo determinato o a tempo indeterminato sono retribuiti secondo il CCNL 1998/2001 per le ore aggiuntive all’insegnamento di cattedra; i docenti o esperti esterni sono assunti con contratto d’opera. Per la retribuzione di queste figure professionali si fa riferimento alla C.M. 446/98 - ufficio di Gabinetto - del 10 novembre 1998, che prevede di potere superare il costo previsto per il pagamento delle ore aggiuntive ai docenti interni solo in presenza di adeguata motivazione.

Per i docenti coinvolti nelle attività di cui al Progetto Lingue 2000 verrà predisposto un piano specifico di formazione/informazione di cui sarà data comunicazione con successiva nota e per il quale si farà ricorso anche ai programmi comunitari.

2. Adempimenti degli Uffici Scolastici provinciali

I Provveditori agli Studi, acquisite le richieste delle istituzioni scolastiche, definiscono il piano complessivo dei corsi da finanziare tenuto conto delle somme a loro assegnate per ciascun ordine e grado di scuola.

Nell’ambito del finanziamento assegnato a ciascun Ufficio scolastico provinciale di cui al prospetto di ripartizione in appendice, sarà riconosciuta la precedenza alle richieste delle scuole nelle quali:

- siano garantite le condizioni per la prosecuzione delle classi e dei gruppi di apprendimento costituiti nell’anno scolastico 1998/99
- si preveda il concorso finanziario degli enti locali e di soggetti esterni
- le iniziative progettuali siano deliberate da reti di scuole
- siano presenti attrezzature multimediali idonee a favorire un approccio all’insegnamento linguistico basato sull’uso integrato delle Nuove Tecnologie e la costituzione di un centro di autoapprendimento
- sia necessario garantire agli allievi continuità di apprendimento della lingua straniera studiata nel ciclo scolastico precedente e lo sviluppo negli anni successivi.
Spetta al Provveditore:

istituire il **gruppo lingue unitario**, che tenga conto delle realtà già operanti sul territorio e che sia espressione delle professionalità della scuola materna alle secondarie di secondo grado, quali docenti, formatori, capi di istituto, ispettori tecnici. Il gruppo può operare anche per singoli segmenti di istruzione ma costituisce la sede unitaria per i dibattiti sulle azioni da intraprendere sul territorio nell'ambito della **legge n.440/97** citata. La struttura può essere l'occasione più idonea per correlare processi e prodotti delle iniziative ex L. 440/97 con quelli delle attività curricolari e opera in stretto collegamento con il nucleo dell'autonomia. Il gruppo lingue collabora con il Provveditore agli studi per tutti gli atti e le iniziative che attengono all'impianto e alla migliore riuscita delle attività progettuali.

indicare il **referente provinciale** per l'attuazione del **Progetto Lingue 2000**, valorizzando il personale già impegnato in tali compiti (ad esempio i componenti dei Gruppi Provinciali di lingua Straniera della scuola elementare o i referenti del progetto seconda lingua comunitaria nella scuola media)

Il nominativo del referente e dei componenti il gruppo deve essere comunicato all'ufficio scrivente, che provvederà a fornire l'elenco a tutti gli uffici centrali (scheda **Allegato 2**) entro il 30 settembre 1999;

individuare una o più scuole che possano fungere come **centro risorse territoriale** (CRT). I centri territoriali già individuati per il progetto della scuola media possono estendere la loro funzione alle scuole di ogni ordine e grado, previa verifica approfondita sulle reali condizioni di operatività. Si richiama la massima attenzione affinché la scelta ricada comunque su una scuola adeguatamente attrezzata allo scopo, che permetta una libera fruizione delle strutture e dei macchinari necessari per svolgere tale ruolo. Si tratta di fotocopiatrici, attrezzature informatiche moderne con possibilità di accesso in rete, attrezzature multimediali che permettano riproduzioni video audio fedeli e utili agli scopi didattici. Il CRT è la sede dove si riunisce il gruppo lingue, dove si svolgono gli incontri di aggiornamento e formazione e tutte le iniziative importanti che attengono alla migliore riuscita del progetto. A tal fine, ove sia ritenuto utile e funzionale, si possono individuare più sedi e/o cambiare quelle individuate nell'anno scolastico 1998/99, in relazione a specifiche esigenze territoriali e compatibilmente con l'ottimizzazione delle risorse finanziarie assegnate a ciascun ufficio provinciale. L'elenco dei centri risorse sarà comunicato a questo ufficio a mezzo della scheda di cui all'allegato 2/1.

Al termine delle operazioni di assegnazione delle risorse alle scuole, i Provveditori avranno cura di far pervenire a questo ufficio l'acclusa scheda (**allegato 3**) che dà conto delle azioni autorizzate e finanziate per le azioni promosse dagli uffici centrali.

Infine, nell'ambito delle conferenze di servizio programmate all'inizio del prossimo anno scolastico, ciascun Provveditore voterà riservare uno spazio specifico alla presentazione delle linee di sviluppo del **Progetto Lingue 2000**.
3. Finanziamenti

La presente circolare riporta in appendice la ripartizione agli uffici provinciali del 90% delle risorse destinate alle attività del Progetto Lingue 2000, articolato per centri di responsabilità di spesa. Esso è destinato alla remunerazione delle attività didattiche nella scuola e all'acquisto di materiali didattici per gli studenti. Per quest'ultima voce è prevista la somma di un massimo di Lit. 600.000 per ogni gruppo o classe di apprendimento. I materiali rimangono di proprietà della scuola.

Le scuole medie che hanno già ricevuto questo finanziamento nell'anno scolastico 1998/99 possono richiedere un ulteriore contributo fino ad un massimo di Lit.150.000 per ogni gruppo di apprendimento.

Il restante 10% dello stanziamento verrà assegnato, con apposito provvedimento, in attuazione di quanto disposto dalla citata direttiva n. 180/99, per azioni perequative e di supporto ed in particolare per:

dotare il centro risorse di adeguati supporti software, materiali didattici e informativi necessari per lo svolgimento delle attività di cui è sede;

retribuire le ore aggiuntive per i docenti e il personale ATA che si impegnino a far funzionare il CRT, in particolare per le attività connesse con le iniziative di formazione dei docenti;

consentire alle scuole che ne fanno richiesta e fino a un massimo di Lit. 2.000.000, l'acquisto di materiali didattici multimediali, a condizione che queste dispongano di una adeguata attrezzatura (centro di autoapprendimento), per la loro fruizione. Non sono ammesse a tale finanziamento le scuole che già ne hanno beneficiato nell'anno scolastico 1998/99;

Per quanto riguarda le iniziative connesse alla certificazione, i cui oneri ricadono entro il 10% indicato, verranno emanate ulteriori disposizioni dopo avere siglato i protocolli d'intesa con gli enti certificatori.

Inoltre, sempre nell'ambito della quota del 10%, il provveditore può finanziare ulteriori acquisti per il migliore successo dell'iniziativa o, nel caso di economie, altre azioni relative al Progetto Lingue 2000.

4. Adempimenti dell'Amministrazione Centrale

L'Amministrazione Centrale, oltre al compito di coordinare e promuovere tutte le attività, provvede a:

impiantare il monitoraggio delle azioni che si intraprendono ai vari livelli
valutare gli esiti di apprendimento, con modalità che saranno successivamente comunicate

diffondere le informazioni sul territorio nazionale

promuovere protocolli di intesa con soggetti terzi con i quali le istituzioni scolastiche potranno successivamente stipulare convenzioni per i servizi che detti soggetti possono offrire.


Le direzioni generali e gli uffici centrali interessati provvederanno autonomamente, se lo riterranno opportuno, a emanare circolari specifiche su singoli problemi attinenti i percorsi tipici degli indirizzi amministrati.

Considerata l’importanza degli adempimenti, si confida nella consueta puntuale collaborazione.

IL MINISTRO
f.to Berlinguer
Appendix 5.6

CEFR as used in the planning for PL2000 by MPI staff

CEFR Table taken from Draft 2 of the CEFR (1997) and translated into Italian by the Ministry team. This table was used in Guidelines for the 2nd language in middle school in 1998, and then for the PL2000 when it was announced 1999.

<table>
<thead>
<tr>
<th>Livello Base</th>
<th>A1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprende e usa espressioni di uso quotidiano e frasi basilari tese a soddisfare bisogni di tipo concreto. Sa presentare se stesso/a e gli altri ed è in grado di fare domande e rispondere su particolari personali come dove abita, le persone che conosce e le cose che possiede. Interagisce in modo semplice purché l’altra persona parli lentamente e chiaramente e sia disposta a collaborare.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Livello Base</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprende frasi ed espressioni usate frequentemente relative ad ambiti di immediata rilevanza (Es. informazioni personali e familiari di base, fare la spesa, la geografia locale, l’occupazione). Comunica in attività semplici e di routine che richiedono un semplice scambio di informazioni su argomenti familiari e comuni. Sa descrivere in termini semplici aspetti del suo background, dell’ambiente circostante sa esprimere bisogni immediati.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Livello Base</th>
<th>B1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprende i punti chiave di argomenti familiari che riguardano la scuola, il tempo libero ecc. Sa muoversi con disinvoltura in situazioni che possono verificarsi mentre viaggia nel paese in cui si parla la lingua. E’ in grado di produrre un testo semplice relativo ad argomenti che siano familiari o di interesse personale. E’ in grado di descrivere esperienze ed avvenimenti, sogni, speranze e ambizioni e spiegare brevemente le ragioni delle sue opinioni e dei suoi progetti.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Livello Base</th>
<th>B2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprende le idee principali di testi complessi su argomenti sia concreti che astratti, comprese le discussioni tecniche nel suo campo di specializzazione. E’ in grado di interagire con una certa scioltezza e spontaneità che rendono possibile un’interazione naturale con i parlanti nativi senza sforzo per l’interlocutore. Sa produrre un testo chiaro e dettagliato su un’ampia gamma di argomenti e spiegare un punto di vista su un argomento fornendo i pro e i contro delle varie opzioni.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Livello Base</th>
<th>C1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprende un’ampia gamma di testi complessi e lunghi e ne sa riconoscere il significato implicito. Si esprime con scioltezza e naturalità. Usa la lingua in modo flessibile ed efficace per scopi sociali, professionali e accademici. Riesce a produrre testi chiari, ben costruiti, dettagliati su argomenti complessi, mostrando un sicuro controllo della struttura testuale, dei connettori e degli elementi di coesione.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Livello Base</th>
<th>C2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Comprende con facilità praticamente tutto ciò che sente e legge. Sa riassumere informazioni provenienti da diverse fonti sia scritte che scritte, ristrutturando gli argomenti in una presentazione coerente. Sa esprimersi spontaneamente, in modo molto scorrevole e preciso, individuando le più sottili sfumature di significato in situazioni complesse.</td>
</tr>
</tbody>
</table>
Appendix 5.7

Profiles of the Case Study schools

There were seven state schools in the study.
Descriptive information was collected from the school Heads in October 2001.
The school data were presented anonymously in the final impact study report.

<table>
<thead>
<tr>
<th>SCHOOL A – Scuola elementare statale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
</tr>
<tr>
<td>Classes</td>
</tr>
<tr>
<td>Teachers</td>
</tr>
<tr>
<td>English teachers/assistants</td>
</tr>
<tr>
<td>PL2000 classes</td>
</tr>
<tr>
<td>PL000 English classes</td>
</tr>
<tr>
<td>PL English hour per week</td>
</tr>
<tr>
<td>Languages taught in the school</td>
</tr>
<tr>
<td>External English tests</td>
</tr>
<tr>
<td>Access to resource centre</td>
</tr>
<tr>
<td>Head’s view on the main impact of PL2000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SCHOOL B - Istituto comprensivo statale (elementary and middle school)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Students</td>
</tr>
<tr>
<td>Classes</td>
</tr>
<tr>
<td>Teachers</td>
</tr>
<tr>
<td>English teachers/assistants</td>
</tr>
<tr>
<td>PL2000 classes</td>
</tr>
<tr>
<td>PL000 English classes</td>
</tr>
<tr>
<td>PL English hour per week</td>
</tr>
<tr>
<td>Languages taught in the school</td>
</tr>
<tr>
<td>External English tests</td>
</tr>
<tr>
<td>Access to resource centre</td>
</tr>
<tr>
<td>Head’s view on the main impact of PL2000</td>
</tr>
</tbody>
</table>
SCHOOL C - Scuola media statale
Students 725
Classes 34
Teachers 73
English teachers/assistants 6
PL2000 classes 23
PL English hour per week 3
Languages taught in the school English, French
External English tests KET (A2) for some selected third year students
Access to resource centre Yes, near the school. Computers, in the school: books, videos, magazines
Head's view on the main impact of PL2000 Students aware of the importance of language as a means of communication

SCHOOL D - Scuola media statale
Students 850
Classes 35
Teachers 70
English teachers/assistants 6
PL2000 classes 2
PL2000 English classes 2
PL English hour per week 3
Languages taught in the school English, French
External English tests Year 3 KET (A2)
Access to resource centre No
Head's view on the main impact of PL2000 More students (and parents) wanting to participate in PL2000 courses

SCHOOL E - Liceo ginnasio statale
Students 905
Classes 43
Teachers 96
English teachers/assistants 9 / 2
PL2000 classes 15
PL2000 English classes 9
PL English hours per week 1 hpw, a.m. 1.5 hpw, p.m.
Languages taught in the school English, French, German, Spanish
External English tests | KET (A2), PET (B1), FCE (B2)
---|---
Access to resource centre | Yes, in the school: books, videos, cassettes, magazines. Computer room used for p.m. PL2000 classes

Head’s view on the main impact of PL2000 | More co-operation, more objectivity in evaluation, which motivates the students

**SCHOOL F - Liceo ginnasio statale**
Students | 1300
---|---
Classes | 53
Teachers | 134
English teachers/assistants | 7 / 2
PL2000 classes | 53
PL2000 English classes | 53
PL English hours per week | 3/4
Languages taught in the school English, French, German, Spanish
External English tests | KET (A2), PET (B1), FCE (B2)
Access to resource centre | Yes, in the school: computer and language lab
Head’s view on the main impact of PL2000 | Improvement in foreign languages

**SCHOOL G - Istituto tecnico statale**
Students | 700
---|---
Classes | 31
Teachers | 81
English teachers/assistants | 8 / 2
PL2000 classes | 16
PL2000 English classes | 16
PL English hours per week | 2 hpw, p.m, 4/5 hpw, a.m.
Languages taught in the school English, French, German, Spanish
External English tests | Year 3: PET (B1); Year 4 or 5: FCE (B2); Year 5: CAE2 (C1)
Access to resource centre | Yes, in the school: library, computers, learning materials, video conferencing disc
Head’s view on the main impact of PL2000 | Motivation (of teachers and students)
| Reduction in the cost of courses
Appendix 5.8

Memorandum of Understanding (renewal 2002)

Signed in 2002 under the new Minister Moratti (Forza Italia) with the examinations board.

All. 1 nota prot. 883/02: PROTOCOLLO DI INTESA

TRA
il Ministero dell'Istruzione dell'Università e della Ricerca, rappresentato dal Ministro Letizia Moratti e i seguenti Enti certificatori:

ALLIANCE FRANÇAISE –
CENTRE PILOTE DELF-DALF

CAMBRIDGE UCLES

CHAMBRE DE COMMERCE ET DE L'INDUSTRIE DE PARIS

CITY & GUILDS INTERNATIONAL (PITMAN)

EDEXCEL INTERNATIONAL LONDON

GOETHE-INSTITUT INTER NATIONES

MINISTERIO DE EDUCACION, CULTURA Y DEPORTE

ÖSTERREICHISCHES SPRACHDIPLOM DEUTSCH, VIENNA

TRINITY COLLEGE LONDON

WEITERBILDUNG-TESTSYSTEME

rappresentato da: Marie-Hélène Estève

rappresentato da: Liam Vint (University of Cambridge Local Examination Syndicate)

rappresentato da: Marie-Hélène Estève

rappresentato da: Liz Sell

rappresentato da: Teresa Jacobs

rappresentato da: Uwe Mohr

rappresentato da: Ángel Maria Sáinz

rappresentato da: Maria Kummer Elisabeth

rappresentato da: Roger Bowers

rappresentato da: Giuseppina Foti

PREMESSO CHE

Il Ministero dell'Istruzione, dell'Università e della Ricerca, nell'ambito del piano per l'arricchimento e l'ampliamento dell'offerta formativa finanziato con i fondi della Legge n.440/97, ha promosso a partire dall'anno scolastico 1999/2000 il Progetto Lingue 2000 per il potenziamento dello studio delle lingue straniere come offerta aggiuntiva extracurricolare, destinato agli allievi delle scuole di ogni ordine e grado.

