Multilingual Frameworks

The construction and use of multilingual proficiency frameworks

For a complete list of titles please visit: http://www.cambridge.org/elt/silt

Also in this series:

Issues in Testing Business English: The revision of the Cambridge Business English Certificates
Barry O'Sullivan

European Language Testing in a Global Context: Proceedings of the ALTE Barcelona Conference July 2001
Edited by Cyril J. Weir and Michael Milanovic

IELTS Collected Papers: Research in speaking and writing assessment
Edited by Lynda Taylor and Peter Falvey

Testing the Spoken English of Young Norwegians: A study of testing validity and the role of 'smallwords' in contributing to pupils' fluency
Angela Hasselgreen

Changing Language Teaching through Language Testing: A washback study
Liying Cheng

The Impact of High-stakes Examinations on Classroom Teaching: A case study using insights from testing and innovation theory
Dianne Wall

Assessing Academic English: Testing English proficiency 1950–1989 – the IELTS solution
Alan Davies

Impact Theory and Practice: Studies of the IELTS test and *Progetto Lingue 2000*
Roger Hawkey

IELTS Washback in Context: Preparation for academic writing in higher education
Anthony Green

Examining Writing: Research and practice in assessing second language writing
Stuart D. Shaw and Cyril J. Weir

Multilingualism and Assessment: Achieving transparency, assuring quality, sustaining diversity – Proceedings of the ALTE Berlin Conference, May 2005
Edited by Lynda Taylor and Cyril J. Weir

Examining FCE and CAE: Key issues and recurring themes in developing the First Certificate in English and Certificate in Advanced English exams
Roger Hawkey

Language Testing Matters: Investigating the wider social and educational impact of assessment – Proceedings of the ALTE Cambridge Conference, April 2008
Edited by Lynda Taylor and Cyril J. Weir

Components of L2 Reading: Linguistic and processing factors in the reading test performances of Japanese EFL Learners
Toshihiko Shiotsu

Aligning Tests with the CEFR: Reflections on using the Council of Europe's draft Manual
Edited by Waldemar Martyniuk

Examining Reading: Research and practice in assessing second language reading
Hanan Khalifa and Cyril J. Weir

Examining Speaking: Research and practice in assessing second language speaking
Edited by Lynda Taylor

IELTS Collected Papers 2: Research in reading and listening assessment
Edited by Lynda Taylor and Cyril J. Weir

Examining Listening: Research and practice in assessing second language listening
Edited by Ardeshir Geranpayeh and Lynda Taylor

Exploring Language Frameworks: Proceedings of the ALTE Kraków Conference, July 2011
Edited by Evelina D Galaczi and Cyril J. Weir

Measured Constructs: A history of Cambridge English language examinations 1913–2012
Cyril J. Weir, Ivana Vidaković, Evelina D Galaczi

Cambridge English Exams – The First Hundred Years: A history of English language assessment from the University of Cambridge 1913–2013
Roger Hawkey and Michael Milanovic

Testing Reading Through Summary: Investigating summary completion tasks for assessing reading comprehension ability
Lynda Taylor

Multilingual Frameworks

The construction and use of multilingual proficiency frameworks

Neil Jones
Consultant, Cambridge English Language Assessment (part of the University of Cambridge)

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107641723

First published 2014

Printed in

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

Library of Congress Cataloguing in Publication data
Jones, Neil, Dr.
Multilingual frameworks : the construction and use of multilingual proficiency frameworks / Neil Jones.
pages cm. -- (Studies in language testing ; 40)
Summary: "Describes 20 years of work at Cambridge English to develop multilingual assessment frameworks. Multilingual Frameworks covers the development of the ALTE Framework and 'Can Do' project; work on the Common European Framework of Reference and the linking of the Cambridge English exam levels to it; Asset Languages - a major educational initiative for UK schools; and the European Survey on Language Competences. It proposes a model for the validity of assessment within a multilingual framework, and while illustrating the constraints which determined the approach taken to each project, makes clear recommendations on methodological good practice. It looks forward to the further extension of assessment frameworks to encompass a model for multilingual education"-- Provided by publisher.
Includes bibliographical references and index.
ISBN 978-1-107-64172-3 (paperback)
1. English language--Ability testing--Evaluation. 2. Language acquisition--Ability testing--Evaluation. 3. English language--Study and teaching--Foreign speakers--Evaluation. 4. Communicative competence--Evaluation. 5. Education, Bilingual. I. Title.

P118.7.J66 2014
428.0076--dc23

2014017113

Contents

List of tables and figures vii
Acknowledgements x
Series Editors' note xii
Abbreviations xx

1 Multilingual frameworks: A practical pursuit 1

2 Constructing a multilingual framework 5
2.1 Construct definition 5
2.2 Scale construction 11
2.3 Alignment across skills, languages and contexts 27
2.4 Standard setting 42
2.5 Validity and validation 46
2.6 Summary 59

3 Scaling comes to Cambridge ESOL: The 1990s 61
3.1 Early scaling developments 61
3.2 Computer-adaptive testing 62
3.3 The Cambridge ESOL Common Scale 69
3.4 The ALTE Framework and Can Do project 76
3.5 Conclusions: A pioneering age 84

4 A universal standard? The Common European Framework of Reference 87
4.1 Origins of the CEFR 87
4.2 The CEFR and Cambridge English 87
4.3 The CEFR as a measurement construct 90
4.4 The contribution of assessment to the CEFR 93
4.5 Conclusions: Taking the CEFR forward 96

5 Asset Languages: A formative framework for language learning 99
5.1 The origins of Asset Languages 100
5.2 Designing the Asset Languages framework 107
5.3 Development of the external assessment: Pilot phase 122
5.4 Developing the scheme 2005–08 135
5.5 Research around Asset Languages 138

5.6 Teacher Assessment: Less formal accreditation of learning 140
5.7 Conclusions: The lessons of Asset Languages 148

6 The European Survey on Language Competences: Informing language policy 154
6.1 The significance of the survey for Cambridge English 154
6.2 Background to the survey 155
6.3 The tender: Language tests and the CEFR 156
6.4 Language test development 159
6.5 Questionnaire development 168
6.6 Sampling 170
6.7 Standard setting 175
6.8 Outcomes of the European Survey on Language Competences 192
6.9 Conclusions: How to interpret the European Survey on Language Competences 198

7 Frameworks for the future 203
7.1 Frameworks so far 203
7.2 Beyond the CEFR: A framework for language education 204
7.3 Engagement in education: Impact studies and the ESLC 208
7.4 Learning Oriented Assessment 211
7.5 In conclusion 215

Appendices
Appendix A: CEFR Common Reference Levels: Qualitative aspects of spoken language use 217
Appendix B: Sample illustrative descriptors 220
ALTE Can Do project: Example statements 220
Cambridge ESOL Common Scale for writing 223
Cambridge ESOL Common Scale for speaking 224
Appendix C: Asset Languages 226
Example Languages Ladder statements (Listening) 226
Sample generic specifications for Breakthrough, Preliminary and Intermediate stages 227
Asset Languages: The final list of languages offered from September 2008 232
Appendix D: The European Survey on Language Competences 233
Task types: A complete list 233
Illustration of CEFR levels: Writing 235

References 247
Author index 259
Subject index 262

List of tables and figures

Figure 2.1 The CEFR's socio-cognitive model of language use (Cambridge ESOL 2011:7)
Figure 2.2 A socio-cognitive framework (based on Weir 2005a)
Figure 2.3 Three basic elements of a testing situation (after Jones and Saville 2007)
Figure 2.4 The Rasch model
Figure 2.5 Items, learners and levels on a measurement scale
Figure 2.6 Keenan and Comrie's Accessibility Hierarchy for two task types
Figure 2.7 Performance level and scaffolding
Figure 2.8 Salience and score variance
Figure 2.9 A multi-faceted Rasch model including raters
Figure 2.10 Item-banking approach to scale construction and use (after Jones and Saville 2007)
Figure 2.11 Two approaches to anchoring
Figure 2.12 Ranking and rating compared for speaking (Breton 2008)
Figure 2.13 Validity evidence of linkage of examination/test results to the CEFR (Council of Europe 2009:8)
Figure 2.14 The salient aspects of CEFR Table 3: Common Reference Levels: Qualitative aspects of spoken language use
Figure 3.1 CB and PB abilities compared
Figure 3.2 Scatterplot showing relation of self-assessment to BULATS score
Figure 3.3 Schematic view of the Common Scale
Figure 3.4 Structure of the ALTE Can Do statements
Figure 3.5 Selected statements at Levels 1–5 from an example Can Do scale
Figure 3.6 Mean self-ratings (Can Do statements, Fluency) by exam grade
Figure 3.7 Probability of a candidate endorsing Can Do statements at the level of the exam taken
Figure 3.8 Data transfer through steps in analysis
Figure 4.1 Different approaches to defining scales (from Jones 2005a)
Figure 5.1 Relation of teacher assessed grades and external assessment stages
Figure 5.2 The major dimensions of the Asset Languages framework

Figure 5.3 Provisional grades derived from teacher ratings
Figure 5.4 The grading model for Asset Reading and Listening tests
Figure 6.1 The targeted test design
Figure 6.2 Marking of writing against exemplars
Figure 6.3 Two steps in the standard-setting procedure
Figure 6.4 Writing cut-offs from round 1–2 (by-task) and round 3 (by-student)
Figure 6.5 Main Study and Alignment study abilities
Figure 6.6 Calibration of 16 Can Do statements
Figure 6.7 Can Do statements, all educational systems, by skill and language tested
Figure 6.8 Can Do scores and test performance by educational system: German Reading and Listening tests
Figure 6.9 First foreign language – Percentage of pupils at each level by educational system using global average of the three skills
Figure 6.10 Second foreign language – Percentage of pupils at each level by educational system using global average of the three skills
Figure 6.11 Parents' target language knowledge (mean)
Figure 6.12 Relation between 'Parents' knowledge of target language' and language scores for English writing
Figure 7.1 Elements of LOA

Table 3.1 Correlation of BULATS scores and self-assessment by language
Table 3.2 Estimated scale adjustments by language (logits)
Table 3.3 Grades of candidates taking CPE and FCE exams, December 1992
Table 3.4 Correlations between exam level, Can Do and Fluency self-ratings, grouping by exam level achieved
Table 4.1 Summary of the development of the CEFR (from Cambridge ESOL 2011:8)
Table 5.1 A Languages Ladder stage (Listening, Breakthrough – Grades 1–3)
Table 5.2 A comparison of general qualifications and Asset Languages
Table 5.3 Languages Ladder stages and CEFR levels
Table 5.4 Asset Languages and General Qualifications in the National Qualification Framework
Table 5.5 Asset Languages and Cambridge English levels within the National Qualifications Framework
Table 5.6 Provision for three contexts of learning across levels
Table 6.1 Domain distribution across Levels A1–B2
Table 6.2 Example of INES model: Concepts related to informal language learning opportunities