Il Progetto Lingue 2000, diramato con c. m. prot. n.160 del 24 giugno 1999, prevede tra
l'altro interventi di certificazione degli esiti di apprendimento con descrizione dei livelli di competenza che si affiancano alla valutazione interna.
La certificazione si configura come "valore aggiunto" e fornisce agli allievi un credito formativo spendibile dentro e fuori il contesto scolastico.
Gli enti certificatori accreditati a livello internazionale rilasciano certificazioni riconosciute che si riferiscono ai livelli di competenza indicati nel Quadro Comune Europeo di riferimento, redatto dal Consiglio d'Europa.
L'accreditamento di ciascun organismo estraniero origina da autorizzazioni governative o di enti nazionali preposti dal Governo del paese alla qualità degli esami.
Un protocollo d'intesa di durata triennale tra il Ministero dell'Istruzione e i sette enti certificatori (ALLIANCE FRANÇAISE - CENTRE PILOTE DELF-DALF, CAMBRIDGE UCLES (University of Cambridge Local Examination Syndicate), MINISTERIO DE EDUCACION, CULTURA Y DEPORTE, CITY & GUILDS INTERNATIONAL (PITMAN), ESB (English Speaking Board), GOETHE-INSTITUT INTER NATIONES, TRINITY COLLEGE - LONDON) è stato siglato il 20 Gennaio 2000 e si ravvisa allo stato attuale la necessità di modificarne alcune sue parti riguardanti il servizio e le tariffe proposte, così come di integrarlo con altre offerte di nuovi enti certificatori parimenti accreditati e riconosciuti.
Tutto ciò premesso, al fine di continuare gli interventi di certificazione viene formalizzato e sottoscritto il seguente Protocollo d'Intesa tra le parti.

ART. 1
Gli Enti certificatori, al fine di consentire il raggiungimento delle finalità di cui in premessa, che è parte integrante al seguente protocollo, si impegnano a:

· sottoscrivere in attuazione del presente Protocollo ed entro i termini della sua validità convenzioni con le istituzioni scolastiche che aderiscono alle iniziative di certificazione esterna previste dal Progetto Lingue 2000, entro i limiti dei costi concordati secondo le modalità indicate nei prospetti allegati;

· offrire attività di supporto e di consulenza alle scuole che abbiano deliberato di aderire alla certificazione esterna, anche mediante la fornitura di materiali didattici, simulazioni di prove di verifica finale;

· somministrare prove di verifica finale atte al rilascio di attestati di livello delle competenze raggiunte.

ART. 2
Le certificazioni conseguite dagli alunni in attuazione del presente Protocollo costituiscono in base alla normativa vigente crediti formativi.
ART. 3

ART. 4
Per la realizzazione delle finalità e degli obiettivi previsti dal presente protocollo è istituito un apposito Comitato di Coordinamento composto da 5 componenti designati dal Ministero e da un rappresentante per ciascuno degli Enti Certificatori.
Ciascuna delle parti provvederà alla designazione del proprio rappresentante, comunicandone per iscritto il nominativo agli altri soggetti.
I rappresentanti degli Enti Culturali di Francia, Germania, Gran Bretagna e Spagna fanno parte di diritto del Comitato di Coordinamento.

ART. 5
Il Ministero dell'Istruzione con l'introduzione delle certificazioni esterne nel sistema scolastico nazionale, collegate all'offerta aggiuntiva di lingua straniera ai sensi della L. 440/97, si impegna a diffondere nelle scuole il presente protocollo, anche al fine di sensibilizzare l'utenza all'importanza di ottenere un certificato trasparente e spendibile oltre il contesto scolastico in quanto attestato di competenza riconosciuto a livello internazionale. Tale processo innovativo avvia al contempo un impatto di cambiamento di cultura e di prassi didattica fra gli insegnanti e può concorrere alla definizione del curricolo linguistico in termini di obiettivi specifici e di competenze, come previsto dall'art. 8 del DPR 8 marzo 1999 n. 275 (Regolamento dell'Autonomia).

ART. 6
I profili organizzativi e di gestione afferenti all'attuazione del protocollo di intesa verranno curati dalla Direzione Generale per gli Ordinamenti, che assicurerà altresì il necessario coordinamento con gli altri Uffici centrali interessati.

ART. 7
Il presente protocollo di intesa entra in vigore alla data della stipula ed ha durata annuale.
Quanto sopra è stato letto, approvato e sottoscritto dalle Parti.
Roma, 16 gennaio 2002
IL MINISTRO
Letizia Moratti
Appendix 5.9

EU White Paper on Education and Training

Teaching and Learning – Towards the Learning Society

PART ONE: THE CHALLENGES

I. THE THREE FACTORS OF UPHEAVAL

A. The impact of the information society

B. The impact of internationalisation

C. The impact of scientific and technological knowledge

II. A FIRST APPROACH: FOCUSING ON A BROAD KNOWLEDGE BASE

A. Grasping the meaning of things

B. Comprehension and creativity

C. Powers of judgement and decision making

III. A SECOND RESPONSE: DEVELOPING EVERYONE'S EMPLOYABILITY AND CAPACITY FOR ECONOMIC LIFE

A. What are the skills required?

B. How can a person become employable?

IV. DIRECTIONS FOR THE FUTURE

A. The end of debate on educational principles

B. The central question: towards greater flexibility

C. Action in the Member States

D. New developments
Part 1 outlines the challenges and Section IV - New Developments (pages 26-7) - is of particular relevance to the background to the PL2000.

Paragraph D points to three main directions, two of which can be seen explicitly in the PL2000 policy documents:

1. the increase in autonomy in education and training
2. evaluation – to justify increased funding
3. attention to disadvantaged groups
Part 2 deals with actions necessary in building the Learning Society. Section I deals with the first general objective which includes a focus on the recognition of skills, mobility and the use of multimedia educational software. Again these underpin Italian reform programme. Section IV is the fourth general objective (page 47 following) which is to develop proficiency in three community languages for all citizens of the community. Sometime referred to after the publication of the White Paper as the Mother Tongue plus 2 policy. This was already present in the Italian language policy before PL2000 where the focus was on introducing a second foreign language into the middle schools.

The following quote from the White Paper has been referred to frequently by the EU Commission since 1996:

In order to make for proficiency in three Community languages, it is desirable for foreign language learning to start at pre-school level. It seems essential for such teaching to be placed on a systematic footing in primary education, with the learning of a second Community foreign language starting in secondary school. It could even be argued that secondary school pupils should study certain subjects in the first foreign language learned, as is the case in the European schools. Upon completing initial training everyone should be proficient in two Community foreign languages.

The last point has also been taken up in some parts of Italy which have been experimenting with CLIL – content and language integrated learning – in some state school contexts (i.e. where curriculum content is delivered through the medium of English).
Appendix 5.10
Language certification in Lombardy

Lombardy became the leading region in Italy (and probably in Europe) in the use of external language certification provided by accredited and internationally recognised examination bodies. This was most notable for English and the use of the Cambridge examinations in particular. A discussion of the data collected and analysed from this region illustrates some of the major impacts of the PL2000 from a quantitative perspective.

The efforts of the Ufficio Scolastico Regionale – USR (the Local Education Authority) for Lombardy can be singled out as a major stakeholder and contributor to these outcomes of the project in that region; institutionally, through the interest and hard work of key individuals, the USR gave major support to the principal objectives of the PL2000. In other words, a commitment to the development of linguistic competence with an aim to ensure that students on leaving school were really able to use one (or perhaps more than one) foreign language at a level which was comparable to European standards (as set by CEFR and compared with other, similar parts of Europe). In particular, Gisella Langé, Technical Inspector and Progetto Lingue Manager, for the USR in Lombardy was influential and was a key individual stakeholder in both the implementation and evaluation of the project.

The USR for Lombardy undertook comparative research on data and results of the English language certification, and this was supported by the Cambridge ESOL impact study team (and the other main local partner for English, the British Council in Milan) as part of the impact study. This research looked at examination data obtained from students in Lombardy compared with data from students in other regions, both within Italy and abroad. In particular, Cambridge was able to provide data from the comparable regions in Europe, known in Italian as the “motors of Europe” because of their high level of economic development, i.e. Catalonia, Rhône-Alpes, Baden-Württemberg, in addition to Lombardy itself. The results of the research were compiled and presented publicly at conferences, for example at the Centro Congressi CARIPOLO of Milan in January 2004 as part of a dissemination programme (see summary report by Langé and Vint 2004).

The results of the research reflected well on Lombardy, as evidenced by the commitment to the project and the outcomes achieved. Out of 52,637 Cambridge ESOL examinations taken at a national level, 11,985 were taken in Lombardy. This represented nearly a quarter of all Italian Cambridge ESOL examination entries at that time. Annual growth of
candidature, furthermore, meant that it was the region which had the highest number of examinations (in the year 2002/03) compared to the other three European regions mentioned above.

Candidates from PL2000 schools by region - (%) for 2002-2003

The researchers at the USR determined six trends in external certification which they considered to be significant:

i). Demand for external certification from state schools continued to grow notwithstanding the official end of the Progetto Lingue 2000 (after 2002);

ii). Enrolments from state schools and the recognised private schools (“paritarie”) in represented the largest number of Cambridge examinations in Italy;

iii). The B1 (PET) examination had the highest candidature but it was noteworthy that demand for the B2 (First Certificate) was increasing faster (perhaps showing that standards were beginning to rise as a result of the project?);

iv). It confirmed a view that Italy had made the most significant effort to date in adopting the principles of the CEFR and that, Lombardy above all was in the avant garde (e.g. with the highest number of Cambridge candidates amongst a group of comparable regions in other countries);

v). The pass rate obtained by the candidates coming from the state schools was very encouraging and underlined the positive impact of the external certification on the language curriculum in schools and on the proficiency levels achieved;
vi). The choice of level B1 as the high school exit level (i.e. after 5-8 years of study of English language) had not been unrealistic and that within a few years a higher level could potentially be set (i.e. B2).

It can be concluded from this part of the study, that there had been an impact on the number of students participating in the use of the external certification and on the success rates in the examinations themselves. These impacts had in turn begun to change attitudes and approaches to the use of language certification within the Italian system.

It is also noteworthy that many of the objectives of the PL2000 persisted, even after the official end of the pilot; this was partly as a result of the commitment shown by the local education authorities and also the systemic effect the project was beginning to have more widely in the north of Italy (e.g. with the universities recognising the language certificates for their own purposes and thus increasing the currency and recognition).

For example, the USR for Lombardy chose to continue a level of funding for language certification after 2002. For the scholastic year 2002/03 they continued to ring fence funds for the certification of language skills and to allocate them to schools of all kinds and grades. For the following year, 2003/2004, it was also decided to allocate such funds but this time to target resources specifically towards the professional institutions (since these schools had had the lowest average of participation in the initiative) to encourage external certification. This reflective yet active intervention helped to ensure that the innovations brought about early in the project did not stop immediately when the funding for the activities needed to be dealt with differently.

By January 2003 the number of candidates for the Cambridge exams had continued on an upward curve; the early increase in the numbers had clearly been connected to the PL2000, and was therefore made up primarily of candidates from state schools. In the year after the official end of the project, the growth remained nearly the same for state schools as the more traditional non-school candidates indicating that the PL2000 was still having an effect.
Cambridge examination in Lombardy – Totals for 1999-2003

<table>
<thead>
<tr>
<th>Examination</th>
<th>1999/00</th>
<th>2000/01</th>
<th>%increase</th>
<th>2001/02</th>
<th>%increase</th>
<th>2002/3</th>
<th>%increase</th>
</tr>
</thead>
<tbody>
<tr>
<td>YLE</td>
<td>514</td>
<td>974</td>
<td>89%</td>
<td>1.319</td>
<td>35%</td>
<td>2.064</td>
<td>56%</td>
</tr>
<tr>
<td>KET</td>
<td>274</td>
<td>1.451</td>
<td>430%</td>
<td>2.531</td>
<td>74%</td>
<td>2.929</td>
<td>16%</td>
</tr>
<tr>
<td>PET</td>
<td>659</td>
<td>2.225</td>
<td>238%</td>
<td>3.725</td>
<td>67%</td>
<td>4.334</td>
<td>16%</td>
</tr>
<tr>
<td>FCE</td>
<td>271</td>
<td>1.438</td>
<td>431%</td>
<td>2.089</td>
<td>45%</td>
<td>2.479</td>
<td>19%</td>
</tr>
<tr>
<td>CAE</td>
<td>-</td>
<td>19</td>
<td>-</td>
<td>61</td>
<td>221%</td>
<td>46</td>
<td>-25%</td>
</tr>
<tr>
<td>CPE</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>1</td>
<td>-67%</td>
<td>5</td>
<td>400%</td>
</tr>
<tr>
<td>BEC</td>
<td>-</td>
<td>76</td>
<td>-</td>
<td>102</td>
<td>34%</td>
<td>97</td>
<td>-5%</td>
</tr>
<tr>
<td>CELS</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>41</td>
<td>-</td>
</tr>
<tr>
<td>Total</td>
<td>1.718</td>
<td>6.186</td>
<td>260%</td>
<td>9.828</td>
<td>59%</td>
<td>11.985</td>
<td>22%</td>
</tr>
</tbody>
</table>

Enrolments of candidates from schools in PL2000: percentage variation compared to previous year 1999-2003

Comparison between Italy as a whole and Lombardy

The Cambridge examination most widely taken in Lombardy was PET at level B1, although it was significant that entries for the higher level, FCE (B2) increased more rapidly (even if the YLE exams for children between 7 and 13 years old had by far highest percentage increase). A potential reason for a growing interest in the higher level qualification was the fact the three universities in Milan required FCE (B2) as the minimum level of language ability for certain purposes. This was also a sign that the “certification culture” was beginning to have wider systemic influences in other parts of society.
Compared to the three “motors of Europe”, the growth of Cambridge exams taken in Lombardy was clearly significant. While only four years earlier, Lombardy represented little more than a quarter of the Cambridge exams taken in Catalonia, by the end of the period the Lombardy figures had nearly reached the historic totals for Catalonia and had actually exceeded their numbers in the fourth year (although there was also a significant fall in numbers in the Spanish region that year).
The pass rates for the Cambridge exams in Lombardy also confirmed that the levels chosen are targets were realistic (at least where best practice existed); it was shown that the schools were able to prepare their students adequately for the challenge of the external certification (covering all four language skills) linked to the target CEFR levels and that this was comparable to a similar region in Germany which already known for its success in English language education in the schools.

Pass rates (%) for B1 - PET in the 4 regions
Pass rates (%) for KET (A2), CAE (C1) and CPE (C2) the 4 regions

<table>
<thead>
<tr>
<th>Region</th>
<th>KET (%)</th>
<th>CAE (%)</th>
<th>CPE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rhônes-Alpes</td>
<td>79</td>
<td>71</td>
<td>83</td>
</tr>
<tr>
<td>Catalunya</td>
<td>79</td>
<td>68</td>
<td>62</td>
</tr>
<tr>
<td>Baden-Württemberg</td>
<td>94</td>
<td>82</td>
<td>80</td>
</tr>
<tr>
<td>Lombardia</td>
<td>84</td>
<td>74</td>
<td>71</td>
</tr>
<tr>
<td>Lombardia PL2000</td>
<td>78</td>
<td>78</td>
<td>100</td>
</tr>
</tbody>
</table>

The analysis of PL2000 student questionnaires completed by all the Cambridge ESOL candidates at the time of the exams, underlined the fact that in Lombardy the number of years of study needed to be able to pass a PET or FCE examination was within the European average. In comparison with other regions looked at, it was only in Baden-Württemberg that most candidates for PET were ready to take the examination after just 5 years of study.