Table 6.3 CEFR Can Do statements included in Student Questionnaire
Table 7.1 Selected volumes in the SiLT series showing changing focus over the period 1995–2013 (titles abbreviated)
Table A1 Common Reference Levels: Qualitative aspects of spoken language use
Table C1 Final list of languages offered from September 2008
Table D1 Main Study Listening task types
Table D2 Main Study Reading task types
Table D3 An A1 level task: Holiday photo
Table D4 An A2 level task: New hobby
Table D5 A B1 level task: Favourite family member
Table D6 A B2 level task: Exchange student

Acknowledgements

Thanks first to the Series' Editors for coming up with the original idea for a book on multilingual framework construction. The idea grew out of work on a Studies in Language Testing (SiLT) series volume devoted to the Asset Languages project (now the subject of Chapter 5). Barrie Hunt, who directed that project, contributed the first literature review and draft text of the full-length book; Karen Ashton did further work on the text. Alas, the passage of time and the demise of Asset Languages made the book less relevant; nonetheless, the Asset Languages story is still well worth telling, and the chapter which I have assembled from Barry and Karen's work contributes much to the multilingual frameworks theme of this volume, and I hope does justice to the scale and ambition of that project. Karen also has my thanks for the European Survey on Language Competences – not so much with respect to the book chapter, but to her work in managing that very demanding project to its successful conclusion.

I must also thank the Series' Editors for agreeing to my use of the first person within this text. I have found it useful to have this option, to make clearer the perspective that this text takes to the work described. Like a number of books in the SiLT series, it offers a historical account of an area of research within Cambridge English Language Assessment, necessarily personal, given my involvement in the work described, and emphasising a practical perspective on the projects described, as much as the theoretical or academic issues involved.

A quick check online finds increasing support for using the first person in academic writing, although one writer suggests that to do so is 'a privilege that must be earned by first demonstrating a scholarly approach to the subject'. The reader must decide if I have achieved that.

Many thanks to the two external reviewers of the manuscript: Dorry Kenyon and Craig Deville, who contributed greatly to the coherence of the final text (and its voice); and to Evelina Galaczi for being such a kind and helpful editor.

Thanks to Blackwell for allowing the reuse of graphics first used in the Spolsky and Hult (Eds) volume *The Handbook of Educational Linguistics*.

The danger of adopting the first-person voice is that it might sound like a claim to have done everything single-handed, whereas of course, in the 21 years at Cambridge English that have now come to an end I have benefited greatly from the collaboration with like-minded colleagues, in the Research

and Validation Group and across the whole organisation, and I make extensive reference to their writing in this text. I hope that I have been successful in communicating some of the creativity and goal-focused enterprise that has characterised the work of the organisation over those twenty-something years.

Finally, my first person voice gives me for the first time a chance to use an acknowledgement page for the time-honoured purpose of thanking my family; so thank you for everything, Iwona, Zosia and Alex.

Neil Jones
Cambridge
May 2014

Series Editors' note

A number of previous volumes in the Studies in Language Testing (SiLT) series have relevance to a discussion of frameworks. A recent volume by Martyniuk (Ed) (2010), SiLT 33, *Aligning Tests with the CEFR* presented a selection of 12 papers which provided a number of perspectives on the process and outcomes of attempts to align examinations to the Common European Framework of Reference (CEFR) using the Manual provided by the Council of Europe.

The CEFR is a widely used, common framework of reference based on six broad reference levels and an 'action-oriented' approach to language teaching and learning. Within a relatively short period of time it has become highly influential in Europe and beyond as a helpful way of articulating objectives for language teaching and learning. The CEFR has certainly helped to raise awareness of language issues and has provided a useful focus for researchers, policy makers, assessment providers, and teachers.

Jones uses the CEFR as a major point of reference for all the projects presented in this volume, even those, like the Cambridge ESOL Common Scale and the ALTE Framework, which predate its publication by several years. He argues that adopting this perspective allows us to see these developments as inevitably heading towards the CEFR.

However, a word of caution is in order. The Series Editors, in their note for SiLT 33, made clear their concerns, clearly shared by Jones in this volume, that the CEFR has been adopted, interpreted or taken on a role as a fixed standard or set of standards, even though it perhaps was not originally designed as such. The CEFR as presently constituted does not provide an adequate framework for language test development or validation. Cambridge English language examinations have sought to make good its deficits and so provide the language testing community with a potentially more useful framework for validation and test development purposes. No claims are made for the Cambridge English approach being better than other contemporary frameworks but its operational usefulness is testimony to the fact that others may derive similar benefit from it.

Lynda Taylor, in SiLT 35, *Examining Listening*, summarises the Cambridge English position succinctly:

> It is only appropriate at this point to acknowledge the existence of other important frameworks and models that are available to language testers

> and examination boards. These include Evidence-Centered Design (ECD) as proposed by Mislevy, Steinberg and Almond (2002, 2003; see also Mislevy, Almond and Lukas 2003), and Assessment Use Argument (AUA), as set out by Bachman (2005) and Bachman and Palmer (2010). Other test providers have found these to be an accessible and fruitful way of guiding their practical test design and validation, as demonstrated by Chapelle, Enwright and Jamieson (2008). However, Cambridge ESOL has found the socio-cognitive model, first offered in Weir (2005) and subsequently refined through the experience of applying it to operational tests, to match well with the kinds of tests the examination board produces, addressing the validation questions that arise and providing some of the answers that are needed. The model has proved to be both theoretically sound and practically useful over a number of years in relation to a variety of different examinations produced by Cambridge ESOL and for this reason is used as the framework for description and analysis in this and the companion volumes (2013:2–3).

From the early 1990s, Cambridge ESOL worked to develop an empirically derived common scale that allowed for the systematic ordering of its examinations according to level (see the Series Editors' note in SiLT 1 (1995)). The empirical underpinning for the system was achieved by introducing an item-banking approach. Item banking is an application of Item Response Theory (IRT). It involves assembling a bank of calibrated items – that is, items of known difficulty. Designs employed for collecting response data ensure a link across items at all levels. The Cambridge ESOL Common Scale, a single measurement scale covering all Cambridge ESOL levels, was constructed with reference to these objective items. The Common Scale thus relates different testing events within a single frame of reference, greatly facilitating the development and consistent application of standards.

Since the inception of the Common Scale many millions of candidates at all proficiency levels have taken the Cambridge English examinations and their responses have allowed the scale to be incrementally refined based on analyses of this data within the framework. (See the paper for the Council of Europe by North and Jones (2009) to accompany the revised Manual; also Maris (2009) for discussion of test equating using IRT in the context of standard setting in the collection of papers edited by Figueras and Noijons (2009).)

The Cambridge English testing system has developed alongside the CEFR over the past two decades; it is now able to provide rich data and analysis to help refine the CEFR as it applies to the English language. This is an important role for a responsible organisation to fulfil and very much in keeping with the original intentions of the Council of Europe. The aim is to facilitate understanding and collaborative activities rather than to regulate or dictate to others what they should or should not do. An example of this in practice is the English Profile Programme (EPP) (see also *Research Notes* 33).

A major objective of the EPP is to analyse learner language to throw more light on what learners of English can and cannot do at different CEFR levels, and to address how well they perform using the linguistic exponents of the language at their disposal (i.e. using the grammar and lexis of English). One of the main inputs to this analysis is provided by the Cambridge Learner Corpus which contains 35 million words of learners' written English from Levels A2 to C2 of the CEFR. The EPP research team are already providing evidence of 'criterial features' of English which are typically found in the writing of learners at the different CEFR levels. Of course this data alone does not provide an adequate sample and so part of the EPP involves the collection of additional data from learners within the 'EP Network', including more written data and also focusing on spoken English as well.

In addition Cambridge English has commissioned a number of 'construct volumes' in the SiLT series to assemble and present additional evidence that the examinations offered by the board are well grounded in the language ability constructs they are attempting to measure. An explicit socio-cognitive test validation framework has been developed which enables examination providers to furnish comprehensive evidence in support of any claims about the soundness of the theoretical basis of their tests (see Weir (2005a), Shaw and Weir (2007), Khalifa and Weir (2009), Taylor (Ed) (2011), and Geranpayeh and Taylor (Eds) (2013) in this series).

Examination boards and other institutions offering high-stakes tests need to demonstrate how they are seeking to meet the demands of validity in their tests and, more specifically, how they actually operationalise criterial distinctions between the tests they offer at different levels on the proficiency continuum. The series of construct volumes develops a theoretical framework for validating tests of second language ability which then informs an attempt to articulate the Cambridge English approach to assessment in the skill area under review. The perceived benefits of a clearly articulated theoretical and practical position for assessing skills in the context of Cambridge English tests are essentially twofold:

- Within Cambridge English – this articulated position will deepen understanding of the current theoretical basis upon which Cambridge English assesses different levels of language proficiency across its range of products, and will inform current and future test development projects in the light of this analysis. It will thereby enhance the development of equivalent test versions and tasks.
- Beyond Cambridge English – it will communicate in the public domain the theoretical basis for the tests and provide a more clearly understood rationale for the way in which Cambridge English operationalises this in its tests. It will provide a framework for others interested in validating their own examinations and thereby offer a more principled basis for

comparison of language examinations across the proficiency range than is currently available.

Cambridge English Language Assessment now feels it is in a position to begin a systematic and empirically based approach to specifying more precisely how the CEFR can be operationalised for English, and this in turn will lead to better and more comprehensive illustrative descriptors (particularly at the bottom and top of the scale). In this way the CEFR will become the useful tool that it was intended to be.

Another volume in the SiLT series relevant to the themes pursued by Jones in this volume is SiLT 36, *Exploring Language Frameworks* (Galaczi and Weir (Eds) 2013) which presents an edited volume of 21 papers from the proceedings of the 4th International ALTE conference held in Krakow covering the area of Frameworks and Social Contexts, Frameworks and Educational contexts and Frameworks and Practical Issues.