Number of years of study of English before attempting PET (B1)

Cambridge researchers also carried out analysis of the PL2000 candidates’ performance at the component level (by skill or paper) and within components by part or individual items. This was able to shed light on which skills or sub-skills were mastered most effectively at each level by the Italian students in comparison with others. The average scores in the Reading component of PET in Lombardy were systematically higher than the average for candidates worldwide. This kind of pattern was also seen for the Use of English in FCE.
Average scores in the Reading component of PET (B1). Percentage for Lombardy compared to Worldwide- June 2003

![Average scores in the Reading component of PET (B1). Percentage for Lombardy compared to Worldwide- June 2003](image)

Average scores in the Use of English of FCE (B2) - Percentage for Lombardy compared to Worldwide- June 2003

![Average scores in the Use of English of FCE (B2) - Percentage for Lombardy compared to Worldwide- June 2003](image)

Recognition of the certificates became an important element of the PL2000 project in that the students' success in the exams could be counted as a “credito formativo” or “formative credit”. These credits could be used within the school system – for example, during the state examination at the end of high school as part of the final assessment, or during the school cycles for passage from one choice of studies to another.

This kind of recognition also gained currency for work and study purposes in Italy outside the schools. The language certification that the students obtained, could be added to a “personal skills portfolio” which was envisaged by the PL2000 planners, or to a version of
the *European Language Portfolio* (ELP) which had been designed to accompany the CEFR in learning contexts in Italy – see Council of Europe website for details. In Lombardy, 7,500 students tried out a version of the ELP (accredited in 2002 by the Council of Europe) which has been designed for students aged from 11-15 years old. In addition, there was also a small-scale pilot by 400 students aged 16 or above in the Istituto Tecnico Commerciale Statale “P. Levi” of Seregno (within an international partnership coordinated by the University of Sofia – Bulgaria) of a professional skills version of the ELP (accredited by the Council of Europe in 2003). Langé and Vint in their report (op cit) concluded that results of these portfolio pilots in the context of PL2000 might enable an analysis and evaluation of all the segments of the educational system (including new elements such as certification, work placement, international exchanges, etc) that contribute to successful language education. Such evaluation could lead to further innovations, which in turn would enable the next generation of learners to become even more competitive at an international level.
Appendices to Chapter 6

Appendix 6.1

The British Institute of Florence (BIF)

The British Institute of Florence is an important teaching and cultural centre with a long tradition in Italy. It is also the largest examination centre in the Tuscany region for the Cambridge ESOL examinations.

It was established in 1917 and granted a Royal Charter in 1923. It was the first of the British cultural institutes to operate overseas and served as a model for the establishment of the British Council in 1934. It was founded by a group of Italian and English men and women including Arthur Acton father of Harold Acton, Walter Ashburner, Edward Hutton, Herbert Trench, G.M. Trevelyan, Lina Waterfield, Guido Biagi, Gaetano Salvemini, Guido Ferrado, Carlo Piaci and Aldo Sorani.

Its objectives as defined in the 1923 Charter were to promote "understanding between the citizens of Italy and the countries of the British Commonwealth through the maintenance in Florence of a library illustrating Italian and British culture and the promotion of the study of the English and Italian language and the cultures of both countries." The British Institute of Florence is recognised under the Anglo Italian Cultural Convention of 1953 and is registered as a charity in the UK (no. 290647).

The British Institute was first housed in the Loggia dei Rucellai, but since 1998 its language teaching has taken place in the Palazzo Strozzi. Today the Institute offers a range of educational and cultural programmes and according to their website (retrieved Dec 2007), their English language courses are linked to the levels of Common European Framework. They cater for all ages the youngest students beginning learning from the age of 5.

The Institute also offers Italian language courses at all levels and provides special courses in conjunction with a number of universities and other institutions, (including the University of Cambridge). Monthly History of Art courses are offered as well as a study abroad programme in conjunction with the University of Bristol.

Today the Harold Acton Library is one of the largest lending collections of English books in continental Europe with over 50,000 volumes on open access shelves.
Appendix 6.2

Summary of BIF class groups and teachers in FLLGP

a) CEFR class levels and learners’ ages

<table>
<thead>
<tr>
<th>Class Level: CEFR level and related Cambridge ESOL exam</th>
<th>Younger learners: Up to age 15</th>
<th>Adults: Age 16 and over</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1 – PET</td>
<td>2 classes</td>
<td>3 classes</td>
</tr>
<tr>
<td>Post B1/pre-B2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-PET / pre-FCE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2 – FCE</td>
<td>2 classes</td>
<td>3 classes</td>
</tr>
</tbody>
</table>

b) BIF classes

<table>
<thead>
<tr>
<th>COURSE ID</th>
<th>YL OR ADULT</th>
<th>CEFR LEVEL</th>
<th>Exam type and CERTIFICATION</th>
<th>HOURS in the course</th>
</tr>
</thead>
<tbody>
<tr>
<td>J4</td>
<td>YL (11-12 yrs)</td>
<td>B1</td>
<td>PET</td>
<td>90</td>
</tr>
<tr>
<td>J5</td>
<td>YL (11-12 yrs)</td>
<td>B2</td>
<td>FCE</td>
<td>90</td>
</tr>
<tr>
<td>M4</td>
<td>YL (13+ yrs)</td>
<td>B1</td>
<td>PET</td>
<td>120</td>
</tr>
<tr>
<td>M5</td>
<td>YL (13 + years)</td>
<td>B2</td>
<td>FCE</td>
<td>120</td>
</tr>
<tr>
<td>E3</td>
<td>ADULT</td>
<td>B1</td>
<td>BIF E3 Exam or PET</td>
<td>120</td>
</tr>
<tr>
<td>E4</td>
<td>ADULT</td>
<td>B2</td>
<td>BIF E4 Exam</td>
<td>120</td>
</tr>
<tr>
<td>FC5</td>
<td>ADULT</td>
<td>B2</td>
<td>FCE</td>
<td>120</td>
</tr>
</tbody>
</table>
### Teachers and BIF Class Groups

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
</table>
| **TEACHER A – LW – female** | FLLGP GROUP 1 | N=11  
BIF CLASS - FC5 A - ADULT Pre-FCE (CEFR B2) |
| **TEACHER B - PD – male** | FLLGP GROUP 2 | N= 10  
BIF GROUP - FC5 B - ADULT Pre-FCE (CEFR B2) |
| **TEACHER C – JG – female** | FLLGP GROUP 3 | N=12  
BIF GROUP - FC5 C ADULT - FCE PREP (CEFR B2) |
| **TEACHER D - JD – male** | FLLGP GROUP 4 | N=10  
BIF GROUP - M5 – YL (over 13 years) - FCE PREP (CEFR B2) |
| **TEACHER E – AL – male** | FLLGP GROUP 10 | N=12  
BIF GROUP - E4 D – ADULT - Post-PET / Pre-FCE (CEFR B2) |
| **TEACHER F - VR – female** | FLLGP GROUP 6 | N=12  
BIF GROUP - J4 A – YL - PET PREP (CEFR B1) |
| **TEACHER G – HB - female** | FLLGP GROUP 7 | N=11  
BIF GROUP J4 B -YL - PET PREP (CEFR B1) |
| **TEACHER H – AC - female** | FLLGP GROUP 8 | N=12  
BIF GROUP E3 A – ADULT - PET PREP – (CEFR B1) |
| **TEACHER I – DB - male** | FLLGP GROUP 9 | N=12  
BIF GROUP E3 B – ADULT - PET PREP (CEFR B1) |
d) Comparisons by age, target level, exam type and teacher

**YL Group – Cambridge exam preparation – target level and teacher variables**

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Exam Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher D</td>
<td>FCE (B2)</td>
</tr>
<tr>
<td>Teacher E*</td>
<td>FCE (B2)</td>
</tr>
<tr>
<td>Teacher F</td>
<td>PET (B1)</td>
</tr>
<tr>
<td>Teacher G</td>
<td>PET (B1)</td>
</tr>
</tbody>
</table>

Teacher E also taught an adult, non-Cambridge B2 group

**Adult Groups – Cambridge exam preparation – target level and teacher variables**

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Exam Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher A</td>
<td>FCE (B2)</td>
</tr>
<tr>
<td>Teacher B*</td>
<td>FCE (B2)</td>
</tr>
<tr>
<td>Teacher B*</td>
<td>PET (B1)</td>
</tr>
<tr>
<td>Teacher C*</td>
<td>FCE (B2)</td>
</tr>
</tbody>
</table>

Teacher B taught two Cambridge, adult groups – one FCE, one PET.
Teacher C also taught a non-Cambridge, adult group at B2 level.

**Adult Group – Non-Cambridge exam**

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Exam Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher C*</td>
<td>BIF exam B2</td>
</tr>
<tr>
<td>Teacher E*</td>
<td>BIF exam B2</td>
</tr>
<tr>
<td>Teacher H</td>
<td>BIF exam B1</td>
</tr>
<tr>
<td>Teacher I</td>
<td>BIF exam B1</td>
</tr>
</tbody>
</table>

e) Summary of classes taught by same teacher (two classes each)

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Same age group</th>
<th>Different Cambridge exam level</th>
<th>Adult FCE (B2) group</th>
<th>Adult PET (B1) group</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix 6.3

Data Summaries and Reports

- **...** = Quantitative data
- **....** = Semi-quantitative data
- **....** = Potential or partially quantitative data
- **.....** = Video recordings

- BRITISH INSTITUTE FLORENCE (BIF) STUDENT GROUPS, WITH LEVELS, TARGET EXAMS, EL TEACHERS
- BIF PROJECT MEMO - OCTOBER 2002
- STUDENT PERMISSION SLIPS
- PARTICIPANT GROUP PROFILE (SUMMARY FROM STUDENT QUESTIONNAIRES PREPARED BY PROJECT COORDINATOR)
- PROJECT COORDINATOR - DATA SUMMARY MATRICES: JANUARY AND APRIL 2003 (VISIT 2 AND VISIT 30
- PROJECT COORDINATOR REPORTS TO CAMBRIDGE (10/02, 10/02 -5/03)
- PROJECT COORDINATOR – FIRST VISIT REPORT - NOVEMBER 2002
- PROJECT COORDINATOR - MARCH 2003 ACTION PLAN,
- PROJECT COORDINATOR - MARCH, MAY 2003 VISIT REPORTS AND ACTION SUMMARY,
- NOVEMBER 2002 COMMUNICAT RESULTS (N=91) COMMUNICAT 2 (n=6)
- 26/11/02 BIF MEMO ON COMMUNICAT RESULTS
- JUNE 2002 PET RESULTS
- JUNE 2003 PET AND FCE RESULTS (N=72)
- BIF PLACEMENT TEST PAPER AND ANSWER SHEET
- BIF ADULT PLACEMENT TEST PAPER AND ANSWER SHEET
- PROJECT COORDINATOR BIF PARTICIPANT ORAL ENGLISH RECORDINGS FRAME, RESULTS AND ANALYSES (N=60)
- LLQ DESCRIPTION 10/01
- EXCEL DATA - SUMMARY SHEET (1)
- EXCEL DATA - SUMMARY SHEET (2)
- PROJECT COORDINATOR REPORT PADS (WITH HIM)
- LLQ FORM FEEDBACK SUMMARY
- RAH TEACHER QUESTIONNAIRES (N=8) AND TEACHER QUESTIONNAIRE DATA SUMMARY
- MID-YEAR FCE MOCK LISTENING EXAM SCRIPTS = READING + WRITING + USE
OF ENGLISH + SCHEDULE

- PROJECT COORDINATOR FLLGP STUDENT ORAL ENGLISH AUDIO CASSETTE LIST

- LLQ RESULTS
  - COMMUNICAT SCORES
  - ST SELF-RATINGS
  - PET + FCE MOCK RESULTS
  - PET + FCE RESULTS
  - BIF INTERNAL EXAM RESULTS (PLACEMENT, MOCK, FINAL)
Appendix 6.4

Summary of the 11 FLLGP/BIF class groups

...... = Quantitative data
...... = Semi-quantitative data
...... = Potential or partially quantitative data
...... = Video recordings

FLLGP GROUP 1
BIF CLASS - FC5A - Pre-FCE
TEACHER A – LW – female (N=11)

REGISTRATIONS (N=10)
PLACEMENT TEST SCRIPTS (N=4)
REGISTER
TEACHER LOG
CLASSROOM VIDEOS AND ANALYSES X 3
ST QUESTIONNAIRES (1) (N=11)
ST SELF-RATING
WRITING WORK 10/11/02 (N=37); 3/03 (n=10); 5/03 (n=9)
FCE MOCK MARKS 3/03
FCE MOCK SCRIPTS INCL. READING AND WRITING 2/03 (n=10)
BIF FINAL INTERNAL EXAM 6/03 (n=7), WRITTEN (n=7)

FLLGP GROUP 2
BIF GROUP - FC5B PreCE
TEACHER B - PD – male (N=10)

REGISTER
PLACEMENT TEST SCORES (n=4)
T LOGS, 02 AND 03
CLASSROOM VIDEOS X 3
LLQ (n=6)
FCE MOCK WRITING 2/03 (n=9)
FCE MOCK LISTENING 2/03 (n=8) (paper + scores)
FCE MOCK SCORE SHEETS (n=9)

FLLGP GROUP 3
BIF GROUP - FC5C FCE PREP
TEACHER C – JG – female (N=12) – First Group

REGISTRATIONS 9(N=12)
PLACEMENT TEST SCORES (n=4)
ST QUESTIONNAIRES (1) (n=10)
ST SELF-RATINGS 10/02 (n=10)
LLQ (n=11)
T LOGS
CLASSROOM VIDEOS X 2
ST WRITING 10/02 (n=11); ST WRITING (?) (N=16), ST WRITING 3/03 (n=11), 4/03 (n=6)
ST APPLICATION LETTERS (n=11), ST FILM REVIEW (N=9)
FCE MOCK 3/03 (n=11) WITH RESULTS
6/03 INTERNAL LISTENING EXAM (n=5), 6/03 INTERNAL WRITTEN EXAM (n=5)

FLLGP GROUP 4
BIF GROUP M5 - FCE PREP YL
TEACHER D - JD – male (N=10)

REGISTRATIONS
TEACHER LOG
CLASSROOM VIDEOS X 2
STUDENT SELF-RATINGS 10/02 (N=10)
ST QUESTIUONNAIRE (1) 2/03 (n=7)
FCE MOCK 3/03 RESULTS (N=10)
LLQ 11/03 (n=3)

FLLGP GROUP 5
BIF GROUP J5 - FCE PREP
TEACHER E – AL – male (N=12) - First Group

REGISTRATIONS (PLACEMENT TEST (n=1)
T LOG (MID-YEAR)
CLASSROOM VIDEO
ST QUESTIONNAIRES (1) (N=12; (2) 5/03) N=7
LLQ (10/02) (N=12)
SELF-RATINGS
FCE MOCK RESULTS (N=12) SCRIPTS, WRITING AND LISTENING (N=12) +
EXAM PAPER
ST WRITING (23/03 (n=8)
ST WRITING 3-5/03 (N=13
ST WRITING 4/03
ST WRITING 5/03 (N= 13
ST WRITING 4/03 (n=11)

FLLGP GROUP 6
BIF GROUP J4A - PET PREP - YL
TEACHER F - VR – female (N=12)

REGISTRATIONS (N=12)
REGISTER
MAIL SELF-RATINGS 10/02 (N=12)
LLQ (N=12)
TEACHER LOG
CLASSROOM VIDEOS X 3
ST QUESTIONNAIRES (1) (N= 12)
MOCK PET RESULTS 2003 (n=10)
ST WRITING 11/02 (n=6)
ST WRITING 2/03 (n=11)
ST WRITING 4/03 (N=12)
ST WRITING 5/03 (n=5)
BIF M3/J4 WRITTENN EXAM 6/03 SCRIPTS (n=3)

FLLGP GROUP 7
BIF GROUP J4B - PET PREP - YL
TEACHER G – HB - female (N=11)

REGISTRATIONS
MID-YEAR T REPORTS + GRADES (N =11)
CLASSROOM VIDEOS X 2
MID-YEAR (3 2003) ST WRITING SAMPLES (N=9)
STUDENT QUESTIONNAIRES (2) (N=6) STUDENT QUESTIONNAIRE (1) (N=1)

FLLGP GROUP 8
BIF GROUP E3A - PET PREP
TEACHER H – AC - female (N=12)

REGISTRATION
REGISTER
ST SELF-RATINGS 10/02 (N=12)
ST QUESTIONNAIRES (1) (N=12), (2) (N= 9)
ST WRITING 1/03 (N=7)
MOCK EXAM SCORES, SCRIPTS 5/03 (n=11)
ST LETTER WRITING SAMPLE 11/02 AND 5/03 (n=11)
FINAL WRITING EXAM 5/03 (N=12), + RESULTS

FLLGP GROUP 9
BIF GROUP E3B - PET PREP
TEACHER I – DB - male (N=12)

REGISTRATION
REGISTER
PET MOCK RESULTS 02/03 (n=9)
T PROGRESS REVIEW 12/02 (n=9)
T LOGS
CLASSROOM VIDEOS X 2
PROGRESS TEST RESULTS 1/03 (n=11)
TEST SAMPLE ANSWERS
ST WRITING SAMPLE 10/02(n= 9) 2/03 (n=6), 5/03 (n=2)
ST SELF-RATING 10/02 (N=12)
LLQ10/02 (N=10)
ST QUESTIONNAIRE (1)(N=12); (2) (n=10)
ST MOCK PET, INCLUDING WRITING 3/03 (N=7); 4/03 (n=8)
PROGRESS TEST 2 SCORES 4/03 (n=9)
FINAL E3 WRITTEN EXAM SCORES 6/03 (n=10) + RESULTS 9N=12

FLLGP GROUP 10
E4D BIF GROUP E4D - Post-PET / Pre-FCE
TEACHER E - AL (N=12) – Second Group
REGISTRATIONS
REGISTER
PLACEMENT TEST SCRIPTS (n=5)
T MID-YEAR NOTES
CLASSROOM VIDEOS X 2
LESSON PLAN
ST SELF-RATING (n=9)
LLQ (N= 6)
ST QUESTIONNAIRES (1) (n=6); (2) (n=6)
BIF FINAL INTERNAL EXAM RESULTS 3/4/03 (N=11)
BIF FINAL EXAM SCRIPTS AND MARKS (n=3)
WRITTEN (n=10) LISTENING (n=10)

FLLGP GROUP 11
BIF GROUP E4E – Post-PET / Pre-FCE
TEACHER C – JG female (N=12) – Second Group

TEACHER’S LOG
REGISTRATIONS (n=9) (INCL. BACKGROUND, YRS OF ENGLISH, PREVIOUS COURSES, PLACEMENT TEST SCORES, ORAL ASSESSMENT SCORES)
ST WRITING (WITH ERROR MARKINGS) SCORED BY T AND RAH (n=10)
NB DETAILED ST PROFILES ON EACH VISIT FROM T (RAH NOTES)
CLASSROOM VIDEOS X 2
ST SELF-RATINGS (N=10)
ST QUESTIONNAIRES (1) (N=10)
BIF FINAL INTERNAL EXAM SCRIPTS AND SCORES
6/03 WRITTEN EXAMS E4: N=8; LISTENING N=8 RESULTS
PLACEMENT TEST SCRIPTS
FOR THIS GROUP DETAILED STUDENT PROFILES WERE PROVIDED FOR EACH VISIT BY THE TEACHER AND THE PROJECT COORDINATOR (RH) MADE NOTES
Appendix 6.5
Summary of responses to Teacher Questionnaire - including teacher profiles

<table>
<thead>
<tr>
<th>Names:</th>
<th>9 TEACHERS altogether but only 8 completing the profile (3 covering two of the target groups):</th>
</tr>
</thead>
</table>
|        | TEACHER A – LW – female  
|        | N=11  
|        | FLLGP GROUP 1 - BIF CLASS - FC5A - Pre-FCE |
|        | TEACHER B - PD – male  
|        | First Group N=10  
|        | FLLGP GROUP 2 - BIF GROUP - FC5B PreCE |
|        | Second Group N =  
|        | BIF GROUP E3G; ADULT PET / BIF EXAM GROUP |
|        | TEACHER C – JG – female  
|        | First Group N=12  
|        | FLLGP GROUP 3 - BIF GROUP - FC5C FCE PREP |
|        | Second Group N=12  
|        | FLLGP GROUP 11 - BIF GROUP E4E – Post-PET / Pre-FCE |
|        | TEACHER D - JD – male  
|        | N=10  
|        | FLLGP GROUP 4 - BIF GROUP M5 - FCE PREP YL |
|        | TEACHER E – AL – male  
|        | First Group N=12  
|        | FLLGP GROUP 10 - BIF GROUP E4D - Post-PET / Pre-FCE |
|        | Second Group N=12  
|        | FLLGP GROUP 5 - BIF GROUP J5 - FCE PREP |
|        | TEACHER F - VR – female  
|        | N=12  
|        | FLLGP GROUP 6 - BIF GROUP J4A - PET PREP - YL |
|        | TEACHER G – HB - female  
|        | N=11  
|        | FLLGP GROUP 7 - BIF GROUP J4B - PET PREP - YL |
|        | TEACHER H – AC - female  
|        | N=12  
|        | FLLGP GROUP 8 - BIF GROUP E3A - PET PREP |
|        | TEACHER I – DB - male  
|        | N=12  
|        | FLLGP GROUP 9  
<p>|        | BIF GROUP E3B - PET PREP |
|        | Teacher G - YL specialist and teacher trainer wouldn’t complete questionnaire or be filmed |</p>
<table>
<thead>
<tr>
<th>Forms of address:</th>
<th>Miss = 2</th>
<th>Mrs = 2</th>
<th>Mr = 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ages:</td>
<td>20-30 = 1</td>
<td>31-40 = 4</td>
<td>41-50 = 3</td>
</tr>
<tr>
<td>Qualifications:</td>
<td>MSc = 1</td>
<td>MA = 1</td>
<td>BA = 2</td>
</tr>
<tr>
<td>Institution(s) of work:</td>
<td>BIF: 8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of years teaching English:</td>
<td>3, 10, 11, 12, 5, 23, 15</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>How often you normally implement the following student activities in your English classes. Please feel free to add comments in the right-hand column.</th>
<th>Frequently</th>
<th>Quite Often</th>
<th>Occasionally</th>
<th>Never</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. discussions in small groups</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>2. listening and taking notes</td>
<td>3</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>3. writing grammar exercises</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>Trouble shooting only; hw;</td>
</tr>
<tr>
<td>4. discussions with the whole class</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>5. listening to the teacher talking to the whole class</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6. watching videos</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>Private viewing encouraged</td>
</tr>
<tr>
<td>7. discussions with a partner</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>8. writing short notes</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>CAE only;</td>
</tr>
<tr>
<td>9. taking practice exams</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>0</td>
<td>Varies;</td>
</tr>
<tr>
<td>10. writing letters</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11. reading then writing answers to questions</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>HW</td>
</tr>
<tr>
<td>12. doing vocabulary exercises</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>13. writing compositions</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>Exam prac; HW 2; not t exam classes;</td>
</tr>
<tr>
<td>14. reading story books</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>Private //;</td>
</tr>
<tr>
<td>15. listening and choosing answers to questions</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>16. discussing exams</td>
<td>0</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>Depends;</td>
</tr>
<tr>
<td>17. using computers</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>CDr;</td>
</tr>
<tr>
<td>Activities</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Modernising the teaching and learning of foreign languages</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Language teaching through modules with clear objectives, numbers of hours, content, expected outcomes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Small homogeneous language learner groups</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Communication skills relevant to students’ language needs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. More self-access language learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Optimal use of modern technology in language learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. The use of external language exams at recognised Council of Europe levels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Improved teacher support such as resource centres and teacher professional development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### How well are the following objectives being achieved in the school

<table>
<thead>
<tr>
<th>Objective</th>
<th>Very well</th>
<th>Well</th>
<th>Not very well</th>
<th>Hardly at all</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Modernising the teaching and learning of foreign languages</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Language teaching through modules with clear objectives, numbers of hours, content, expected outcomes</td>
<td>3</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>Well established; but not in modules</td>
</tr>
<tr>
<td>Small homogeneous language learner groups</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>Max 12</td>
</tr>
<tr>
<td>Communication skills relevant to students’ language needs</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>More self-access language learning</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Optimal use of modern technology in language learning</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>C room (n=12)</td>
</tr>
<tr>
<td>The use of external language exams at recognised Council of Europe levels</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Camb centre (Tuscany)</td>
</tr>
<tr>
<td>Improved teacher support such as resource centres and teacher professional development</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>DOS keen; lots of opps;</td>
</tr>
</tbody>
</table>

Please summarise here your most important beliefs and practices in English language teaching:

- NB comm (above acc); confidence building activities; acceptable pron; language taught and learnt in blocks; >> ss feel good about themselves when sp TL
- T high level skills e.g. guessing M of words in context; good pronunciation; making exam practice interesting;
- People learn best when genuinely motivated, quite relaxed, having fun but challenged enough and can see own progress
- To develop CALT & s-centred classes
- Aim to provide stimulating and/or interesting and effective teaching in a relaxed and supportive LL environment
- There may be wrong ways but there aren’t right ways!
- Balance authentic spoken comm and varied, interesting IP of TL. NB Interaction at all stages of a lesson
- Communicative, maximum exposure (actively and passively), continuous (progress) assessment, objective (standardised)

Please write here how you think your English language teaching and attitudes have changed over the past year:

- NB for Ts and ss, real; more focus on formulaic TL (collocations, vocab, integrated
with gr Teaching; "natural sounding E, break away from trad gr syll; CEF
  ▪ More faith in technology, more willing to use CD (e.g. FCE prac CDRom;
  ▪ PET examinership helps give different perspective to exam course; greater flexibility
    needed for wide range of ss
  ▪ Quite new to ELT; mainly YL, more aware of different approach needed
  ▪ No; should they have?
  ▪ NB the idea that practice makes perfect' applied to speaking work
  ▪ Take lead more from ss and feel that listening and assessing needs, preferences
    can help T plan better course, teach what they need to know, instead of this and
    what they know already well; , more flexible, open to adapting mats to suit ss needs
  ▪ No significant change
Appendix 6.6

Participant profile grid - the learners (N = 134)

Compiled by the Project Coordinator (April 2003)

<table>
<thead>
<tr>
<th>Classes</th>
<th>Adult</th>
<th>Young Learners</th>
<th>Intermediate</th>
<th>Upper Intermediate</th>
<th>KET</th>
<th>PET</th>
<th>FCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11</td>
<td>8</td>
<td>3</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Participant Gender</th>
<th>Male</th>
<th>Female</th>
<th>Ages 11-15</th>
<th>16-19</th>
<th>20s</th>
<th>30s</th>
<th>40s</th>
<th>50s</th>
<th>60s</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>48</td>
<td>86</td>
<td>33</td>
<td>22</td>
<td>28</td>
<td>22</td>
<td>11</td>
<td>10</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Home Languages</th>
<th>Italian</th>
<th>Catalan</th>
<th>Chinese</th>
<th>English</th>
<th>French</th>
<th>Japanese</th>
<th>Polish</th>
<th>Portuguese</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>111</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English Study</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>More</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>18</td>
<td>18</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>7</td>
<td>6</td>
<td>8</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Previous Tests, Exams Taken</th>
<th>KET</th>
<th>PET</th>
<th>BIF</th>
<th>Trinity</th>
<th>University</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>18</td>
<td>26</td>
<td>26</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Most Recent Previous Course</th>
<th>BIF regular</th>
<th>BIF summer</th>
<th>American Inst</th>
<th>University</th>
<th>Overseas</th>
<th>Private</th>
<th>Long gap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>72</td>
<td>6</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Future Intended Exams</th>
<th>KET</th>
<th>PET</th>
<th>FCE</th>
<th>CAE</th>
<th>CPE</th>
<th>BIF</th>
<th>TOEFL</th>
<th>Don’t know</th>
<th>No answer</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>18</td>
<td>56</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>27</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EFL Courses Abroad?</th>
<th>Yes</th>
<th>No</th>
<th>Where?</th>
<th>How long?</th>
<th>Using EL as a tourist</th>
<th>Yes</th>
<th>No</th>
<th>Where?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>40</td>
<td>73</td>
<td>UK &amp; Ireland</td>
<td>Most 2-3 weeks</td>
<td>82</td>
<td>19</td>
<td>UK, Ireland 55; Europe 35; USA16; Asia 5; Mid East 3; Africa 2</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>EL Used in Job?</th>
<th>Yes</th>
<th>No</th>
<th>Where?</th>
<th>What?</th>
<th>Present Jobs</th>
<th>Future Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>31</td>
<td>47</td>
<td>All occasional, as part of job, mainly locally, apart from 1; Yrs not included</td>
<td>accountant: 2; architect: 1; business: 2; civil servant: 2; dentist: 1; doctor: 3; hotelier: 3; IT: 1; lawyer: 1; librarian: 2; police: 1; priest: 1; researcher: 4; scientist: 1; secretary: 3; shopkeeper: 2; taxi driver: 1; 26 of the respondents are current school students; 15 are current university students</td>
<td>architect: 1; artist: 2; business: 7; dentist: 1; doctor: 5; engineer: 3; footballer: 1; interior design: 3; IT: 2; journalist: 2; lawyer: 9; pilot: 2; PR: 1; scientist: 2; psychologist: 1; singer: 1; social worker: 1; tourism: 1; writer: 1;</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>University</th>
<th>Yes</th>
<th>No</th>
<th>Study Areas (past or present)</th>
<th>Study Areas (future)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>346</td>
<td>Yes</td>
<td>Study Areas (past or present)</td>
<td>Study Areas (future)</td>
</tr>
</tbody>
</table>
### Show how often you did the following activities in your previous English language course.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Frequently</th>
<th>Quite Often</th>
<th>Sometimes</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. discussions in small groups</td>
<td>39</td>
<td>40</td>
<td>36</td>
<td>7</td>
</tr>
<tr>
<td>2. listening and taking notes</td>
<td>37</td>
<td>28</td>
<td>40</td>
<td>6</td>
</tr>
<tr>
<td>3. writing grammar exercises</td>
<td>50</td>
<td>39</td>
<td>20</td>
<td>1</td>
</tr>
<tr>
<td>4. discussions with the whole class</td>
<td>16</td>
<td>29</td>
<td>47</td>
<td>7</td>
</tr>
<tr>
<td>5. listening to the teacher talking to whole class</td>
<td>60</td>
<td>26</td>
<td>27</td>
<td>1</td>
</tr>
<tr>
<td>6. watching videos</td>
<td>2</td>
<td>2</td>
<td>59</td>
<td>45</td>
</tr>
<tr>
<td>7. discussions with a partner</td>
<td>52</td>
<td>25</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>8. writing short notes</td>
<td>25</td>
<td>45</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>9. taking practice exams</td>
<td>7</td>
<td>30</td>
<td>55</td>
<td>14</td>
</tr>
<tr>
<td>10. writing letters</td>
<td>17</td>
<td>11</td>
<td>60</td>
<td>12</td>
</tr>
<tr>
<td>11. reading then writing answers to questions</td>
<td>23</td>
<td>36</td>
<td>48</td>
<td>2</td>
</tr>
<tr>
<td>12. doing vocabulary exercises</td>
<td>16</td>
<td>39</td>
<td>53</td>
<td>8</td>
</tr>
<tr>
<td>13. writing compositions</td>
<td>13</td>
<td>32</td>
<td>51</td>
<td>13</td>
</tr>
<tr>
<td>14. reading story books</td>
<td>8</td>
<td>16</td>
<td>58</td>
<td>28</td>
</tr>
<tr>
<td>15. listening and choosing answers to questions</td>
<td>23</td>
<td>36</td>
<td>49</td>
<td>4</td>
</tr>
<tr>
<td>16. discussing exams</td>
<td>4</td>
<td>10</td>
<td>62</td>
<td>32</td>
</tr>
<tr>
<td>17. using computers</td>
<td>4</td>
<td>6</td>
<td>53</td>
<td>48</td>
</tr>
<tr>
<td>18. other activities in your English classes: drama, games</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Show how often you do the following activities in English outside your school.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Daily</th>
<th>Occasionally</th>
<th>Almost</th>
<th>Never</th>
</tr>
</thead>
<tbody>
<tr>
<td>a reading books</td>
<td>17</td>
<td>50</td>
<td>22</td>
<td>26</td>
</tr>
<tr>
<td>b newspapers, magazines</td>
<td>11</td>
<td>36</td>
<td>38</td>
<td>28</td>
</tr>
<tr>
<td>c writing letters</td>
<td>12</td>
<td>31</td>
<td>35</td>
<td>35</td>
</tr>
<tr>
<td>d using email</td>
<td>31</td>
<td>20</td>
<td>23</td>
<td>39</td>
</tr>
<tr>
<td>e watching TV</td>
<td>19</td>
<td>35</td>
<td>33</td>
<td>22</td>
</tr>
<tr>
<td>f listening to the radio</td>
<td>28</td>
<td>20</td>
<td>18</td>
<td>45</td>
</tr>
<tr>
<td>g using the Internet</td>
<td>30</td>
<td>30</td>
<td>12</td>
<td>31</td>
</tr>
<tr>
<td>h watching movies</td>
<td>18</td>
<td>44</td>
<td>28</td>
<td>20</td>
</tr>
<tr>
<td>i going to shows</td>
<td>5</td>
<td>16</td>
<td>25</td>
<td>58</td>
</tr>
<tr>
<td>j talking with visitors</td>
<td>13</td>
<td>55</td>
<td>30</td>
<td>14</td>
</tr>
<tr>
<td>k Other activities using English?</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   translating for others 2;
songs: 3;
talking to friends: 2;
| My English has improved ........... in the last year. | a. a lot 38  
| b. a little 37  
| c. very little 14  
| d. quite a lot 21  |
| My English language skill that has improved most is... | a. reading 27  
| b. listening 33  
| c. writing 22  
| d. speaking 51  |
| My main reason for learning English is ..... | a. to get a better job 31  
| b. for holidays abroad 14  
| c. because I like it 33  
| d. because it is international 47  |
Appendix 6.7
Learner Profiles based on teacher logs

These learner profiles were taken from the teacher logs and were summarised by the Project Coordinator (May 2003). They have been abridged/anonymised by the current author.