ALTE, the Association of Language Testers in Europe, includes many of the world's leading language assessment bodies among its 34 members. Together with over 40 affiliates, ALTE members represent the testing of 27 European languages. In its work to promote common standards and the transnational recognition of language skills certification, ALTE has done much to encourage quality and fairness in language testing. The development of ALTE's Code of Practice and Quality Management System are key milestones in the organisation's history, as are the previous ALTE conferences held in Cambridge, Berlin and Barcelona. This fourth conference built upon the success of three previous ALTE Conferences: the first held in Barcelona in July 2001, hosted by the Generalitat de Catalunya, on the theme of 'European Language Testing in a Global Context' to celebrate the European Year of Languages; the second in Berlin in May 2005, hosted by the Goethe-Institut, on the theme of 'Language Assessment in a Multilingual Context' to support the 50th Anniversary of the European Cultural Convention and the third held in Cambridge in April 2008, hosted by Cambridge ESOL. Edited proceedings from these events were published as Volumes 18, 27 and 31 in the SiLT series. Most recently, the 5th International ALTE conference on 'Language Assessment for Multilingualism: Promoting Linguistic Diversity and Intercultural Communication' was held at Cité Internationale Universitaire de Paris in April 2014 and papers from this event will be published in due course.

These various collections of papers in the SiLT series covered many topics in the area of multilingual assessment and a number of papers in them have invariably received positive reviews for their usefulness to practitioners in the field. However, their very number and diversity precludes them from offering a clear, comprehensive, unified overview of the area of multilingual assessment. Neil Jones sets about making good this gap by providing a multilingual

proficiency framework based on his experience of working in this area for Cambridge English Language Assessment over a period of more than 20 years.

It is this modern period in the history of Cambridge's involvement with language testing that the book addresses; the period starting in 1989 which Weir (2013a) labelled the birth of professionalism at Cambridge, following the catalyst provided by the Cambridge-TOEFL Comparability Study (Bachman, Davidson, Ryan and Choi 1995, SiLT Volume 1).

In **Chapter 1** Jones deals with a number of issues that multilingual assessment must face up to. The introduction considers the nature of comparability within an assessment framework, emphasising that practical utility is the primary goal.

Chapter 2 offers an overview of the process of constructing a multilingual framework. It advances a model of how in the best possible world one would order the stages in the process. He argues we should never think of a framework as complete: validation is a continuous process, and there are always further questions to ask. His model identifies the following stages:

- scoping and construct definition
- scale construction
- alignment across languages and contexts
- standard setting
- validation.

These principles in Chapter 2 are distilled from a rich body of practice, and it is to the more detailed historical account of that practice that he then turns. The subsequent chapters follow the development of related thematic strands or particular projects, beginning with the earliest work on scale construction at Cambridge, from the beginning of the 1990s.

Chapter 3 covers a range of early research and development projects conducted by Cambridge ESOL. Topics covered in Chapter 3 include:

- computer-adaptive testing (CAT)
- the Cambridge ESOL Common Scale
- the ALTE Can Do project.

He describes CAT as offering quite a practical low-stakes context for doing basic research on plurilingual competence. The Common Scale project was a long-term endeavour to place all Cambridge ESOL assessments on a single proficiency scale based on an IRT measurement model. One of ALTE's stated aims was 'to establish common levels of proficiency in order to promote the transnational recognition of certification in Europe'. To this end, a long-term project was envisaged with the final aim of establishing a framework of levels within which meaningful comparisons between qualifications in different languages, gained in various states of the European Union, could be made.

Chapter 4 critically examines whether the CEFR can be considered a universal standard. In presenting the CEFR Jones looks critically at this descriptive framework, and considers the nature of levels, interpretation and description of performance. He looks at how the development of the CEFR intersects with that of the Cambridge English exam suite, and with other scaling and framework projects that Cambridge ESOL and partners in ALTE embarked on in the 1990s. He then examines the problems experienced with using the CEFR as a basis for test construction.

Chapter 5 extends the range of enquiry of multilingual frameworks beyond the concerns of testing. Jones argues that the frameworks described in the first four chapters of the volume are for the most part focused on assessment, specifically of language ability, and that they stand outside the educational process. He argues that as exam providers move to involve themselves more closely in learning, and begin to define their business as that of 'language education' rather than 'proficiency testing', they will need different, broader frameworks. He explores such possibilities through the lens of the Asset Languages Scheme, a case study of a multilingual framework which set out to promote communicative ability as the primary goal of language education. He describes in detail the context of its development and the practical constraints that were encountered.

Chapter 6 examines the case study of Cambridge's involvement with the European Survey on Language Competences (ESLC) project. Jones (2013b: 5) argues that this truly international project provided a new context for applying the well-developed theoretical models and the operational experience of working with multilingual frameworks accumulated in the previous two decades. The project provided a compelling portrait of the successful language learner:

> a language is learned better where motivation is high, where learners perceive it to be useful, and where it is indeed used outside school, for example in communicating over the internet, for watching TV, or travelling on holiday. Also, the more teachers and students use the language in class, the better it is learned.

He notes that the conclusions from the project were perhaps not surprising: they probably confirmed what we already believed. However, he argues it is an important achievement that the ESLC has provided empirical evidence in support of them. The key for Jones is language being used for motivated, purposeful communication. It is this which favours learning: we learn in order to communicate, and we learn by communicating. Moreover, the ESLC showed that this ideal learning situation is approximated only in some countries, and effectively, only for English.

In **Chapter 7** Jones first describes how over the period 1989–2013 the focus of work on scales, scaling and framework construction has continually

shifted, reflecting the evolving priorities of the Research and Validation Group in Cambridge English Language Assessment since its inception in 1989. The initial focus was on the analysis, the administrative systems and the skills needed to develop measurement. Ensuring quality of measurement and a grasp on reliability was an urgent issue, as the Cambridge-TOEFL Comparability Study had shown (Bachman et al 1995 SiLT volume 1). Next there came a greater focus on interpretation. The ALTE Can Do project and work to link this to the CEFR aimed to provide accessible characterisations of what it means to achieve a particular Cambridge English exam level. This then shifted to a more explicit consideration of construct validity, culminating in the four construct volumes in the SiLT series we described above: Shaw and Weir (2007), Khalifa and Weir (2009), Taylor (Ed) (2011), and Geranpayeh and Taylor (Eds) (2013). Progress having been made on these major aspects of validity and reliability, attention has shifted more recently to aspects of test use, and the development of impact studies as a research priority. Learning Oriented Assessment (LOA) is a new focus which develops logically out of the study of test impact.

Jones looks beyond the CEFR towards a framework for language education. He argues that two foreign language learning contexts in particular are not best treated by the CEFR:

- young learners, because there is no explicit treatment of cognitive stages
- Content and Language Integrated Learning (CLIL) because language for learning is not clearly distinguished from language for social use.

Taking these two factors into account requires us to supplement the proficiency dimension by two additional dimensions – age and academic content area – thereby enabling us to describe a learner at a specific proficiency level, at a specific age, studying a specific subject.

He feels that in the past we were concerned with measuring foreign language proficiency, rather than with providing a comprehensive system within which language policies can be developed and implemented. He presents a model of LOA which Cambridge English is currently developing, as a theory of action for achieving positive impact, particularly where Cambridge English exams are adopted in institutional educational settings, that is, as a significant intervention in language learning. Cambridge English is devoting considerable research effort to studying the impact of its exams in such settings, where it is possible to work with local partners to create maximally beneficial links between classroom practice and the examination which is a final target, with the aim of achieving 'positive impact by design' (Saville 2012).

For Jones, LOA offers a systemic model of assessment operating on multiple levels in an educational context and takes on many different forms. It encompasses both the macro level of framing educational goals and

evaluating outcomes, and the micro level of individual learning interactions which take place in the classroom or outside it. It defines complementary roles for teaching and assessment expertise. Successful language learning requires a focus on language for purposeful communication. Language exams which test purposeful communication may move teaching in the right direction, and the complex, coherent and inclusive conception of assessment provided by LOA represents a more explicit model of intervention within which to pursue closer alignment of assessment and teaching.

In sum this volume is a valuable addition to the SiLT series and will be of interest to everybody working in the field of multilingual assessment. Jones brings together in a single volume a description of 20 years of work at Cambridge English Language Assessment to develop multilingual assessment frameworks. He covers the development of the ALTE Framework and Can Do project; work on the Common European Framework of Reference and the linking of the Cambridge English exam levels to it; Asset Languages – a major educational initiative for UK schools; and the European Survey on Language Competences. He proposes a model for the validity of assessment within a multilingual framework, and while illustrating the constraints which determined the approach taken to each project, makes clear recommendations on methodological good practice. He looks forward to the further extension of assessment frameworks to encompass a model for multilingual education.

Cyril J Weir and Michael Milanovic
April 2014

Abbreviations

ACTFL	American Council on the Teaching of Foreign Languages
ALTA	Automated Language Teaching and Assessment
ALTE	Association of Language Testers in Europe
BEC	Business English Certificate
BICS	Basic Interpersonal Communication Skills
BULATS	Business Language Testing Service
CAE	Certificate in Advanced English
CALP	Cognitive Academic Language Proficiency
CAT	Computer-adaptive Testing
CBT	Computer-based Testing
CEFR	Common European Framework of Reference
CELS	Certificates in English Language Skills
CIEP	le Centre international d'études pédagogiques
CiLT	The National Centre for Languages
Cito	National Institute for Educational Measurement
CLIL	Content and Language Integrated Learning
CPE	Certificate of Proficiency in English
DfES	Department for Education and Skills
EASA	European Academic Software Award
EFL	English as a Foreign Language
EPMA	Educational and Psychological Measurement and Assessment
EPP	English Profile Programme
EPS	Exam Processing System
ESLC	European Survey on Language Competences
FCE	First Certificate in English
FINGS	Finnish, Irish, Norwegian, Greek and Swedish
FSI	Foreign Service Institute
GERM	Global Education Reform Movement
ICT	Information and Communication Technology
IEA	International Association for the Evaluation of Educational Achievement
IELTS	International English Language Testing System
INES	International Indicators of Education Systems
IRT	Item Response Theory
ISCED	International Standard Classification of Education
KET	Key English Test

KoBaLT	Komputer Based Language Testing
L1	First Language
L2	Second Language
LIBS	Local Item Banking System
LLL	Learning Ladder for Languages
LOA	Learning Oriented Assessment
LTT	Latent Trait Theory
LUCIDE	Languages in Urban Communities – Integration and Diversity for Europe
MCQ	Multiple-choice Question
NQF	National Qualifications Framework
OECD	Organisation for Economic Cooperation and Development
Ofsted	Office for Standards in Education
PB	Paper-based
PET	Preliminary English Test
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PPS	Probability Proportional to Size
QCA	Qualifications and Curriculum Authority
SD	Standard Deviation
SiLT	Studies in Language Testing
TIMSS	Trends in International Mathematics and Science Study
TOEFL	Test of English as a Foreign Language
UCLES	University of Cambridge Local Examinations Syndicate

1 Multilingual frameworks: A practical pursuit

The origins of language testing at Cambridge English go back to 1913, when the Certificate of Proficiency in English (CPE) examination was launched. On the occasion of its centenary in 2013 Cambridge English Language Assessment published two volumes in the Studies in Language Testing (SiLT) series: *Cambridge English Exams – the First Hundred Years* (Hawkey and Milanovic 2013), a history of Cambridge English through its staff members, and *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012* (Weir, Vidaković and Galaczi 2013). The present volume may be seen as another contribution to the historical record, covering what might be called the modern era of Cambridge English.