<table>
<thead>
<tr>
<th>Teacher A – LW</th>
<th>BIF Class: FC5A - ADULT FCE GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ages:</td>
<td>Gender:</td>
</tr>
<tr>
<td>20s: 4</td>
<td>Male: 1</td>
</tr>
<tr>
<td>50s: 2</td>
<td>Female: 9</td>
</tr>
<tr>
<td>30s: 3</td>
<td>Which language at home:</td>
</tr>
<tr>
<td></td>
<td>Italian 11</td>
</tr>
<tr>
<td></td>
<td>French 1</td>
</tr>
<tr>
<td></td>
<td>Japanese 1</td>
</tr>
<tr>
<td>Years of English</td>
<td>At which school:</td>
</tr>
<tr>
<td>3 years: 2</td>
<td>elementary: 2</td>
</tr>
<tr>
<td>4 years: 3</td>
<td>middle: 5</td>
</tr>
<tr>
<td>5 years: 3</td>
<td>high: 5</td>
</tr>
<tr>
<td>9 years: 1</td>
<td>college: 1</td>
</tr>
<tr>
<td>10 years: 1</td>
<td>BIF: 5</td>
</tr>
<tr>
<td>13 years: 1</td>
<td></td>
</tr>
</tbody>
</table>

"Nice group; very hard-working, doing extra work voluntarily, together". T will take FCE in June 03. T uses phonetics, re-visits and re-cycles new words using post-it slips; gives both compulsory and optional writing homework.

<table>
<thead>
<tr>
<th>Learner 1 – female EG</th>
<th>43; 4PC</th>
<th>Probably won’t take FCE in June 03; domestic problems; has taken extra year, i.e. G4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 2 – female FM</td>
<td>33; ?PC</td>
<td>Back from London with increased confidence</td>
</tr>
<tr>
<td>Learner 3 – female SR</td>
<td>23; U student; 2PC</td>
<td>Needs the extra year, but should pass FCE; Tries hard; U student but not academic; speaking better, sells herself well, with good pronunciation;</td>
</tr>
<tr>
<td>Learner 4 – female CF</td>
<td>28; U student</td>
<td></td>
</tr>
<tr>
<td>Learner 5 – female UP</td>
<td>32; 1PC; went to UK for three weeks</td>
<td>Making big progress (5/03); does homework, pushed by group;</td>
</tr>
<tr>
<td>Learner 6 – female AN</td>
<td>29; civil servant;</td>
<td>Left group FOR Master’s studies.</td>
</tr>
<tr>
<td>Learner 7 – male FC</td>
<td>?; 3PC</td>
<td></td>
</tr>
<tr>
<td>Learner 8 – female MD</td>
<td>54; 4PC</td>
<td>Lacks TL exposure; Improving 5/03; has taken extra year, i.e. G4</td>
</tr>
<tr>
<td>Learner 9 – female CS</td>
<td>23; 2PC</td>
<td></td>
</tr>
<tr>
<td>Learner 10 – female MK</td>
<td>34; interpreter;</td>
<td>Writing improving (5/03), spelling more competent; LC, RC good</td>
</tr>
</tbody>
</table>
Learner 11 – female RS

Had taken extra year, i.e. G4; dropped out but will return for CAE

Teacher B – PD (First Group) BIF Class: FC5B - ADULT FCE GROUP

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Which language at home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>20s: 6</td>
<td>Male: 1</td>
<td>Italian: 9</td>
</tr>
<tr>
<td>30s: 1</td>
<td>Female: 8</td>
<td></td>
</tr>
<tr>
<td>40s: 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50s: 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 year : 2</td>
<td>middle: 2</td>
</tr>
<tr>
<td>3 years: 2</td>
<td>high: 4</td>
</tr>
<tr>
<td>5 years: 1</td>
<td>BIF: 2</td>
</tr>
<tr>
<td>8 years: 1</td>
<td>American Inst:1</td>
</tr>
<tr>
<td>9 years: 2</td>
<td></td>
</tr>
<tr>
<td>10 years: 1</td>
<td></td>
</tr>
</tbody>
</table>

November 02: All enjoy talking about English and Italian culture, like travel and have all been to an English-speaking country, on holiday or to study; interested in music, cinema, reading, food / wine. English for pleasure and travel; work-related reasons secondary, tho the dance teacher and the musician hope to live in England at some stage and the taxi driver hopes to work as a private tour guide / chauffeur

Enjoy open class work, like typical FCE course book activities, pair work, discussions; good intra group relations, very high M. All but RO doing FCE after only one year from PET level; Lesson structure dictated by the course book but a lot of open class conversation and pair work discussions user as warmers. Vocab revision games and re-cycling activities and songs to break up the course book. E Regular visits to computer room for CD Rom work (First Cert grammar Rom) and Internet project work (when time allows); Class has a weakness, lack of adequate vocabulary range necessary at the required level. I identified this just before Christmas and made a decision to deal with this in the New Year.
March 03: Gradual and steady progress since NY, excellent M maintained; Attendance good apart from DP and RO; Course book lessons predominate due to syllabus coverage concerns. Core members from teacher Peter's PET class from previous year (AF, RO, RR, GS); new members mainly conform to the “group mould” – studious and book bound approach to learning; remain the dominant feature and chosen pattern. Homogeneous group with good study skills. With exception of RO and DP, structured, book-oriented approach favoured, homework revolving round a religious attention to detail. Written work predominantly in the form of formal/informal letters and compositions standard of written work very good. Listening remains a challenge and SDM and RR face particular problems in dealing with the rate of quick delivery, colloquial listening texts tested in FCE past papers. After Christmas I presented group with set of individual “Vocab books”. 30 minutes heightening awareness of lexis issues and discussing the importance of “remembering” words and what “knowing a word involves. We discussed ways of recording vocabulary and we drew up a list of lexical areas from the course book and my own list. Homework was to divide up their “Vocab hook” and then to transfer “old” vocab into it and keep up the process. It is noticeable that those students who have kept to the system have a wider capacity to recycle and use vocab learnt during the course. The students who have stuck to a systematic approach to vocab recording are AF, GB, AMP and RR.

Their “Use of English” skills seem to have improved a little and LC and conversation has improved noticeably in the case of AF, G and AMP. Rather strange habit of resorting to Italian in class and in particular when speaking to me! DP is the main culprit but also GT (due to shyness) tends to use Italian. AF, AMP have this habit to a lesser extent. The amount of Italian used in class has grown and I may have to take some blame for not doing anything about this. I may also have erred a little in this area. Having a coffee in the break is usually “in Italian” and the habit has spilled over a bit into class. Laziness is the main reason and I am in the middle of correcting this habit. Of all the skills, speaking has suffered in terms of classroom time. The ruthless nature of following the course book and correcting homework and dealing with the variety of exam skills has meant a reduced amount of time for speaking. I am trying to redress the balance as I am very satisfied with the “mock” results and will soon be working in a systematic way on Paper 5 speaking skills in class.

Outside conversation practice: I dedicated a whole lesson to introducing students to an Irish girl called K (an Estate Agent) who was studying at BIF for two months. I asked K to come in and give a 15 minute presentation about herself and her job and her reasons for learning Italian. After this, we had a “questions time” and then went to have coffee in the break. I asked K to talk at a normal pace (a little slower perhaps) and to use natural stress patterns and intonation and not try to be a “teacher” giving a lecture. My students really enjoyed the experience and learnt some interesting new vocab. The point of the exercise was not just a listening task but also to arrange out of class conversations exchanges with K and my students. This worked very well and 5 of my students met with K for two hours on a few occasions. The first hour was “in Italian” and was as BIF. My students gave grammar explanations to K and helped her with her Italian pronunciation. The “English hour” was held in the bar outside where my group of 2 or 3 students chatted to K in English about a range of topics.

Result: In terms of confidence and realising how much they could say/understand (even if it meant comparing themselves favourably in English to the limited capacity of K in Italian!) the exercise was a huge success. These students who met up for 2 or 3 sessions (AF/AMP/GB) got a real psychological lift. Even if their vocab remained roughly the same and their fluency is very much the same too after only 3 sessions- their ability and skills and confidence to cope with a conversation with a person they don't know has increased. At the same moment, I found a conversation partner for RR for 2 weeks. SDM too shy to accept K's offer; and indeed her lack of confidence was cause for concern; previous experience of Florence LS left her struggling, limited vocab, lack of confidence in speaking, still struggling.
| Learner 1 – male DP | 23; student; "wild card"; good colloquial knowledge of vocab; great speaking; a bit unpredictable; gaps in grammar | The ‘non-academic’ learners. Both DP and RO strong in speaking skills and have good vocabulary and listening skills. Both spent 6 weeks in Edinburgh on BIF scholarship Summer English course programme. Wild time at the Festival, English lessons at the University, and learnt a lot of slang in the surrounding pubs. However neither are adapted to “serious study”. Hard to get any written work from either of them and although DP’s lack of attendance and work record is a bit of a concern. Wants to take FCE but often absent, finds homework a chore. Has gift for language and yet a lack of a systematic approach. I will need to have few words when I give his ‘mock’ results back. He remains motivated and cheerful at all times! | Does almost no written work |
| Learner 2 – female GT | 27; assistant at BIF; | Joined BIF Dec 02. Excellent FCE level student and particularly good pronunciation. Confident but a little conscious about using her English. “Big” progress; more TL exposure at BIF office? |
| Learner 3 – female SDM | 24; maths student; v shy but excellent in written work and grammar; speaking less confident; SDM lacks confidence and feels that AMP and AF are better than her. Her written work remains one of the best but her listening and speaking needs practice (lack of vocab is her main weak point). |
| Learner 4 – male RO | 40; taxi driver; Danish girlfriend and scholarship award to Edinburgh give him a head start in speaking; v fluent; grammar not yet FCE level; will not take exam; | RO NOT doing FCE and is only in the class for social and “basic English” reasons, |
| Learner 5 – female RR | 59; retired primary school T; vg FCE level student; fluent speaker with good writing skills; good grammar | RR is the more “mature” student. Speaking skills and written skills are very good but over the last 3 months has improved perhaps the least; various commitments at home and has missed some lessons. She finds the FCE type of task a challenge but her reading skills and vocabulary arc by far the best/widest in the class. Her passive and active vocabulary is by far the best in the class. She loves reading and listening to readers with cassettes. |
| Learner 6 – female CO | Maths student; good solid all-rounder; leaving class soon; |
| Learner 7 – female AMP | 32; literature and philosophy student; speaks Swedish and does Scandinavian studies; good pronunciation and vg grammar; a touch shy but a good all-rounder. AMP and others regularly meet an hour before class in the Study Centre to check and go over homework. AF/AMP and GB Have improved the most and are very studious. |
### Learner 8 – female  
**AF**  
26; dance teacher; confident easy going (tho’ objected to last video-ing); good all round skills, one of the best at speaking, lacks vocab but good pronunciation and fluency; AF and others regularly meet an hour before class in the Study Centre to check and go over homework  
doesn’t like the video

### Learner 9 – female  
**GB**  
21; Economics student; shy but v friendly; good language control but speaking lags a bit behind other skills; GB, AF, AMP and others regularly meet an hour before class in the Study Centre to check and go over homework  
GB has improved the most.

### Learner 10 – female  
**GS**  
Later joiner? Jan 03; economics graduate; joined Jan 03; student in my PET level class last year. Weak Intermediate, just passed PET; 6 months in Holland on an Erasmus University project. speaking skills have improved enormously and jump in her spoken level of English and colloquial vocab noticeable. Lots of English speaking friends in NL and attended lectures in English. Feels “behind” in FCE terms has missed a large chunk of the course but finds Listening task of FCE exam very difficult (more so than others) and her Use of English ‘mock’ pretty disastrous. It seems that in terms of FCE skills the “English abroad” has been of limited benefit. she has also lost some of her initial confidence when she arrived fresh faced and fluent from 6 months abroad.

**Netherlands; ; big progress but grammar?**

### Learner 11 female  
**MM**  
Joined she has transferred from Teacher F’s FCE class; her spoken and listening skills good but use of English is a bit ropey. she joined us 2 chapters behind in the book and has found it a struggle to catch up. Her ‘mock’ was good though and she is settling well. She feels a little behind and needs extra encouragement.

---

<table>
<thead>
<tr>
<th><strong>Teacher B – PD</strong> (Second Group)</th>
<th>BIF Class: E3G - ADULT PET / BIF EXAM GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Age:</strong></td>
<td><strong>Gender:</strong></td>
</tr>
<tr>
<td>Under20: 1</td>
<td>Male: 6</td>
</tr>
<tr>
<td>20s: 3</td>
<td>Female: 3</td>
</tr>
<tr>
<td>30s: 3</td>
<td></td>
</tr>
<tr>
<td>40s: 2</td>
<td></td>
</tr>
<tr>
<td><strong>Years of English:</strong></td>
<td><strong>At which school:</strong></td>
</tr>
<tr>
<td>1 year: 2</td>
<td>high school: 2</td>
</tr>
<tr>
<td>2 years: 2</td>
<td>Uni: 1</td>
</tr>
<tr>
<td>3 years: 2</td>
<td>BIF: 5</td>
</tr>
<tr>
<td>4 years: 1</td>
<td></td>
</tr>
<tr>
<td>5 years: 1</td>
<td></td>
</tr>
</tbody>
</table>
A strong group of Intermediate speakers with a couple of exceptions. All enjoy *English File 3* book based lessons but also love IT and Conversation lessons and using songs in class. In fact due to their wide knowledge and interest in English music from 60s and 70s the use of songs as part of language and lexical development and consolidation is a popular choice and a regular feature used within the course.

Most students have been at BIF before which means that such features as pair and group work come naturally to them. They really enjoy speaking in English to me and to each other as much as possible. They are also committed to 100% communication between themselves in English (more so than my FCE group!).