In terms of personalities the modern era started with the appointment of Peter Hargreaves as Director of English as a Foreign Language (EFL) in 1988. An EFL Evaluation Unit was set up in the summer of 1989, headed by Mike Milanovic, working with Nick Saville. This unit was to establish a validation programme and research agenda specifically focusing on the EFL examinations. It was the first unit of its kind within a UK EFL examinations board. An article entitled *EFL research at UCLES* (University of Cambridge Local Examinations Syndicate 2000) describes the range of research projects undertaken in those early years.

One major project was the Cambridge-TOEFL Comparability Study (Bachman, Davidson, Ryan and Choi 1995), one outcome of which was that 'far greater attention was paid to scoring validity to bring Cambridge's procedures more in line with the psychometrically sophisticated approach that had long been part of professional language assessment in the United States' (Weir 2013b:423). The developments presented in this volume might be seen to have originated in such concerns.

The focus is on a recurring and evolving theme in the development of a Cambridge English approach to assessment: the construction and interpretation of assessment frameworks, particularly those with a multilingual dimension. It is the purpose and use of such frameworks which will be foregrounded. This poses the question of how the work described in this volume should be evaluated, in terms of its effect, quality and utility. Two points of reference are relevant: firstly, models of validity and validation, and secondly, a particular standpoint within the philosophy of social science.

The validity perspective chiefly adopted here is that of the socio-cognitive model as outlined in the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001) and more fully developed by Weir and Cambridge English collaborators in a series of volumes in the SiLT series (Weir 2005a, Shaw and Weir 2007, Khalifa and Weir 2009, Taylor (Ed) 2011, Geranpayeh and Taylor (Eds) 2013). The model is introduced in more detail in section 2.1. It is chosen in preference to, say, the assessment use model (Bachman 2005) perhaps more familiar to some readers, as it has impacted most directly on Cambridge English practice.

Regarding the second, Cambridge English Language Assessment has laid claim to a realist philosophical stance (Sayer 1992), which will also be introduced in more detail in Section 2.2.1, but in broad outline might be described as a commitment to do useful work with the best tools practically available. This book attempts to illustrate the approach in action.

Both perspectives are necessary and both rank consequential outcomes highly in evaluating the impact of an assessment on a context of learning. As the later chapters will make clear, this is an empirical endeavour: a search for means that serve practical ends. The approach adopted in this book emphasises that language testing is a human endeavour which proceeds within a real world of practical constraints. Psychometric models capture just a part of it.

This book does not set out to provide a comprehensive, academic, technical account of scaling and measurement, partly because I am not qualified to write such a book, but chiefly because it is intended to be a rather different kind of book, focused on the nature of language testing as a practical pursuit. It is largely devoted to assessment frameworks in the experience of Cambridge English Language Assessment (and Cambridge ESOL as it was formerly known), but that defines quite a broad and multilingual range. It includes collaboration with partners in the Association of Language Testers in Europe (ALTE) (see Section 3.4), engagement with language learning in English schools through the 25-language Asset Languages project (Chapter 5), and the delivery of the European Survey on Language Competences (ESLC) (Chapter 6). It also includes an increasing number of bilateral collaborations with ministries of education or other educational institutions, several of which are not simply concerned with the testing of English, but focus on approaches to language education more widely construed. These multilingual collaborations thus highlight the relevance of the work described in this volume for language education policy and the promotion of better language learning (Chapter 7). As noted above, this volume also contributes something to the historical record, and hopefully aspects of this are not without some intrinsic interest. It is organised as follows.

The major part of the text (Chapters 3 through 6) offers a chronological account of the major scaling and framework construction projects undertaken by Cambridge English in the last 20 or so years. In preparation for

this some introduction seems necessary to the concepts and the measurement technology common to all these projects. This is the purpose of Chapter 2, which defines terms concerning scales and frameworks, and breaks down framework construction into a series of steps – even if, as the later chapters show, few projects actually follow such an ideal sequence, given the practical constraints within which all developments operate. The proposed sequence begins with construct definition – attempting to be explicit about the nature of the skill to be scaled. It proceeds through scale construction, which requires the choice of appropriate psychometric statistical approaches. Cambridge English Language Assessment has opted to use forms of the Rasch model for this purpose, and it is this model which is presented here. Further steps include the alignment of different language tests and learning contexts, addressing their location in a framework relative to each other – something which ideally would be established prior to the setting of absolute standards. The chapter concludes with a discussion of validity relating to how frameworks can contribute to interpretation of learners' language skills.

Chapter 3 begins the chronological account with the adoption of the Rasch model by Cambridge English Language Assessment in the early 1990s, and the gradual transition towards an item-banking approach to test construction and score interpretation. Of central importance at this early stage was a project to align the multi-level suite of Cambridge English examinations to a single proficiency scale. At the same time several projects to develop computer-adaptive tests also went ahead, precisely because they were seen as of more marginal importance. A multilingual dimension was opened up in collaboration with the recently founded ALTE, aimed at aligning the partners' exam systems to levels within a single framework.

Chapter 4 is devoted to the CEFR, the influential interpretive framework to which Cambridge English examinations and the ALTE Framework were eventually to claim formal alignment. It treats the CEFR as an extensible, open framework of reference, and points out the positive contribution that language assessment expertise has made and continues to make to the process of implementing the CEFR and extracting practical benefits from working with it.

Chapter 5 offers a detailed account of the Asset Languages project, a unique opportunity to develop a complex multilingual assessment framework on behalf of the Department for Education in England, as an alternative to standard school examinations. This collaboration between Cambridge English and Oxford, Cambridge and RSA Examinations (OCR), the sister organisation within Cambridge Assessment which provides school examinations, set out to implement the vision of a learning-oriented language framework with a strong focus on communicative language skills. That this vision finally found too little acceptance offers important lessons as to the necessary conditions for bringing about profound systemic changes in language education.

Chapter 6 gives an account of the first European Survey on Language Competences (ESLC), delivered on behalf of the European Commission by a multinational consortium led by Cambridge English. The survey reported on levels of achievement in first and second foreign languages, studied to the end of lower secondary education, and using the CEFR as the reporting framework. It encompassed the five most-taught languages in Europe (English, French, German, Italian and Spanish), and from a language assessment perspective it is significant not simply for the technical challenge of developing comparable tests and a common standard across these languages, but also in terms of its concrete outcomes, which demonstrated the huge difference in levels of achievement in language learning across the participating countries.

This theme is picked up in Chapter 7, which reviews the achievements of the last 20 years, and looks forward to the multilingual frameworks of the future. These, we predict, will build on their summative, outcomes-focused strengths to develop a much stronger focus on supporting the process of learning.

2 Constructing a multilingual framework

This chapter offers an overview of the processes involved in constructing a multilingual framework. It advances a model of how in the best possible world one would order the stages in the process, and to that extent offers an accessible expositional approach. In practice, as described in later chapters concerning actual cases, we never operate in the best possible world. Various constraints apply, and a project is likely to proceed on a number of levels simultaneously, working with provisional conclusions and relying on an iterative process of progressive approximation. Indeed, we should never think of a framework as complete: validation is a continuous process, and there are always further questions to ask.

The model advanced below identifies the following stages:

1. Construct definition
2. Scale construction
3. Alignment across languages and contexts
4. Standard setting
5. Validation

2.1 Construct definition

2.1.1 Achievement and proficiency tests

First of all we should make clear that the measurement scales and frameworks presented in this volume all share a focus on testing *proficiency*, rather than *achievement* (or so at least we would like to believe). To introduce this traditional distinction let us consider how scales arise in different contexts for different purposes. As North (2000:13) points out, language proficiency scales have originated in three contexts:

1. As *rating scales* – most influentially, that of the US Foreign Service Institute (FSI), dating from the 1950s.
2. As *examination levels*, for example, the development over time of the Cambridge English exams.

3. As *stages of attainment*, associated with the objectives of an educational system or course of instruction, for example, as expressed by the UK's National Curriculum.

Different purposes are evident here. The FSI rating scale effectively serves to *define* levels of language proficiency, together with a tool for assessment. The evolution of the Cambridge exam levels or the ALTE Framework can be seen as approaches to enhancing the *interpretation* of existing assessments, by relating exam levels to the observable abilities of candidates in the world beyond the test. The stages of attainment defined in the UK National Curriculum start from what is to be taught – the nature of the subject itself. They reflect, among other things, a judgement as to what constitutes a rational, desirable and practical sequence of teaching objectives.

Therefore, while all scales doubtless start from some basis in observation and analysis, we can distinguish two orientations: a descriptive approach based on how we think learners really progress, or a prescriptive approach based on how we wish them to progress. Thus, when the descriptors of the American Council on the Teaching of Foreign Languages (ACTFL) Guidelines are criticised as being 'groundless, made up—arbitrarily' (Lantolf and Frawley 1992:35, quoted by North 2000:30), it is because they are found to have the latter orientation, when in the judgement of the critics they ought to have the former.

Proficiency tests aim to describe how learners can be expected to progress, quite possibly based on observations in first or second language acquisition, or on theories of cognitive development. Achievement tests aim to determine whether learners have progressed in relation to the prescriptions of a syllabus or course, quite possibly defined in terms of different levels of linguistic content.

In practice, of course, it may be impossible to distinguish clearly between the two purposes, because in an educational setting with the goal of teaching communicative language skills, both may well be addressed by a single test, simultaneously measuring not only mastery of what has been taught (achievement), but what use the learner can make of it (proficiency). The distinction remains an important one, because ideally proficiency tests should enable us to interpret performance in a language test better: generalising beyond the test tasks themselves, to indicate ability to manage in a wider range of situations. But everything depends on how proficiency is defined and measured; something we will return to (see Section 2.2.2.2).

2.1.2 A socio-cognitive model of language proficiency

The frameworks presented in this volume share the proficiency orientation, where proficiency is understood as communicative language ability: an ability to use language purposefully. Proficiency tests assume that all learners

for whom a test is designed can be expected to progress in roughly the same way.

Validity in language testing has been defined in terms of supporting inference to performance in some more or less defined target language use situation (Bachman 1990). What this entails is an explicit model of the skill or skills to be tested, showing how they are expected to progress and develop: that is, a *construct*.