They are mostly young, professional people who live and work in Florence. They get on well and like to make fun of each other. The atmosphere is very relaxed and even more laid back than in my FCE class. Both of my LGP groups have wonderful students, but this group likes to chill out more and the lack of an Exam at this stage in the year is noticeable. I imagine they are good enough to do well at PET level. However, depending on ‘Mock’ results in next couple of weeks, only 2 or 3 students are likely to take the PET exam in June. Work commitments and their belief in the lack of tangible benefits of having PET means they will choose to wait for FCE. None of these students have the need for an immediate Cambridge qualification.

**Lively, sociable group, “non-academic”, not exam-motivated** - they decide whether they take exam or not; teacher doesn’t push PET; group like a “social club”; v communicative; group behind re the course book because they do “other things”, e.g. songs every two weeks, e.g. analysing, translating Beatles songs or videos; 2 of the group doing PET; most will “wait for” FCE, e.g. for Erasmus purposes

| Learner 1 – male | 31; Big on commitment.; religiously does homework, good at grammar and wider vocabulary than most.; slightly strange and forced way of speaking sometimes but very accurate and quite fluent; gradual and steady improvement over the last 3 or 4 months; huge fan of conversation exchanges / parties and would happily talk the hind leg off any English speaking donkey (blond and pretty preferred); class joker and teaser; well liked but sometime O.T.T; family favourite. | v fluent |
| Learner 2 – male | 17; Secondary school student; one of the strongest but quiet and laid back; doesn’t speak much in open class; accurate and fluent tho’ vocab lacking; studies E at school, grammar and writing good; follows “Coldplay” LC very good | “will walk PET” |
| Learner 3 – male | 27; Quietly spoken and less accurate than Learner 1 (DC); good grammar and vocabulary though. Grammar wise one of the strongest students Likes pair work but a little shy in open class; fluency limited before Christmas but over last months this has improved noticeably. | |
| Learner 4 – female | 23; economics student Nicknamed “the woman of doubts” by friend Learner 1 (DC). Conscientious student who likes to check all her work and is indeed “full of little doubts” when doing class exercises. However, her little doubts pay off as she is generally accurate in her language work. She likes to process things more slowly but is nevertheless very thorough. Shyness makes her fluency seem less evident than in other students. However she is accurate. | “woman of doubts”, grammar quite accurate, but shy |
| Learner 5 – male | 34; student; Very nice guy, doing E3 for second year; did E2 class with T 2 years ago; weakest student in class, English motivation seems to be more for social reasons; lacked commitment, concentration last year (according to his T). This year great class dynamics and the fact that T has made an effort to include him and give him extra attention has helped him settle; has done 50% of *English File* 3 before but still struggle at times; speaking has improved since Christmas but remains weak for this level. Loves conversation exchanges/parties. | Repeating PET; there to socialise (o-o) |
| Learner 6 – male | 47; civil servant; V confident and fluent; leads role plays with confidence; Works in a State office. Mad about London and British 60s music; Dad from Naples, transferred to Malta; AS grew up in Malta speaking English, went to an English speaking school; very confident speaker, without doubt the most fluent speaker; pronunciation very good indeed, however makes a surprising number of spoken errors and many of these scented engrained and not easy to wash out.; above average in grammar, no more; vocabulary however is excellent. | |
| Learner 7 – female | 32; biologist; researcher in Biology at University; very keen good “all rounder”, level wise one of the best; good pronunciation and is very conscientious; enjoys speaking although quite shy to begin with.; very much come out of her shell and has shown herself to he very confident and determined. | Missed quiet a few lessons; confident? |
| Learner 8 – male | 32; primary school teacher, ex-law student; strong E2 student last year, struggled to start with in October in E3; T’s strongest E2 student but found the “cut and thrust” of the class personalities a bit daunting; new students from outside the BIF (Learner 2 and Learner 6 were strong) and other E2 students all knew each other well. She is shy, speaking not strong. However as the course has progressed, her language awareness and grammar control have improved significantly.; now very much level of the others although her speaking skills still lag behind a little. Whereas the others have improved slowly and gradually, Francesca has perhaps improved the most — even if this means she has “caught up” a little in terms of the level of the other students. | |
| Learner 9 – male | 46; chemist, very precise; | |
| Learner 10 – male | 23; economics student; The “wide boy”; leant the word “bunk” off" re his school days; but quite traditional in learning style; dictionary constant for looking up words and v little communication in class with colleagues; v studious seems bored and not personal in expressing ideas; potentially disruptive? Generally co-operative, polite takes part in p[air work, answers all Qs in open class; sense of humour | Will take PET |
| Learner 10 – female | Architect; Late, non-P student (jan 03); E3 repeater; “scatty”; independent; speaking good, LC above average; grammar confused? Lapses in concentration. | |
### Teacher C – JG (First Group)

**BIF Class:** FC5C - ADULT FCE EXAM GROUP

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Which language at home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 20: 1</td>
<td>Male: 3</td>
<td>Italian 10</td>
</tr>
<tr>
<td>20s: 5</td>
<td>Female: 7</td>
<td></td>
</tr>
<tr>
<td>30s: 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50s: 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years: 2</td>
<td>elementary: 1</td>
</tr>
<tr>
<td>3 years: 3</td>
<td>high: 7</td>
</tr>
<tr>
<td>4 years: 1</td>
<td>BIF: 4</td>
</tr>
<tr>
<td>5 years: 1</td>
<td>Private schools: 2</td>
</tr>
<tr>
<td>6 years: 2</td>
<td>University: 1</td>
</tr>
<tr>
<td>10 years: 1</td>
<td></td>
</tr>
</tbody>
</table>

Italian ss good at exams so like to take them. But ss shouldn't take Cambridge exams too early, n better to work their way through the levels. PET a “beautiful” exam but ss can get B at FCE and not be able to write emails; CAE seems to have Cs at a range of levels but all get C grades?? W how can ss with so much practice (e.g. pronunciation) still make such basic errors? NB this T has a deep knowledge of the Cambridge ESOL and BIF systems.

Sample lesson: Writing composition: revision of description adjectives to describe personality / appearance: crossword; Activity: gap fill from book to describe person in criminal investigation;; Flower FCE book: ; role play: “fun”: ss to describe a dangerous criminal; prep for composition: components needed in a composition: describe a member of your family or friend; say how they have affected your life; write plan; write composition; ask ss to do peer evaluation; hand in

<table>
<thead>
<tr>
<th>Learner 1 – male TA</th>
<th>22: law student; excellent, good mock, always does writing tho it is weaker; ambitions for CAE</th>
<th>Borderline, shy, LC a problem; has been thru the levels with BIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 2 – male FC</td>
<td>17; dentist; B in mock; exposed to LC; depends on topic</td>
<td></td>
</tr>
<tr>
<td>Learner 3 – female VC</td>
<td>22; student, natural science; not improving as much as she should?; attendance not so good; still at school.</td>
<td></td>
</tr>
<tr>
<td>Learner 4 – female SDG</td>
<td>22; D Eng degree; very tired; good all round except for writing; not a producer? Needs to learn the right approaches</td>
<td>Strong FCE; poor pronunciation; doesn’t need a T!</td>
</tr>
<tr>
<td>Learner 5 – male SF</td>
<td>34; statistician; only 60% attendance; RC usually OK</td>
<td></td>
</tr>
<tr>
<td>Learner 6 – female FP</td>
<td>59; librarian; needs El but negative about it; gave up in mock LC; no writing done</td>
<td></td>
</tr>
<tr>
<td>Learner 7 – male PP</td>
<td>32; has left class</td>
<td>Bright, made progress; will take FCE in December not June</td>
</tr>
<tr>
<td>Learner 8 – female CR</td>
<td>22; bright student but weak in LC; has cracked the writing; B in mock; could get A in FCE, high achiever</td>
<td></td>
</tr>
<tr>
<td>Learner 9 – female IW</td>
<td>22; first year law student; C in mock; disappointing, not improving? attendance poor; problems at U?</td>
<td>Has been to UK several times so LC better</td>
</tr>
<tr>
<td>Learner 10 – female ST</td>
<td>39; attendance poor; Mock reading VG; insignificant in class, frustrated in the lessons? Not doing much work</td>
<td>Rarely comes</td>
</tr>
<tr>
<td>Learner 1 – female MP</td>
<td>Nice; Wr, LC weaker? a bit quiet; oral results may not be so good; at FCE; but making progress</td>
<td>Seems to have left</td>
</tr>
<tr>
<td>-----------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>------------------</td>
</tr>
<tr>
<td>Learner 2 – male PLM</td>
<td>Late for mock; ambitious for FCE; in a hurry; PET scores disappointing (no merit) although he did well</td>
<td>Rarely comes; Did a UK Erasmus, up from PET but only passed.</td>
</tr>
</tbody>
</table>

**Teacher C – JG (Second group)**

<table>
<thead>
<tr>
<th>BIF Class: E4E - ADULT BIF EXAM GROUP</th>
</tr>
</thead>
</table>

### Ages:

| Pre-20: | 1 |
| 20s:    | 2 |
| 30s:    | 4 |
| 40s:    | 3 |

### Gender:

| Male: | 4 |
| Female: | 6 |

### Which language at home:

- Italian: 10

### Years of English:

- 2 years: 4
- 3 years: 2
- 4 years: 1
- 5 years: 2
- 10 years: 1

### At which school:

- middle: 1
- high: 3
- BIF: 7

The class is very diverse, with ages ranging from 18 to 45, and a wide range of English proficiency. They enjoy the CALT but dislike writing compositions. **Very good class dynamic and work well together.** Interested in reading, music, sailing, travel.

Typical lesson structure is: recycling / warm-up; introduction: context, topic; Using reading or listening; answer comprehension questions; answer more detailed questions on grammar or vocab; Speaking activity; Game / fun

E.g. 12/02: future tenses: 1) revision of vocab from last lesson; 2) overview of future tenses: activity from *English File 4*; 3) speaking about the future from *English File* and *Reward Up*, Intermediate; 4) Fun activity: fortune telling

<table>
<thead>
<tr>
<th>Learner 1 – female PL</th>
<th>Missed a term; involved, communicative</th>
<th>Making progress, makes own effort at home; yet no transfer? Or is it speaking vs. writing?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 2 – male CB</td>
<td>42; forensic Dr, works v hard but not a linguist, making progress</td>
<td>Relaxed re spoken English, fewer errors; written work much better; law background helping where language is so close and analytical, applying chunks and synthesising?</td>
</tr>
<tr>
<td>Learner 3 – male TD</td>
<td>33; lawyer; overcoming initial shyness and making progress; can joke in TL now having passed a threshold; enjoying himself but might complain later?</td>
<td>Environmental health problem; left school at 14; discipline and method needed</td>
</tr>
<tr>
<td>Learner 4 – female EF</td>
<td>33; could do better but doesn’t have enough time to study? Poor health, thyroid problems</td>
<td>Bright, v good at communicating</td>
</tr>
<tr>
<td>Learner 5 – female EI</td>
<td>21; psychology student young and scatty? T L OK but poor attendance; bright and good at self-correction</td>
<td></td>
</tr>
<tr>
<td>Learner 6 – male LN</td>
<td>Engineer; bright</td>
<td></td>
</tr>
</tbody>
</table>
| Learner 7 – male  
AP  | Doctor / dentist; VG student; knows a lot of idiomatic expressions; good grammar and writes clear and logical essays |
| Learner 8 – female  
DP  | V poor attendance |
| Learner 9 – male  
RP  | 42; doctor; doctor; pronunciation problems and makes elementary errors; not a natural LL but making progress, learns fast and is an achiever  
Needs EL class, NB pronunciation |
| Learner 10 – female  
RP  | 27; art history student; slow at understanding new grammar but gets there, tho’ tends to repeat errors; lovely; needs 1 : 1 tuition?  
Still lots of errors; obstinate re grammar? But has made progress since E2 |
| Learner 11 – female  
NV  | 21; law student |
| Learner 12 – female  
SZ  | 34; police woman; methodical and works well but poor attendance; is improving |
| Learner 13 – female  
SP  | Agricultural scientist; Works hard and improving  
Terrible pronunciation; nice; should pass internal exam OK |

**Teacher D - JD**  
**BIF Class: M5 - YL FCE GROUP**

| Ages:  
11 to 15: 2  
16 to 19: 8 | Gender:  
Male: 4  
Female: 6 | Language used at home:  
All Italian | At which school:  
elementary: 1; high school: 6; middle: 1 |

| Years of English:  
3 years: 3  
5 years: 2  
6 years: 2  
7 years: 2  
9 years: 1 | All school students, interests, according to T. "sex, drugs and rock and roll"!  
Reasons for studying TL at BIF parental pressure, self-M for jobs and snobbery.  
The students don’t like doing homework, particularly writing, which they don’t do enough of; they like video, films; The class are seen by the t as co-operative, friendly, motivated, cheerful and "not exaggeratedly industrious"; all intended to take FCE (in one year); FCE seen as a benchmark; NB the problem of school exams at the same time | Typical lesson text book based, tho’ T is behind in the book, mixing four skills where possible: grammar IP and practice; speaking practice in pairs or group; listening; occasional progress tests and reviews; as June approaches, more and more integration of exam practice |

| Learner 1 - female  
GB  | 17; student;  
Best worker, a little cocky, determinedly left-wing and not unexpectedly immature |
<table>
<thead>
<tr>
<th>Learner 2 - female CC</th>
<th>1 6, student</th>
<th>Quiet and very accurate; hard-working, quick to pick up</th>
<th>Strong and getting stronger</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 3 - female ChC</td>
<td>14, student;</td>
<td>Always asks the right questions; has fun resolutely</td>
<td>Much better</td>
</tr>
<tr>
<td>Learner 4 - female CF</td>
<td>18, student;</td>
<td>“Flower child” who is somewhat below the level of the others; making efforts to catch up</td>
<td></td>
</tr>
<tr>
<td>Learner 5 - female FF</td>
<td>17, student;</td>
<td>Quiet and conscientious</td>
<td>Better to makes frequent errors;</td>
</tr>
<tr>
<td>Learner 6 - male NG</td>
<td>16, student;</td>
<td>Habitual absentee, works reasonably well; pays more attention to his hormones than his m homophones!</td>
<td></td>
</tr>
<tr>
<td>Learner 7 - female RG</td>
<td>17, student;</td>
<td>Distracted, wildly inaccurate, but has fun and does what she should, tho remembers little</td>
<td></td>
</tr>
<tr>
<td>Learner 8 - male SM</td>
<td>18; student;</td>
<td>Quiet and mysterious; very shy, had a motivation problem, but has overcome it</td>
<td>Weak; pressure from mother</td>
</tr>
<tr>
<td>Learner 9 - male AP</td>
<td>16, student;</td>
<td>Hard worker, perhaps too hard, considered a bit of a swot &amp; somewhat alienated by others (new to the group)</td>
<td></td>
</tr>
<tr>
<td>Learner 10 - male ES</td>
<td>19, student;</td>
<td>Rampant playboy; oldest but most immature in class; sometimes despairs of his own efforts; has ability but has not yet tuned into it</td>
<td>Weak, worried may not pass; pressure from his mother</td>
</tr>
</tbody>
</table>

**Teacher E - AL (First Group)**  BIF Class: E4D - ADULT BIF EXAM GROUP

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Language at home?</th>
</tr>
</thead>
<tbody>
<tr>
<td>20s: 4</td>
<td>Male: 3</td>
<td>Italian: 9</td>
</tr>
<tr>
<td>30s: 2</td>
<td>Female: 6</td>
<td></td>
</tr>
<tr>
<td>50s: 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60+: 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school:</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;A lot&quot;: 3 2: /; 3; /</td>
<td>elementary: 1</td>
</tr>
<tr>
<td>5;/;7;/; 8;/; 10</td>
<td>middle: 6</td>
</tr>
<tr>
<td></td>
<td>high: 6</td>
</tr>
<tr>
<td></td>
<td>BIF: 2</td>
</tr>
</tbody>
</table>
An enthusiastic and well-motivated group with independent LL skills. Their priorities are aural skills; reading and writing are of secondary importance. They seem quite disciplined and consistent in their use of English in the classroom during activities, when asking for clarification etc. They seem to respond well to a variety of teaching techniques and activity types. They appear to enjoy working in pairs and small groups and seem to understand the value of this type of interaction pattern. My aim is to give the learners the opportunity to improve their language in a positive and supportive environment, trying as much as possible to find an opportunity to personalize materials and topics, thus promoting more “authentic” communication. They are conscientious in doing homework and can do it with the minimum of explanations. It generally consists of work from the English File IV workbook or vocabulary work. Sometimes I ask them to study the grammar focus point from their EF IV Student’s Book independently before dealing with it in class.