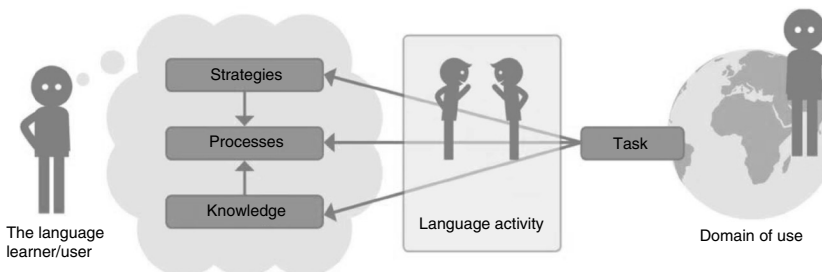
The much-quoted model of language use and learning provided by the CEFR is a good starting point. It refers to:

> . . . the actions performed by persons who as individuals and as social agents develop a range of ***competences***, both ***general*** and in particular ***communicative language competences***. They draw on the competences at their disposal in various contexts under various ***conditions*** and ***constraints*** to engage in ***language activities*** involving ***language processes*** to produce and/or receive ***texts*** in relation to ***themes*** in specific ***domains***, activating those ***strategies*** which seem most appropriate for carrying out the ***tasks*** to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences (Council of Europe 2001:9, emphases in original).

This paragraph describes a learner's cognitive apparatus (general knowledge, language competences, strategies), which develops through engaging with the communicative tasks that arise through social interaction over a range of contexts and purposes. It is a *socio-cognitive* model of language use – that is, it is about cognition demonstrated in social action, recognising that: 'language use constitutes both a socially situated and a cognitively processed phenomenon . . .' (Weir 2013b:443). A Cambridge English introduction to the CEFR (Cambridge ESOL 2011:7) offers the following graphical rendition of the above paragraph (Figure 2.1).

Figure 2.1 places language activity motivated by a communicative task squarely in the centre, as the basis for making judgements about the learner;

Figure 2.1 The CEFR's socio-cognitive model of language use (Cambridge ESOL 2011:7)



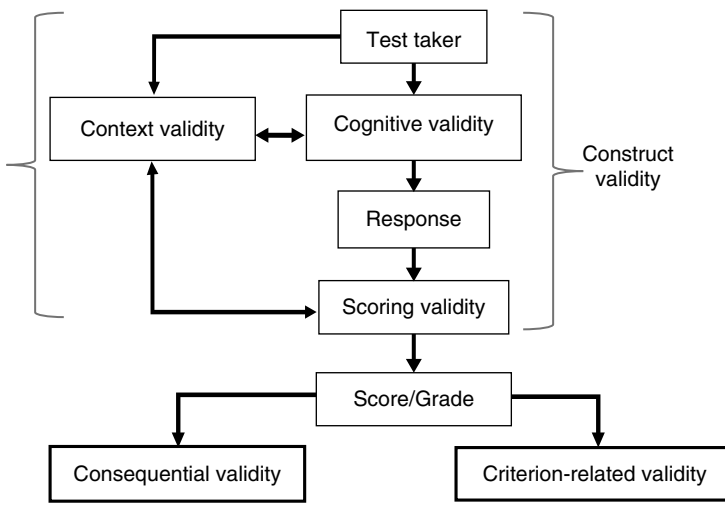
but it also indicates two possible emphases for such judgements. We might focus more on the task and the degree to which it is successfully addressed, or alternatively on the learner's cognition and what we can deduce about that. The former might be appropriate in a context where a limited range of communicative interactions can be defined, as was done in the 'Beijing speaks English' programme, which set out in the context of the Beijing Olympics to equip taxi drivers, for example, with a repertoire to communicate with their passengers. The latter might be more relevant where communicative language ability needs to be defined in broader terms. But at all events the model places motivated language activity at the centre.

Practically, to define test constructs requires considerably more detail than the CEFR provides, particularly in respect of models of cognition which would enable us to construct appropriate tasks over a wide range of proficiency levels. The CEFR itself does not set out to address the needs of language assessment at that level of detail, as pointed out by several authors (Alderson, Figueras, Kuijper, Nold, Takala and Tardieu 2004, Huhta, Luoma, Oscarson, Sajavaara, Takala and Teasdale 2002, Jones 2002, Little, Lazenby Simpson and O'Connor 2002, Morrow (Ed) 2004, Weir 2005b).

Weir (2005a) develops a socio-cognitive validity framework which is useful for an exam provider in that it maps to steps in the test design and administration process, each of which impacts on validity, and which must therefore be a focus of attention. Figure 2.2 presents this framework in outline.

Construct validity relates to what the test actually measures and whether this is what we intend, and so can be seen as the essence of validity. Weir

Figure 2.2 A socio-cognitive framework (based on Weir 2005a)



stresses 'the importance for any successful assessment system of seeking and assembling validity evidence on each of these three core aspects of validity: cognitive, context, scoring, which together constitute test construct validity' (Weir 2013b:443). Test tasks must be designed to elicit performance which gives evidence of the learner's language abilities in relation to context and cognitive validity – a model of the processes involved in performing tasks at given levels of proficiency. The theory must reflect the context of the learner – their educational background, purpose in learning the language, and so on.

Further elements within Weir's socio-cognitive framework include:

- Scoring validity, which relates to how test responses are processed and interpreted to provide a score or grade in a test. This is at the heart of scale construction – turning scores, which are essentially arbitrary numbers, into points on a measurement scale.
- Criterion-related validity, which relates to how points on the measurement scale can be interpreted as indicating abilities or performance levels defined in real-world terms.
- Consequential validity, which relates to whether the impact of the exam is positive for individual test takers or society more widely.

Four volumes in the SiLT series offer detailed analyses of the skills constructs as defined in Cambridge English exams over a range of CEFR levels: Shaw and Weir (2007) for writing, Khalifa and Weir (2009) for reading, Taylor (Ed) (2011) for speaking and Geranpayeh and Taylor (Eds) (2013) for listening. Organised around Weir's validity model, these volumes set out to supply the useful level of detail which the CEFR itself does not (see also Weir et al 2013). Weir (2005b) uses this validity model to identify omissions and shortcomings in the CEFR:

- that the CEFR descriptor scales give an incomplete picture of contextual variables and performance conditions (context validity)
- that little account is taken of the nature of cognitive processing at different levels of ability (cognitive, or theory-based validity)
- that the quality of actual performance expected to complete an activity is not made clear (scoring validity).

Weir contributes to a rich literature on the interpretation of test scores, associated with issues of social value and consequences. Messick (1989) is the influential starting point. Bachman has applied Messick's approach to language testing, through the notion of inference to domains of language use (Bachman 1990, Bachman and Palmer 1996). In the early 1990s Cambridge ESOL came under the strong influence of Messick through collaborations with Bachman. In educational measurement, Mislevy, Kane and colleagues have picked up and developed Messick's thinking: Kane (1992, 2004),

Mislevy (1996), and Mislevy, Steinberg and Almond (2002, 2003). Construct validity is discussed further in the context of the CEFR in Section 4.4.

The CEFR's 'action-oriented' socio-cognitive construct of communicative language ability cited above applies quite well to all the framework-building projects described in this volume. What varies is the degree of control or ownership of the framework. In the case of the CEFR, being an open point of reference for anyone who cares to use it as such, no control is possible, and proposals to create a 'CEFR police' to ensure its proper use have (fortunately, I would judge) not been taken up. The Council of Europe, whose Languages Policy Division sponsored the development of the CEFR, has consistently rejected such a role, even though in a related field – the European Language Portfolio – they did for some years maintain a role in certifying national versions of the document. Inevitably this has led to the development of context- or language-specific interpretations of the CEFR, which are most evident in relation to understanding of the CEFR levels, although they doubtless concern wider issues. Evidence for such divergence over levels appeared early on, in the illustrative materials for the levels provided by representatives of different languages. It was awareness of this issue which prompted one serious attempt to achieve cross-language consensus regarding the performance skill of writing (see Section 2.3.3). In 2012 the ESLC (Chapter 6) still found clear evidence for the context-relative way in which CEFR levels may be understood.

Other frameworks presented in this volume are more controlled. The ALTE Framework (see Section 3.4), which located members' exams according to a set of levels (later aligned to the CEFR) was a judgemental construction based on close analysis of test task features, and thus a collaboration involving all interested parties. Likewise the Asset Languages project (Chapter 5) and the ESLC (Chapter 6) were owned and controlled by their developers (although in the case of Asset Languages asserting control over so many languages was to become an issue). Thus all of the framework projects presented here can appeal to the same basic construct: communicative language ability, an action-oriented approach. But that construct is defined and implemented with differing degrees of explicit control, depending on the context.

What may be seen as a limitation of the frameworks described in this volume is that they are almost entirely focused on assessment. Other aspects of language teaching and learning, such as curricular objectives, teacher training, teaching materials, and so on, are largely excluded, although the Asset Languages project in particular (Chapter 5) illustrates an attempt to exert a more explicitly formative effect, and in fact points up difficulties that arise in developing valid assessments explicitly to support learning from a low proficiency level, in the absence of control or influence over specific syllabus content.

This volume thus reflects the role of assessment in relation to teaching and learning as understood by Cambridge English in the period covered. However, Chapter 7 looks ahead to possible models for defining a more complex, integrated relationship between teaching, learning and assessment, under the name of Learning Oriented Assessment (LOA), and considers the nature of an extended framework, going beyond the communicative competences described in the CEFR, which would address the needs of language education more widely construed.

2.2 Scale construction

The purpose of construct definition is to enable the development of test tasks which will validly measure the skills of interest. The next step is to administer the test tasks, collect responses, and use them to generate a measure, or scale.

In this volume the term 'scale' is used rather loosely, for any kind of depiction of a progression from less to more. A scale may be quantitative, based on some use of numbers, or qualitative, based on a description of levels, or both of these.

Regarding quantitative scales, some hold more intrinsic meaning than others. The numbers 1 to 10 may just be arbitrary labels, or they may constitute a measurement scale, like the numbers on a ruler, which divide the property of length into equal units. Measurement scales in this sense are the focus of this section.

Qualitative scales interpret progression descriptively. They may be entirely subjective, or they may be developed from data. For example, the CEFR's descriptive scales are derived from a Rasch scaling of descriptors using data from teacher ratings. Such descriptive scales have been usefully classified according to four purposes, each purpose implying a different interpretation: *user-oriented*, i.e. to make clear to end-users what exam results mean; *assessor-oriented*, i.e. to assist raters in evaluating test performance; *constructor-oriented*, i.e. to assist item writers to construct items; and *diagnosis-oriented*, i.e. to diagnose the skills profile of a learner on the basis of test results (Alderson 1991, Pollitt and Murray 1996).

Clearly, the most useful scales are those that combine effectively the quantitative and qualitative: accurate measurement and meaningful interpretation. This is the major goal of scale construction as presented in this volume. This section focuses on the statistical process of constructing measurement scales that capture the difficulty of test tasks and the ability of learners. The presentation of item banking (see Section 2.2.4) will then show how the availability of the measurement scale facilitates the development of the criterion-referenced interpretive framework.

This section introduces the Rasch model, a statistical model based on Item Response Theory (IRT), also referred to as Latent Trait Theory (LTT).

Much of the work described in this volume uses the Rasch model, which we treat as belonging to the family of IRT models. Scale construction is based on data – primarily, the responses of candidates to test tasks. This introduction focuses on objectively marked tests (that is, tests constructed of items which can be marked right or wrong, or scored on a scale, often but not always using automated marking). For the Cambridge English exam suite that primarily means Reading, Listening, and sections of the Use of English papers. Performance skills are discussed in more detail in Section 2.2.3.

2.2.1 Trait-based measurement

Before looking at the concepts involved in working with IRT it is worth considering the conception of measurement which it embodies.

The term *measurement* has considerable metaphorical power, suggesting as it does that measuring language proficiency is straightforwardly analogous to measuring physical attributes like weight, length or temperature. However, in the case of mental constructs like language proficiency things cannot be that simple. The statistical approaches described in this volume seek to enable objective measurement by reducing the differences between people – the objects of measurement – to a simple quantitative measure. That is, they define *traits* – dimensions on which individuals may be ranked lower or higher.

Constructing a trait depends on identifying broad patterns of regularity in data. Of course, people differ in many complex ways, so that the regularity will always be limited. However, the greater the regularity, the more interpretable and useful the trait, and so we will naturally attempt to maximise the regularity in the data we collect, for example in the way we construct test items.

This approach has its critics. Goldstein (2012:152) finds the Rasch model to be unrealistically simplistic, 'not only content with ignoring other explanatory variables but also insistent that only a single dimension is needed in any given application'. Goldstein argues that 'a better understanding is needed of the difference between data that "confirms" a theory by providing a good model fit, and data that allows us to explain observed data patterns using as much potentially falsifiable information as possible' (2012:156). In other words, by accepting as reality the imperfect patterns in data, we in fact see things that are not there.

The counter to this argument is to be sought not in the technical details of the statistical model, be it IRT or any other approach, but rather in the nature of the world and our ability to understand it – that is, in considerations of ontology and epistemology. In the introduction we asserted the importance of locating the practice of language assessment within a philosophy of social science. The approach which Cambridge English takes to research and to its

operational work is a realist one (Sayer 1992). Realism is a position which seeks to find a middle path between naïve objectivism on the one hand (that is, a common sense view of reality) which would lead us to act uncritically, and extreme relativism on the other, which sees objective understanding as unattainable and would prevent us acting at all. A realist view implies that:

- We may assume that there are realities in the world, even if our ability to apprehend them is fallible. That language proficiency exists, such that individuals have less or more of it in respect of a given language, would be one such reality.
- Knowledge is not so much a representation of the world as a way of doing things in it.
- We should be aware that all observation is theory laden, in that we inevitably bring previous experience and belief to bear on deciding what to observe, how to observe it and how to interpret our observations. However, being aware of this we can set out to structure our observations precisely. We can define the constructs of language proficiency explicitly, adapting them as necessary to achieve relevance to particular contexts.
- This requires a rational approach to analysing contexts and purposes, which in turn requires interpretive understanding. The role of qualitative analysis should thus be emphasised. Quantitative outcomes require qualitative interpretation.
- We should avoid thinking of language proficiency as a thing or product, but locate it in the social context of language use. We should recognise it as an emergent power which cannot be broken into its constituent parts.
- Concerning the truth of our models, the notion of falsifiability may be explicitly rejected, because there are no absolute laws in social science. We may model the mechanisms which cause things to happen, but in an open system we cannot fully control the conditions which determine whether they happen or not. Judgements of truth should be based on 'practical adequacy' (Sayer 1992:65), which we take to mean: having sufficient explanatory power to move us forward.

Therefore, the counter to Goldstein (2012) is that we have a sufficiently clear understanding of how our assessments work – a model of the cognition of learners and the features of test tasks – to defend the notion that our interpretation of test responses as indicating more or less language proficiency is meaningful and useful. This point is the critical one. Language assessment has important roles in society, and finally we should evaluate it in terms of the balance of positive and negative impacts. Or to go further, we should be aiming to maximise the positive and minimise or eliminate the negative impacts by studying how assessments are used, and progressively effecting changes. Saville (2009, 2012) describes this as 'positive impact by design'.

Goldstein in fact played a key role in obstructing the use of the Rasch model in educational assessment in England. As recounted in *The assessment revolution that has passed England by* (Panayides, Robinson and Tymms 2009), there were debates in 1981 about the use of the Rasch model which the anti-Rasch camp, led by Goldstein, was held to have won (Lacey and Lawton 1981). In consequence the Rasch model was effectively banished from educational assessment in the UK. Since then Cambridge English Language Assessment has successfully put the Rasch model at the centre of its assessment methodology. From our Cambridge English perspective it is clear that scaling has enabled us to develop stable criterion-referenced interpretations of test performance, and that criterion reference is critically important. Without it exams inevitably become self-referential, the results uninterpretable, and the impact on learning potentially disastrous. Support for this thesis is provided in Section 6.8, where the results of England in the 2012 ESLC are discussed.

Of course, the metaphor of measurement as applied to mental traits must not be taken too literally. In a discussion of trait-based measurement and its relation to the concepts of reliability and validity Jones (2012:351) points out that:

> Firstly, it suggests that language proficiency is something real in a person's head which can be quantified, like their height or weight. Such a conception does not accord with current understanding of human abilities. Currently language assessment takes a more socio-cultural view of proficiency, as something which is situated in social use and interaction, and to some extent inseparable from it.
>
> Secondly the metaphor implies that language proficiency, like temperature, has a single unique meaning. If this were true then a single language test would be enough to measure all learners in all contexts of language learning. However, we recognise that every context of learning needs to be treated on its own terms. A one-size-fits-all approach is not possible.

Or in other words: unlike temperature or length, measurement of language ability must begin with construct definition. Jones (2012) argues that in trait-based approaches, conceptions of validity and reliability are closely linked. They both relate to the idea of measuring one thing at a time, for example, by identifying skills such as Reading, Writing, Listening and Speaking as distinct *unidimensional* traits, and testing them separately. When all items in a test measure the same thing they demonstrate high *internal consistency* – a feature also commonly used as a convenient proxy for reliability. Objectively marked tests such as of Reading and Listening tend to be composed of a large number of items – the more items, the higher the reliability. This is because the items collaborate with each other to accurately locate a learner on a proficiency scale. By the same line of trait-based reasoning, low internal

consistency may imply not just low reliability, but that the construct itself lacks clear definition, or in fact is made up of a number of components (Weir, Huizhong and Yan (Eds) 2000).

Should validity really be associated with internal consistency in this way, given the complex, multidimensional nature of language ability? In fact it seems that despite this complexity, test items can elicit responses which fit a simple trait model quite well. This is because different aspects of language competence are interrelated and tend to correlate with each other. Thus a test can demonstrate good internal consistency and reliability, even if it contains tasks which focus on different aspects of the same skill, or elicit different kinds of response. As Wood (1991:138) points out, high internal consistency need not imply homogeneity. This is fortunate, because communicative language ability is of its nature complex and multidimensional, and cannot be broken down to its constituent parts without destroying the very thing we wish to measure.

The assumption of unidimensionality in a trait-based approach might seem unrealistic when the trait spans several proficiency levels, given that skills change fundamentally as they develop. Each level has different salient features. Reading at A1 seems a completely different activity to reading at C2. In fact, it is perfectly practical to construct a trait covering a range of levels. While it is impossible to compare directly A1 learners with C2 learners, because there is no single task on which both could demonstrate their ability, it *is* possible to compare individuals at every small step along the path from A1 to C2. In an item-banking approach proficiency scales are constructed by linking carefully across levels. Unidimensionality has meaning only within the range of ability where individuals can reasonably be compared.

The approach to trait-based measurement presented in this volume is based on Item Response Theory (IRT), introduced in Section 2.2.2. However, all those more traditional approaches to testing which involve scoring learners' performance in a test and ranking them according to that score can also be considered to be trait based. Trait-based testing done in one way or another, more or less amenable to meaningful interpretation, is the familiar norm in school examinations and similar assessments.

Whether constructs such as 'knowledge', 'skill' or 'ability' can be adequately measured using trait-based conceptions of measurement is something that has been questioned (e.g. Frederiksen, Mislevy and Bejar 1993). Trait-based approaches are more appropriate to *summative* assessment purposes – that is, assessments which provide a summary measure of the outcomes of some stage of learning. Problems arise when the assessment purpose is *formative* – that is, when we wish to achieve a more complex understanding of learners' current state of knowledge in order to help them learn more. Classroom assessments most often have a primarily formative purpose. Summative assessments look backwards, asking: what has the learner

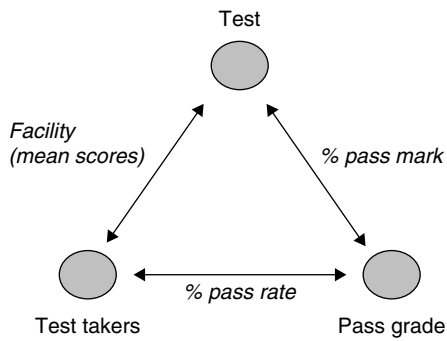
achieved up to the present moment? Formative assessment attempts to help the learner move forwards from where they are. The projects described in this volume are mostly about measurement used summatively, although the Asset Languages assessment framework (Chapter 5) had an explicitly formative strand. In Chapter 7 however, we shall consider future directions of development, which envisage supporting the process of language learning more closely. Learning Oriented Assessment (LOA) (see Section 7.4) is concerned not simply with measuring learning outcomes but intervening to improve them. This will not require us to abandon trait-based conceptions of measurement, but to complement them with other conceptions of the evidence that needs to be collected and interpreted.