A number of the students (MS, GS, FM and SDG) regularly borrow graded readers from the BIF library. When asked their opinion of the course book, they responded favourably to it. It appears to cover topics which are of relevance to them and which can generate a minimum of language.

The course syllabus is based on their book. We are behind in terms of units which should have been covered up to now but I am happy with their progress so far. I haven’t felt the need for testing as they are quite clearly enjoying and improving their English.

Group dynamics are interesting. MS and DP tend to monopolize student-to-teacher interaction. DP is at times antagonistic towards MS which occasionally raises the temperature in class. The others students are forthcoming in asking me questions but less vociferous. I let them decide on where they sit, thus largely determining who they will be working with during pair work. I think this is conducive to a more relaxed atmosphere and so more likely to generate language. Learning appears to be taking place in a relaxed and enjoyable environment and I am very happy with their progress. This helps to build my confidence as a teacher which in turn is motivates me.

Relationship of FCE and YLs? Programme should reflect learner interests; important that they should be motivated, but this group’s motivation is not exams; group members do little writing; YL version of FCE too structured; FCE academic but LC and speaking relatively easy; in Use of English, examiners don’t know what to ask the Cs! so the exam splits hairs. NB book vs. non-book, graded readers; NB speaking, which T likes to encourage. Group take the BIF internal exam, level 5 next; whether they then take FCE depends on the individuals; T wouldn’t propose PET for students learning for pleasure; FCE level (of content) more suitable for this group?

| Learner 1 – male | 24; |
| Learner 2 – female | 21; architect student; Uses EL at University |
| Learner 3 – male | 23 | Has Irish girl-friend |
| Learner 4 – female | 58? |
| Learner 5 – female | 31; professional; Doping well; works outside; minimal time |
| Learner 6 – female | 36; primary school EL teacher; Tries hard but misses every third lesson. |
| Learner 7 – male | 62; Married Couple – MS/GS |
| Learner 8 – female | 61 | MS and his wife GS seem to be progressing; inaccurate but good LC and RC; they read TL outside |
| Learner 9 – male CM | 21; economics student; | Keen |
| Learner 10 – female SDG | 34; Engineer; | serious |
| Learner 11 – male BB | 22; economics student; | Initially reticent, better now; NB Erasmus and progress |
| Learner 12 – female IS | 25; law student; 3 yrs BIF; | Fairly weak; not much TL outside |

**Teacher E – AL (Second Group)**

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Which language at home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>12 to 15: 8</td>
<td>Male: 4</td>
<td>Italian: 9</td>
</tr>
<tr>
<td>16 to 18: 2</td>
<td>Female: 8</td>
<td>Catalan: 1</td>
</tr>
<tr>
<td>8 yrs: 2</td>
<td></td>
<td>Spanish: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 years: 4</td>
<td>elementary: 7</td>
</tr>
<tr>
<td>6 years: 1</td>
<td>middle: 3</td>
</tr>
<tr>
<td>7 years: 2</td>
<td>high: 5</td>
</tr>
<tr>
<td>8 years: 2</td>
<td>BIF: 10</td>
</tr>
<tr>
<td>9 years: 3</td>
<td>American Inst: 1</td>
</tr>
</tbody>
</table>

A young learner’s group with a range of language ability levels. See FCE exam results sheet.

**Their main priority is passing FCE.** Though this is determined by BIF levels scheme.

They are quite disciplined and consistent in their use of English in the classroom during activities, when asking for clarification etc. Being chatty young people, there is a tendency to break into Italian in-between activities. They seem to respond well to a variety of teaching techniques and activity types. Though FCE tends to be rather dry and structured in approach. They are used to working in pairs and small groups and seem to accept this type of interaction pattern. My aim is to give this group of learners the opportunity to practise and improve their language within the context of exam preparation. They are fairly conscientious in doing homework and do it with a minimum of explanations. It generally consists of work from either the FCE Gold student’s book or the Exam Maximiser. Being young learners, I don’t think that they have time for additional study such as using graded readers. The course syllabus is based on the FCE course book. I move around the book a lot in order to introduce an element of variety to a somewhat monotonous diet. Most of the students are forthcoming in asking me questions when they need help with vocabulary or when they don’t understand. **Nobody dominates in class, despite the differences in level.** I let them decide on where they sit, thus largely determining who they will be working with during pair work. If chatting gets out of hand, then I will review this practice. **My impression is that they understand the importance of learning English: whether or not they actually enjoy it is another matter!** Learning appears to be taking place in an environment conducive to learning. I am fairly confident that they will be successful in their FO Exams in June.

<p>| Learner 1 - female BA | 17, student |
| Learner 2 - male NA | 15, student |
| Learner 3 - female FC | 14, student |
| Learner 4 - male LC | 14, student |
| Learner 5 - male CC | 13, student |
| Learner 6 - female | 16, student |</p>
<table>
<thead>
<tr>
<th>Learner 7 - female</th>
<th>12, student</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 8 - male</td>
<td>15, student</td>
</tr>
<tr>
<td>Learner 9 - female</td>
<td>14, student</td>
</tr>
<tr>
<td>Learner 10 - female</td>
<td>15, student</td>
</tr>
<tr>
<td>Learner 11 - female</td>
<td>17, student</td>
</tr>
<tr>
<td>Learner 12 - female</td>
<td>17; student</td>
</tr>
</tbody>
</table>

**Teacher F - VR**  
**BIF Class: J4A - YOUNG LEARNERS - PET GROUP**

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Language used at home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 12: 1</td>
<td>Male: 7</td>
<td>Italian 12</td>
</tr>
<tr>
<td>12 to15: 11</td>
<td>Female: 5</td>
<td>French 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>English 1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Portuguese 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school:</th>
</tr>
</thead>
<tbody>
<tr>
<td>9 years :1</td>
<td>Elementary: 6</td>
</tr>
<tr>
<td>8 years :3</td>
<td>middle:1</td>
</tr>
<tr>
<td>6 years :3</td>
<td>high: 5</td>
</tr>
<tr>
<td>7 years: 2</td>
<td></td>
</tr>
<tr>
<td>4 years: 3</td>
<td></td>
</tr>
<tr>
<td>2 years: 1</td>
<td></td>
</tr>
</tbody>
</table>

YL class who have been studying together at BIF for a couple of years. As a result they all get on very well and the class has a fairly positive atmosphere, as it means that they can meet up with their friends. The students are all very motivated, particularly for young learners, and seem to enjoy studying English. I think this is helped by the fact that their parents give them a lot of support, but without applying excess pressure, at home.

Generally, lessons begin in a very focused manner, with homework correction, or, recently, exam practice. Introduction to the task or activity is quite teacher-centred, with an explanation or demonstration of what to do, which allows for any questions about the task, vocab, etc to be answered.

The lesson then follows a general PPP approach, based on the course book "Kickstart", which covers all the skills and also has sections which focus on the exam. Nearly all the tasks are done in pairs or small groups, with feedback as a class, but occasionally, students are asked to work alone. (This is usually when the class gets quite boisterous and a calming activity is needed.) However, we still feedback as a class. At the end of the lesson, the last 15 minutes or so, students usually have the opportunity to relax with a “fun” activity such as a song or a game, although this is possible only when all the activities are completed.
<p>| Learner 1 - male | 14, student; strength is speaking, good vocabulary and lots of confidence when communicating in TL; can get his point across easily; but difficult to hold his attention for any amount of time, very easily distracted; so considerably weaker in other areas of English. When he does apply himself, his work is very good, but at the moment, this rarely happens. | Wants to wait till FCE, more credits at school; didn't do well at Mock; settled down; no exam but contributes well, more consistent homework, volunteers to RA etc. |
| Learner 2 – female | ??; lot of enthusiasm when for English; half-sister in London (waiting to take A levels) which motivates her to study, would like to live in England herself transferred from another BIF class at the beginning of the course, so to begin with she was quite shy; confidence has increased over the last few months, aided by her peers often ask her for help; all rounder, who pays close attention in class and always completes homework assignments; making steady progress in English. | Practises at home with model exams; absent a lot but keeps ahead |
| Learner 3 – male | 14, student; very motivated, making steady progress in all areas; NB writing skills have improved considerably since October, mainly as a result of writing his Diary regularly. Quite good at speaking and can communicate very well in TL; e areas he has trouble with are listening and reading, which was reflected in the PET mock. | Lovely, will not take PET; forgot to register; E for football purposes; has had a trial for Chelsea; speaking good, LC not so good (on cassette); Mock scores low, NB RC and writing |
| Learner 4 – male | 13, student; Shy; youngest, like CB, transferred from another class at the BIF October 02. but although CB has overcome her shyness, FB hasn’t, remains very quiet in class; does like learning TL, but I think finds it difficult to be in a class like this, slightly dominated by the older boys. Very good at speaking and listening, having trouble with reading and writing. I often find myself working with him individually when there’s an odd number of students; re LC and speaking in Feb 03 Mock. | T worried; speaking Ok but too quiet voice; LC OK, RC, Wr hard; poor vocab? Tenses confused, no punctuation! |</p>
<table>
<thead>
<tr>
<th>Learner 5 – female</th>
<th>GC 13; student; good all and seems to enjoy the lessons, always does the activities well, NB exceptionally good at reading in English; very natural intonation pattern, also comprehends listening fairly quickly. Occasionally she is distracted by CC, which means she hasn’t progressed as much as she could have in the area of writing.</th>
<th>V good, attends well, v motivated tho’ quiet; writing VG, speaks clearly; Dad at BIF; she has a house in UK, but not spoilt!</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 6 – female</td>
<td>CC 13; student; very laid-back approach to English, and she rarely does her homework, finds most activities easy and skims through most tasks quickly, usually without exploring their full potential, particularly true for speaking activities. I wonder if the fact that her mother is also an English teacher affects her attitude towards learning?</td>
<td>Mother and elder brother at BIF; like her brother, CC can be disruptive; finds ELL easy</td>
</tr>
<tr>
<td>Learner 7 – male</td>
<td>EDS 13; student; Brazilian, having moved to Italy with his mother several years ago, and so speaks 3 languages fluently, don't think this is affecting his progress in English, his frequent absence from class is; strengths speaking and listening, followed by reading and writing; fairly good grasp of grammar.</td>
<td>Portuguese spoken at home; more mature, calm now, helps to settle GM down; writing not so good1, handwriting poor; absence period but catching up because his TL level is high; speaking VG, reading good.</td>
</tr>
<tr>
<td>Learner 8 – male</td>
<td>BHS 15, student; lots of enthusiasm for TL; one of strongest; may be helped by an English Grandfather, but also because he takes the time to watch films and read books in TL; so, VG all rounder, like CC and BL; oldest in e class, occasionally taken responsibility for class activities, e.g., Christmas play, which he co-wrote with Andrea, also leads them into chaos sometimes, usually at 6:45 on Fridays!</td>
<td>Was better, more confident than others, but they have caught up; wouldn't listen to correction etc; went backwards in ELL; not so good in prog test, which jolted him; Now more attentive; still VG; this year is a revision year.</td>
</tr>
<tr>
<td>Learner 9 – female</td>
<td>BL</td>
<td>13; student; strong student, makes every effort to complete all tasks set to best of her ability.; good all rounder, but can be a little shy, will only volunteer information when nominated; seems to enjoy studying TL; regularly completes Diary and all other homework assignments.</td>
</tr>
<tr>
<td>-------------------</td>
<td>--</td>
<td>-----------------------------------------------</td>
</tr>
<tr>
<td>Learner 10 - male</td>
<td>GM</td>
<td>14 student; probably most confident speaker, always willing to answer any question or to volunteer to read aloud.; enthusiasm often contagious, cheeky attitude is quite endearing.; hardly ever does homework, always completes tasks set in class very well; weakest area writing, which hasn't helped rarely does his Diary, recently, has started to improve.</td>
</tr>
<tr>
<td>Learner 11 – female</td>
<td>LM</td>
<td>12; student; tries very hard in English, results slightly hit and miss! When listening or reading, often guesses answer, writing not as good as other members in the class; may be because she 'jumped' a year at the BIF. When speaking, she can communicate well, although she is a little shy, so has to be nominated.</td>
</tr>
<tr>
<td>Learner 12 – male</td>
<td>PM</td>
<td>15; very motivated when learning English, good in all areas; always completes homework assignments well, but is easily distracted in class, particularly when LA is there, which is impeding his progress a little.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Teacher G - HB</th>
<th>BIF Class: J4B - YOUNG LEARNERS PET GROUP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ages:</td>
<td>(not filmed)</td>
</tr>
<tr>
<td>12 to 14</td>
<td>Gender:</td>
</tr>
<tr>
<td></td>
<td>Male: 2</td>
</tr>
<tr>
<td></td>
<td>Female: 9</td>
</tr>
</tbody>
</table>
**Not an exam-based approach:** YLs should not be pushed; Some of the group have taken KET; T doesn’t enter ss for exams unless it seems right for them; not much specific PET-driven practice, but next year the students in the group will be going for FCE; most of their parents are “flexible”; T can only advise on whether the ss may take the exam or not. Some of the group are older, e.g. have done M3, M4 in-between years.

YL courses should be presented in terms of what they will achieve towards FCE, 5/6; PET a yardstick; you shouldn’t have one, some taking the external exam and some not or their confidence is affected. IC, CDS, GB, RVS, NF could take FCE in one year; better for all to stay together and do it in 5 years. PET >> FCE a big step, in terms of maturity; NB writing has improved.

<table>
<thead>
<tr>
<th>Learner 1 – female AA</th>
<th>Capable of making more progress than at this stage; when concentrated, capable of producing good work; good LC in MY test; needs to be determined to make the effort if she wants to pass PET June 03 exam. Mature, good, tho’ hesitant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 2 – female GB</td>
<td>Excellent progress; good PET mock grade; enthusiastic, spoken TL particularly good; keep up hard work to do well in June PET</td>
</tr>
<tr>
<td>Learner 3 – female IC</td>
<td>Steady progress, especially in RC and LC; v conscientious, little shy, under-confident, affecting speaking and writing; hard work can improve these areas for June PET pass.</td>
</tr>
<tr>
<td>Learner 4 – female CDS</td>
<td>Always works hard, enthusiastic. Making v good progress in all areas; should pay attention to accuracy in speaking and writing; Mock show she should do well in PET June 03</td>
</tr>
<tr>
<td>Learner 5 - male NF</td>
<td>12, student, Fluent, will take PET; v enthusiastic, curious, making v good progress I all areas; need top practise writing as can be erratic; PET Mock good so hard work and should do well in June</td>
</tr>
<tr>
<td>Learner 6 - male CF</td>
<td>Excellent progress in all areas; writing improved, speaks with ease; Mock PET v good so PET June should do well</td>
</tr>
<tr>
<td>Learner 7 - female FL</td>
<td>Very capable, potential to do well in June PET; sometimes does not realise her ability and hides behind shyness; Mock PET show she can do well, should work hard and with confidence</td>
</tr>
<tr>
<td>Learner 8 - female MO</td>
<td>Must really want to do PET to make necessary improvements to pass comfortably; capable but speaking and writing skills plagued with elementary errors, which when pointed out she understands easily; LC good; need to be determined and confident</td>
</tr>
<tr>
<td>Learner 9 - female LO</td>
<td>VG progress, enthusiastic, motivated; speaking and writing developing well; needs to work on accuracy in grammar and LC for PET.</td>
</tr>
<tr>
<td>Learner 10 – female GT</td>
<td>Good commitment and progress; speaking a little weak because she needs more confidence in her own ability; Good PET mock; try for it in June.</td>
</tr>
<tr>
<td>Learner 11 - female RVS</td>
<td>--</td>
</tr>
</tbody>
</table>

366
### Teacher H - AC  
**BIF Class: E3A - ADULT PET/BIF EXAM GROUP**

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Language used at home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under 20: 2</td>
<td>Male: 4</td>
<td>Italian: 12</td>
</tr>
<tr>
<td>20s: 1</td>
<td>Female: 8</td>
<td>Chinese: 2</td>
</tr>
<tr>
<td>30s: 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40s: 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50s: 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60+: 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 years: 6</td>
<td>BIF: 10</td>
</tr>
<tr>
<td>3 years: 4</td>
<td>Uni: 1</td>
</tr>
<tr>
<td>4 years: 1</td>
<td></td>
</tr>
<tr>
<td>7 years: 1</td>
<td></td>
</tr>
</tbody>
</table>

All took Mock PET; T thought 8/11 would take PET but in fact 4/11 will; T sees PET as useful “lower middle level exam”, encouraging because at this level students think they are weaker than they really are; PET has a good grammar balance. NB LC. T presents PET positively and fully but without pressurising students to take it. Mock results are not the key to whether students take the exam or not. A common view is “Do FCE but don’t bother with PET”.