2.2.2 An introduction to Item Response Theory

2.2.2.1 The problem with classical statistics

Figure 2.3 illustrates three simple statistical concepts which are familiar to anyone who has taken tests (that is, everyone): *facility*, or the mean score on a test, the *pass mark*, perhaps stated as a percentage, and the *pass rate*. Interpretation of these concepts is straightforward: as more people score more than the pass mark, so more of them will pass. But as Figure 2.3 makes clear, these statistics are not informative, because each of them only reflects a relationship between two underlying factors. Thus, facility reflects a relationship between the test takers and the test: specifically, between the *ability* of the test takers and the *difficulty* of the test items. The pass mark reflects a relationship specifically between the *difficulty* of the test items and the *standard* or passing grade which is applied. The pass rate reflects a relationship specifically between the ability of the test takers and the standard applied.

Figure 2.3 Three basic elements of a testing situation (after Jones and Saville 2007)



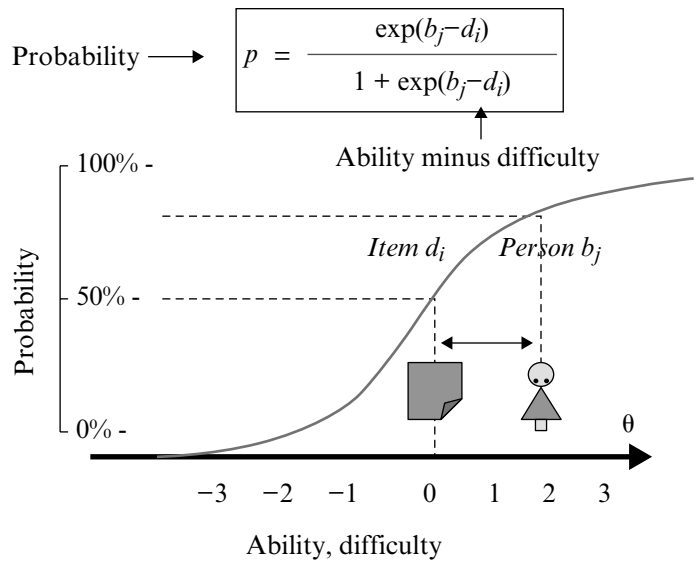
In other words, facility, pass mark and pass rate are relative concepts with no intrinsic meaning. What we are interested in knowing is the *ability* of test takers, the *difficulty* of the test and the level of the *standard*. These are (potentially) absolute values with intrinsic meaning: for example, a standard can be set in terms of a CEFR level, and a test taker can be located at that level, or below or above it, by a known margin.

2.2.2.2 IRT concepts

So, in an IRT view, what interests us are not scores as such, but the underlying features of learners and test items which lead to those scores being observed. The language proficiency continuum is a *latent trait* – that is, an underlying, invisible dimension – upon which learners, items and criterion levels of ability (standards) can all be located. To derive such abilities and difficulties from test response data requires the use of a specific statistical model. Cambridge English Language Assessment uses the Rasch model, which belongs to a class of models within IRT (Bond and Fox 2001, Hambleton, Swaminathan and Rogers 1991, Wright and Stone 1979).

Figure 2.4 includes the basic Rasch model equation, and illustrates the relationship it defines between the *probability* of a learner responding correctly to an item (the vertical y axis) and the *difference* between the item's difficulty and the learner's ability on the horizontal axis (also called the *theta scale*, hence the Greek character on the right). Probability is a value between zero (certainly

Figure 2.4 The Rasch model



wrong) and 100% (certainly right), while the horizontal ability/difficulty scale is linear, with limits of plus or minus infinity, for total test scores of 100% or zero respectively. The scale units are called *logits*, and will be further discussed below. This explains the S-shape of the curve which describes the relationship. When a person and an item are at exactly the same point on the scale the person has a 50% chance of responding correctly. The higher the person is on the scale, the higher the probability of responding correctly (and vice versa). The relation defined by the model is quite intuitive: when the person is relatively higher on the scale than the item, they are more likely than not to get it right, and when they are relatively lower they are more likely to get it wrong.

To construct a scale we must start from test data – the correct and incorrect responses given by a group of people to a group of items. The higher the total score of each person, the higher their ability. The higher the total score on each item, the lower its difficulty. That provides enough information to estimate the most likely values for all the abilities and difficulties, something that dedicated statistical software can do.

The software is necessary because estimation is not straightforward: it seeks to find the best possible fit of the data to the model, and that fit is only ever approximate. Students will get some items wrong which they would have been expected to get right, given their overall ability, and vice versa. So even after estimating the most likely ability and difficulty values for each person and item, individual responses will not be perfectly predicted by those estimated values. But this does not mean that the measurement is somehow faulty – the model works precisely because it depends on probabilities, and given enough data, probabilities can produce very accurate results. A coin is expected to land heads-up half the time, and the more throws, the closer the observed result will approach that expected outcome. Similarly, tests can produce results which are accurate to within a knowable degree of error, that error depending chiefly on the number of observations (items) in the test. Goodness of fit is important in evaluating whether the model has produced a useful measurement or not. Badly fitting data doesn't support substantive interpretation. Good measurement depends on well-defined constructs and well-written items, and you can only measure one thing at a time – hence the importance of testing language skills separately.

Finding the difficulty of test items is called *calibration*. Because the whole scale is defined through the relative *difference* in position between items and persons (ability minus difficulty) there is no meaningful zero point. So at the very beginning of scale construction we set an arbitrary point and ensure that every subsequent data set can be linked to it, by including some items which have already been calibrated. This is called *anchoring*. Developing suitably practical schemes for anchoring is one of the basic and most important steps in constructing a measurement scale.

The above description shows that in an IRT view, ability and difficulty

define each other: they arise in the interactions of learners and tasks. This notion is in fact clearly analogous to the socio-cognitive view of validity on which Cambridge English exams are based (see Section 2.1), where ability is seen to reflect observable interactions between the cognition and skills of a learner and the demands of a task. Good model fit may thus strengthen the claim for the *interactional authenticity* of test tasks (see Section 2.5.4.2).

In thinking about measurement scales it is worth trying to keep separate in our minds the measure, which is a number indicating a point on a proficiency scale, and the thing measured, which reflects cognitive attributes of the learner, as elicited by content attributes of the tasks. Of course, our focus in testing is on the learners, but the test tasks define completely what we can expect to discover about them.

So the term 'proficiency' is defined here in terms of a measure, and interpretations drawn on the basis of the measure. Proficiency thus defined does not exist until someone measures it. We shall use it consistently in this sense to distinguish it from terms identifying various kinds of ability or competence which may be used in defining the construct of what is tested. These describe posited properties of learners which exist independently of whether they are measured or not. I believe that insisting on this functional definition of proficiency, together with the socio-cognitive model, actually renders superfluous a considerable amount of unsatisfactory theorising and model-building concerning the relationship of 'competence' and 'performance'.

The argument for the validity of measures must eventually come back to our theoretical model of cognition and the interactions with test tasks that we predict we will observe, given the features designed into them. To the extent that test performance empirically confirms these predictions then our claim for the validity of the test is strengthened.

2.2.2.3 How task difficulty, ability and standards relate

As explained above, a scale co-locates three things: test tasks, learners, and standards, that is, points on the scale which indicate achievement of some criterion-referenced level. Taking the levels of the CEFR as examples of standards we can see how these three notions interact.

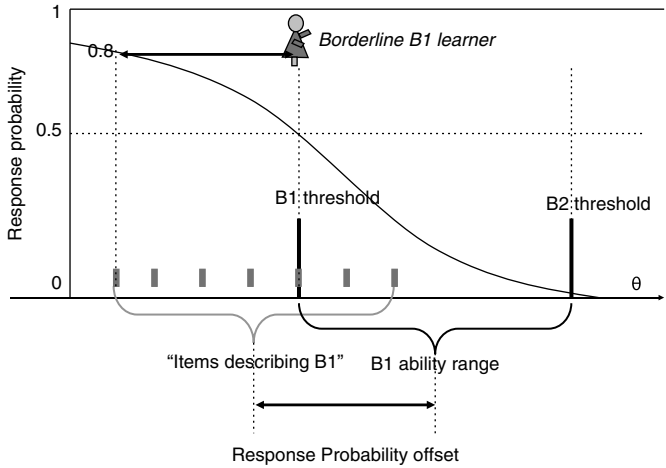
The tasks define what is tested. Grouping tasks by level allows us to characterise each level in terms of the sort of things learners can do. But learners define levels too: we understand levels not only in terms of *what* things learners can do but also *how well* they can do them. Another way of looking at Figure 2.4 above defines a point on a proficiency scale in terms of task difficulty *plus* performance level.

Performance level is more easily understood in relation to the performance skills of writing or speaking: a task such as *describing your holiday* does not relate strongly to a level. It sounds like an appropriate kind of task for a learner at CEFR A2 or B1, but every level of performance on the task is imaginable.

Performance on item-based tests such as of Reading and Listening is more simply evaluated – items are answered right or wrong. Here performance level must be understood in terms of the probability of getting an item at a certain level correct. The expected total score in a test is simply the sum of the probabilities across items, so that higher performance will relate to higher total scores.

Figure 2.5 illustrates two levels: CEFR B1 and B2. Each level has a lower threshold at which it is achieved, and a higher threshold where the next level takes over. A borderline B1 learner is shown. This learner has just achieved B1 level, and will continue to be at B1 until they achieve the next level up.

Figure 2.5 Items, learners and levels on a measurement scale



What do we expect of a B1 learner, even a borderline one, in terms of being able to do the kind of tasks which describe B1? We recognise that there are easier and harder B1 tasks, because B1 covers the whole range of difficulty between the end of A2 and the beginning of B2. We would expect the borderline learner to have mastered the easiest B1 tasks but not the harder ones. But thinking of mastery creates an apparent problem, because as illustrated in Figure 2.5 above, this learner has only a 50% chance of responding correctly to an item at the same point on the scale as they are, and their chance on more difficult items is even lower. This does not square with our idea of mastery – surely they should have a much higher chance (i.e. *response probability*) on the easiest B1 items? An 80% probability is a frequent rule-of-thumb definition of mastery, although this is of course an arbitrary choice.