The class are generally fine, keen and more relaxed, motivated; Some of them have made contact with US, UK students and continue this. T gives both compulsory and optional written homework, mainly on Wednesdays to give more time and ss respond to this; **writing practice not usually PET-restricted.**

<table>
<thead>
<tr>
<th>Learner 1 - female SB</th>
<th>51; fairly good but not consistent; LC hard; seems to have mental blocks about some things; may be because he is getting on a bit.</th>
<th>Finds LC hard; uses cassettes for practice outside class; so improving, tho’ not fast; not regular with written work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 2 - female MDB</td>
<td>43; architect; quite strong but absent quite a lot; LC weaker</td>
<td>Q strong; will do well at PET; but won’t take it; Thinks a lot, maybe too much? She and LV do extra fluency work</td>
</tr>
<tr>
<td>Learner 3 - female VC</td>
<td>32; usherette; late and often absent; middle of the class; doesn’t often give in writing assignments</td>
<td>Still not much writing; won’t take PET, will wait till FCE</td>
</tr>
<tr>
<td>Learner 4 YM</td>
<td>18; student; has withdrawn</td>
<td></td>
</tr>
<tr>
<td>Learner 5 - female AM</td>
<td>33; absent a lot; confident in TL but not accurate</td>
<td>Accuracy a bit better? (there’s grammar in the BIF test). Grammar</td>
</tr>
<tr>
<td>Learner 6 - female AM</td>
<td>55; lacks confidence, not as weak as she thinks;</td>
<td>Asked to do PET; happier now. Should stay as writing quite good, tho’ oral / aural not so good. OK in pairs but not in front of the class</td>
</tr>
<tr>
<td>Learner 7 - male MP</td>
<td></td>
<td>NB sounds; will take PET; dominant in class; too old?</td>
</tr>
<tr>
<td>Learner 8 - male IP</td>
<td>57? Regular attender but among the weaker ss; doesn’t apply what he has learnt; should pass Mock;</td>
<td>Slow progress, quite nervous; writing not that strong?</td>
</tr>
</tbody>
</table>
Learner 9 - female
CS

32; doing EOP, for tourism; quite good; LC a problem?

Will take PET; doing very well; asks T for communicative help; holiday home; speaking stronger; LC improving; writing OK.

Learner 10 - male
LV

31; Hotelier; on study leave; stronger all round; should be OK.

Returned from Edinburgh; will do PET; speaking good, LC good, grammar fine; writing and speaking poor; doesn’t often do writing; does do other extra work.

Learner 11 - male
DZ

54; silversmith; started at low level; confident, reads upper intermediate; Mock LC a problem; speaking PET; Writing good

Works hard, improving a lot but won’t pass interview; no PET; failed mock; will take the BIF test; good attendance; in E2 he took the year twice and failed the exam.

Learner 12 - female
XY

17; student; bubbly but “all over the place”; may not pass; double interference?

Grammar not good; misses classes; speaking stronger and LC good, perhaps because she listens to pop.

Will take PET; just passed Mock.

---

**Teacher I - DB**

**BIF Class: E3B - ADULT PET /BIF EXAM GROUP**

<table>
<thead>
<tr>
<th>Ages:</th>
<th>Gender:</th>
<th>Language at home:</th>
</tr>
</thead>
<tbody>
<tr>
<td>20s: 3</td>
<td>Male: 5</td>
<td>Italian: 10; Chinese: 1; Spanish: 1; Polish: 1</td>
</tr>
<tr>
<td>30s: 3</td>
<td>Female: F</td>
<td></td>
</tr>
<tr>
<td>40s: 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>50s: 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60+: 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Years of English:</th>
<th>At which school:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 year: 4</td>
<td></td>
</tr>
<tr>
<td>2 years: 3</td>
<td></td>
</tr>
<tr>
<td>3 years: 1</td>
<td></td>
</tr>
<tr>
<td>4 years: 2</td>
<td></td>
</tr>
<tr>
<td>5 years: 2</td>
<td></td>
</tr>
<tr>
<td>Secondary: 3; BIF: 9</td>
<td></td>
</tr>
</tbody>
</table>

Group of mixed personalities; Learner 5 (LM) changed her job and had to move to a Saturday class; only two turned up to a lesson and too much L1 spoken; T read riot act; 2 more moved to Saturday; Learner 4 (SI) and Learner 12 (SP) vs. Learner 7 (BV) and Learner 11 (GP)?

Now only 9 in group; T focuses on the noisy students.

Repeating a problem; NB a.m. vs. p.m. classes; evening attendance for more personal reasons, social, freelances?

**Revolution over**; 4 of the group will take PET; E3 not explicitly PET (internal exam for all); YL courses are PET courses; T offers extra PET classes Fridays, for those who will take it; NB different reasons for taking external exams, serious attitudes, wish for hurdles, certification, quite normal. Some have instrumental reasons for taking exams, others learn Ls for social reasons, “to keep their minds alive”, + those who change roles, e.g. Learner 3 (PF) who on day one asked about PET because her daughter had done KET; PF found it v expensive so wouldn’t do it but will now because of her objectives. Don’t do PET if you get less than 42/75 on the oral?
<table>
<thead>
<tr>
<th>Learner 1 – female DB</th>
<th>31; young mother; likes lessons; OK? straightforward, right placement; feeling of progress; likes explanation; practises in shop; uses Murphy</th>
<th>Good attendance, uses class hours well, active; writing OK but progress test disappointing and she didn’t bother; speaking should be gaining.</th>
<th>Good attendance, keen, interested but little work outside(T doesn’t insist); Progress test OK will do PET, extra classes by T free; PET may motivate her</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learner 2 – female CB</td>
<td>23; art / design student, from Colombia; well-heeled family enthusiastic, speaking better than writing? Lots of progress in RC, pronunciation; happy</td>
<td>Enthusiastic as ever; uses her brain, applies rules, very thorough; wants to be sure of new words pronunciation; pronunciation and syntax a problem when speaking (Spanish interference?) speaking not improving?</td>
<td>Good attendance, enthusiastic still, pro-T; writing improving more than before, but speaking? T stresses impact on reading e.g. for applications, appreciation; tolerance ion writing lower than for speaking. No PET</td>
</tr>
<tr>
<td>Learner 3 – female PF</td>
<td>50; gynaecologist; lacks confidence; reads well; bright; LC problems, some progress</td>
<td>Missed three lessons because operating; active, participatory, progress test OK</td>
<td>Will take PET; good chance; has been ill;</td>
</tr>
<tr>
<td>Learner 4 – female SI</td>
<td>Not noticed, timid; should be OK; now has afternoon job; busy; afraid of speaking</td>
<td>Has left class (2/03); had been conscientious, some good writing, asked questions of T; serious, shy, speaking weak so she got worried and resentful</td>
<td>N/A</td>
</tr>
<tr>
<td>Learner 5 – female LM</td>
<td>Young shop assistant; outgoing and keen; should do OK</td>
<td>LEFT CL;ASS (2/03) Changed job; missed lessons; serious, making progress</td>
<td>N/A</td>
</tr>
<tr>
<td>Learner 6 – female EM</td>
<td>45; (ex-Professor?) Computer illiterate; lacks confidence; speaking problems; should be OK; happy</td>
<td>Always late, TL weak, lacks confidence</td>
<td>Wants to do PET (with PF); speaking poor, writing better; (cf her can-do’s negative??)Eyes all over the place looking at others’ work.</td>
</tr>
<tr>
<td>Learner 7 – male BV</td>
<td>65; retired businessman; E for travel etc; weak TL but settling in; chatters; not much progress pron.</td>
<td>How did he get into the class; GP’s friend; sometimes doesn’t follow;</td>
<td>Not interested in PET; no instrumental M; more interested in homework; slow</td>
</tr>
<tr>
<td>Learner 8 – male AW</td>
<td>37; Polish priest; new; bright, likes T explanations!, hard-working; grammar improving, speaking difficulty</td>
<td>Cheerful and positive, stays late with extra work group; learner but conscientious; T relieved he’s not doing PET because of his LC and speaking; Has been away; no PET</td>
<td></td>
</tr>
<tr>
<td>---------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>Learner 9 – male GZ</td>
<td>21; Chinese student; in business? Liked Communicat</td>
<td>Went back to PRC;; missed a lot of lessons</td>
<td></td>
</tr>
<tr>
<td>Learner 10 – male EF</td>
<td>52; obstetrician; was level 6, seems more like 3; misplaced; LC weak</td>
<td>Still “crazy”; erratic attendance; now at right level; not fluent but now intervening better and more appositely; getting more out of class now; “Floating in and out”, late etc; pleased she’s down to a class at this level; When she intervenes, her Qs are relevant; No PET yet; next year?</td>
<td></td>
</tr>
<tr>
<td>Learner 11 – male GP</td>
<td>22; U student; wrong level? failed last year, big issue as EL important for him; writing better; grammar problem; T can tell him which units of Murphy to use</td>
<td>Good attendance but still the same problems; doesn’t listen, talks to neighbour; T call him by name now so he feels picked on; sometimes wants to argue re vocab, e.g. anglicisations; Still “crazy things” in his writing; What next? Last year he failed the exam, retook, just passed; tried harder but did he improve? Not academic.</td>
<td></td>
</tr>
<tr>
<td>Learner 12 – female SP</td>
<td>Needs to do more homework; LC difficulty; feeling discouraged; comes to get away from family situation, too.</td>
<td>Left for E3S, 2/03</td>
<td></td>
</tr>
</tbody>
</table>
Appendix 6.8

Responses to the LLQ

The questionnaires completed by 60 BIF learners gives the following averages for the main criteria:

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Mean scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitude</td>
<td>2.5</td>
</tr>
<tr>
<td>Anxiety</td>
<td>1.7</td>
</tr>
<tr>
<td>Motivation and Effort</td>
<td>3.48</td>
</tr>
<tr>
<td>Doing strategies</td>
<td>2.79</td>
</tr>
<tr>
<td>Thinking strategies</td>
<td>3.18</td>
</tr>
</tbody>
</table>

The mid-range score for attitude towards the language and its speakers suggests that this was not a particularly relevant feature for these learners (neither negative nor very positive). The low score for anxiety suggests that the BIF students who completed the LLQ were quite relaxed about learning the target language and not particularly nervous about making mistakes. They also seemed to be relatively calm about taking tests.

The relatively high average on motivation and effort (3.5) suggests that they were generally well-motivated to learn the target language. It also suggests that tests encouraged them to study a little harder than normal.

The scores for the “doing” and “thinking” strategies suggest that they used a range of learning strategies to help them understand and remember things: e.g. asking questions; looking for contextual clues to help make sense, making associations; learning rules and trying them out in different contexts; making written or mental lists: etc.

The LLQ had been designed to be used in a mixed method approach with smaller sample sizes; this kind of use was modelled on other questionnaire instruments (such as the SHL Occupational Personality Questionnaire – OPQ3). Responses and scores from such
questionnaires are used in social contexts where decisions are made about individuals on the basis of their personal traits, especially for staff recruitment. However, the scores are not usually used in isolation, but form part of a decision-making process which typically includes a debriefing of the respondent by a trained interlocutor: e.g. “You responded to question X in the following way: …… This suggests that you are Y. What do you think about this? Is this how you see yourself?” This approach allows for the traits to be examined in greater depth and for the interlocutor to determine the validity and interpretation of the responses given. In the FLLGP, it was not possible to carry out this stage systematically, although to some extent the Project Coordinator performed a similar role by interviewing many of the participants.

The SHL OPQ32 (Saville and Holdsworth: www.shl.com) is one of the most popular and widely used occupational personality questionnaires and is available in over 30 languages. SHL claim that it is suitable for a range of HR purposes: e.g. selection, training needs analysis and individual development, career counselling, succession planning, team building and development, and organisational change.
Appendix 6.9
Note on latent growth and multilevel modelling

Latent growth modelling is a statistical technique used in the structural equation modelling (SEM) framework to estimate growth over a period of time. It is also called latent growth curve analysis and various SEM software programs are available to estimate the growth trajectory (e.g. AMOS, M-plus, LISREL and EQS). It is now widely used by researchers in the field of behavioral science, social science and education.

Latent growth models require a longitudinal design with repeated measures of the dependent variables as a function of time and other measures. In language education, this technique can be applied to learning, and proficiency gains can be considered the dependent variable to be investigated. Applications of this approach can reveal systematic change in a group of learners, as well as variability between individuals with respect to this change.

Systematic data collection over time is important for this approach and the size of the sample determine the range of analyses which can be applied. With large data sets (e.g. 1000 learners) it is possible to examine how changes in one or more of the variables affect the changes in another variable. It is also possible to do exploratory analyses to locate “clusters” of cases that have very different growth curves; for example, there might be a sub-group of learners that does not change in line with the group, a sub-group that changes in the opposite direction of the overall pattern, and a sub-group that changes more rapidly than the others.

Similar questions can also be addressed using hierarchical or multi-level modelling. This approach allows variance in outcomes (e.g. learning) to be analysed at multiple hierarchical levels. In educational research the data that is required for this analysis is considered to be “nested”; for example, learners are nested within classrooms which in turn are nested within schools.
Bibliography


Alderson, J C (1995) Ideas for research into impact, washback and IELTS, Paper commissioned by Cambridge ESOL.


Association of Language Testers in Europe (1995) *Development and descriptive checklists for tasks and examinations*, Produced by the members of ALTE with the support of a grant from the Lingua Bureau, Brussels.


Berry, V (1994) Current assessment issues and practices in Hong Kong: A review, in Nunan, D, Berry, R and Berry, V (Eds) Bringing about change in language education, Hong Kong: The University of Hong Kong, Department of Curriculum Studies.


379


CEDE (Centro Europeo dell’Educazione) 2001: www.cede.it.


Center for Continuous Training in European Dimension. April 2001 - www.ceses.it


Cheng, L (1997a) How does washback influence teaching?, Implications for Hong Kong, Language and Education 11(1), 38-54.


Green, A (2005b) EAP study recommendations and score gains on the IELTS Academic Writing test, *Assessing Writing* 10(1), 44-60.

Green, A (2007a) IELTS Washback in Context; Preparation for academic writing in higher education, Cambridge: Cambridge ESOL and Cambridge University Press.


Kunnan, A (2000) IELTS impact study project report prepared for the University of Cambridge Local Examinations Syndicate.


Mislevy, RJ (2002)


Nunan, D, Berry, R and Berry, V (Eds) (1996) Bringing about change in language education, Hong Kong: The University of Hong Kong, Department of Curriculum Studies.


Spolsky, B (1996) The examination-classroom backwash cycle: some historical cases, In D. Nunan, D, Berry, R, and Berry, V (Eds) Bringing about change in language education (pp. 55-66). Hong Kong: The University of Hong Kong, Department of Curriculum Studies.


Strevens, P D (1980) The paradox of individualised instruction: it takes better teachers to focus on the learner, In Altman H and James, C (Eds) 1980.


Wall, D (1999) *The impact of high-stakes examinations on classroom teaching: a case study using insights from testing and innovation theory.* Unpublished PhD thesis Department of Linguistics and Modern English Language Lancaster University