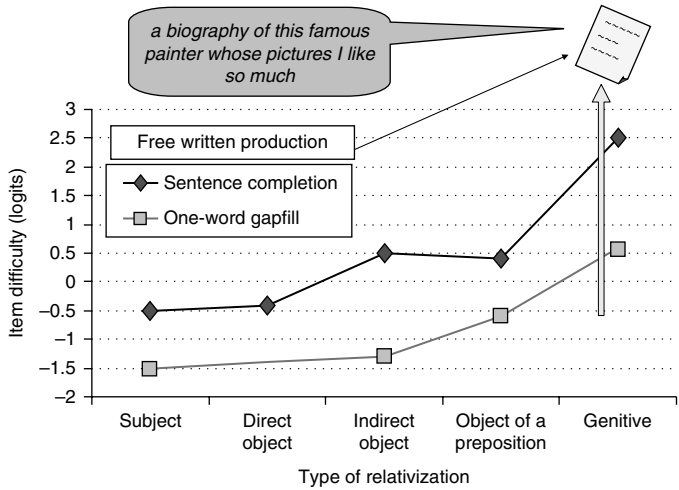
The conclusion is clear: the tasks which we take to describe B1 level reflect an expectation that learners at that level will be able to perform them reasonably well. This is true both of objective test tasks and performance-based tasks such as writing. An adequate performance level is built into our

understanding of the task. Thus the B1 level threshold defines the point at which learners can be said to be 'at B1 level', but it is confusing to talk of *items* as being 'at a level'. Better to speak of describing the level, or providing information about learners at the level. In terms of their location on the measurement scale items which we take to describe the level will be offset downwards from the level thresholds, as illustrated in Figure 2.5. The size of that offset reflects a judgement about the response probability which we choose to specify as indicating minimum mastery.

Another way of looking at this quite important concept is in terms of the notion of scaffolding. Scaffolding refers generally to the classroom setting, and describes the way a teacher sets up a communication exercise, preparing the students so that they can successfully achieve a learning task. To this extent, learning tasks are not like tasks in the real world. Similarly, test tasks are not quite like real-world tasks. In testing, the important thing is to enable learners to demonstrate their ability, and this requires a kind of scaffolding. Scaffolding is less evident with the performance skills, because a Writing test task is not dissimilar to a real-world writing task – for example: a holiday postcard, or an email to a friend describing an accident you had. But with the indirectly observable skills of listening and reading a test task is markedly different from a real-world task, however hard we may try to make it 'authentic'. Answering multiple-choice questions to show understanding of a listening passage is not something we do in the real world. The mechanisms by which we enable learners to demonstrate their ability in reading and listening can be seen as a kind of scaffolding, with the important purpose of ensuring that the test tasks have an appropriate facility for the targeted level of the learner. Imagine a Listening test in which B1 level is achieved with a score of 25% – that is, with the borderline candidate getting three-quarters of the items wrong. This might reflect that candidate's practical efficacy in some real world, but it would offer the learner a miserable experience and also would not work well from a validity or a measurement point of view. So sufficient scaffolding must be built into the construction of test tasks to offer an experience which is positive in educational terms, to the extent that the assessment relates to an educational context, and also provides good valid measurement.

An investigation of the components of task difficulty was the research focus of my PhD (Jones 1992), based on a bank of largely lexico-grammatical items. One example from this study will illustrate the role that item response format plays in scaffolding. Figure 2.6 represents Keenan and Comrie's (1977) hierarchy of difficulty for different types of relativisation. Keenan and Comrie posited that some kinds of relative clauses were inherently more difficult than others, the easiest involving sentence subjects (*I saw John, who had just got back*) and the hardest a genitive (*I saw John, whose car had just been stolen*). Data collected from the item bank of language-focused tasks (Jones 1992:191) confirmed the hierarchy reasonably well, while showing

Figure 2.6 Keenan and Comrie's Accessibility Hierarchy for two task types



that a one-word gap-fill task type is systematically easier than a sentence-completion task type. It revealed the difference in difficulty between writing one word or several words to complete a sentence.

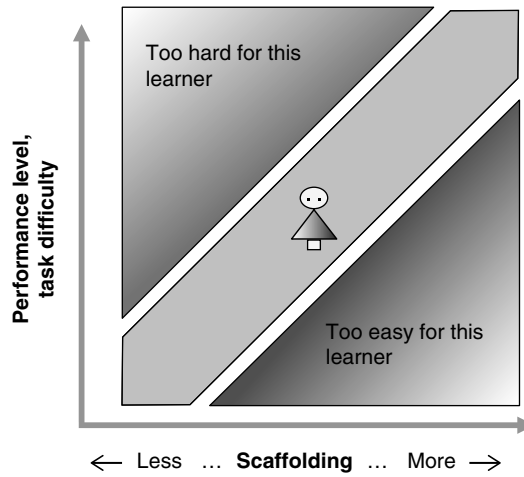
Keenan and Comrie's hierarchy is also confirmed by data from a corpus of writing analysed in the English Profile Programme (EPP) (Hawkins and Filipović 2012:33). The objectively marked task types can be seen to test different degrees of mastery of an area of English grammar, which could potentially be linked to, and predict, productive use in writing, as captured in English Profile corpus data, e.g. 'a biography of this famous painter whose pictures I like so much'.

Figure 2.7 illustrates the relation of scaffolding to performance level, underlining the relevance of the concept to both learning and assessment. The diagonal zone is the area within which learning can happen, with the learner able to perform on easier tasks with less scaffolding, and on more difficult tasks with more scaffolding. Outside this zone there is too much or too little challenge for learning to take place. In testing terms it is the zone where the learner can best demonstrate their ability. Figure 2.7 indicates that the same level of ability can be measured through different configurations of task difficulty and scaffolding, so that the test constructor can manipulate the degree of scaffolding in order to efficiently target an ability of interest.

2.2.2.4 Salience and information

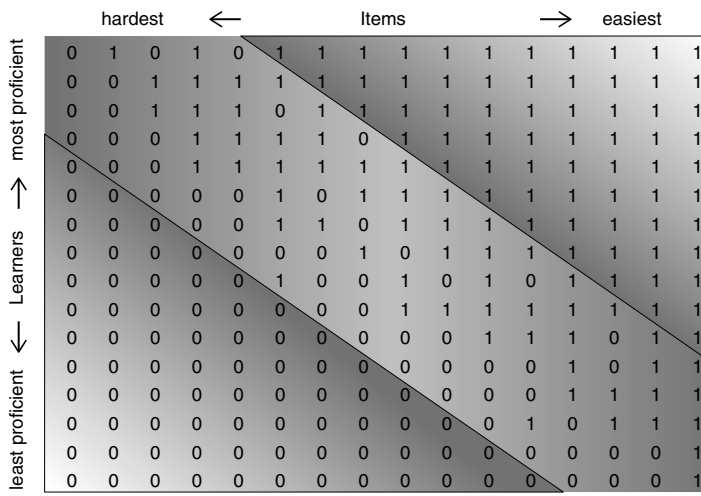
Salience is a term often used in describing the characteristic features of a proficiency level. Salience emerges from the patterns observed when learners respond to test tasks – it reflects that level of difficulty best matched to

Figure 2.7 Performance level and scaffolding



the ability of a learner, which thus produces the greatest score variance for that learner. Figure 2.8 illustrates this. It shows a matrix of responses to test items (1 = correct, 0 = incorrect), with the items sorted from hardest to easiest, and the learners sorted from most proficient to least proficient. The two triangles show areas of the matrix where all the responses are either right or wrong. These areas do nothing to discriminate between learners at each level. The diagonal band in which learners are getting some items right and

Figure 2.8 Salience and score variance



some wrong is what causes variance in scores and thus discriminates between learners at that level. Thus the hardest items discriminate between the most proficient learners, and the easiest items discriminate between the least proficient. Figure 2.8 clearly resembles Figure 2.7, which was used to illustrate the idea that differing degrees of scaffolding make tasks accessible to students at different levels.

Score variance creates *information* – tests which are well targeted at the level of the learner create good score variance and provide more information to discriminate between learners at that level. In contrast, responses within the two triangles in Figure 2.8 provide no information to discriminate within that level, even though they do separate high scorers generally from low scorers.

2.2.3 Scale construction for performance skills

2.2.3.1 Scaling performance skills; multi-faceted Rasch measurement

Writing in 2014, we can observe trial applications by Cambridge English Language Assessment of computer-based (CB) assessment of speaking and writing using artificial intelligence techniques. However, for high-stakes operational purposes Cambridge English Language Assessment still bases the marking of the performance skills on human judgement. It is a process which depends on the principle that markers, having been trained in the use of subjective rating scales, understand the standards, and therefore that the marks they award translate directly into measures of performance.

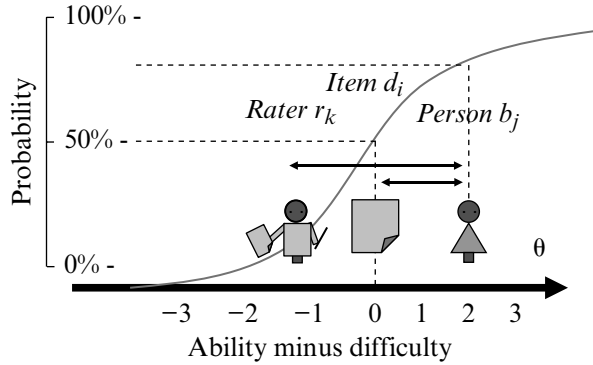
But even with careful training the accuracy of human marking cannot be taken for granted. Quality control procedures are necessary. The simple Rasch model illustrated in Figure 2.4 has been extended into a multi-faceted model which allows 'difficulty' to be decomposed into any number of elements, or facets (Linacre 2011). A common use of multi-faceted Rasch measurement is to include both raters and tasks as facets. Just as tasks can be easier or more difficult, raters can be more or less severe. Both facets impact on the learner's mark. Figure 2.9 below illustrates such a model.

In a multi-faceted model the design must be carefully done to ensure that all the facets can be separately estimated. If one rater marked half the candidates and a second rater marked the other half this would not be a workable design, because it would be impossible to distinguish the severity of the two raters from the overall level of the two groups of candidates. But if a third rater marked some candidates from both groups then the design would work. Then the learner's ability can be correctly estimated, with both rater severity and task difficulty being taken into account.

The Facets approach is attractive because it allows the possibility of correcting for rater leniency or severity without bringing the problem to the rater's attention. Such feedback has been observed to have unpredictable

Figure 2.9 A multi-faceted Rasch model including raters

$$p = \frac{\exp(b_j - d_i - r_k)}{1 + \exp(b_j - d_i - r_k)}$$



results, possibly leading to over-compensation on the part of the rater, or undermining the rater's confidence, leading to more erratic rating. There is the additional problem that providing feedback in the course of rating negates the idea of calculating a simple *post hoc* correction, given that the rater's behaviour can be expected to have changed mid-way. Facets can thus deal elegantly with stable differences in rater severity, but cannot of course correct for erratic rater behaviour. What it can do however is help identify such behaviour through analysis of fit statistics. Wolfe (2004) discusses the use of the multi-faceted Rasch model to identify a range of rater effects.

2.2.4 Item banking

Item banking is a methodology for constructing tests and interpreting test outcomes based on an IRT model. Its great value is that it creates an interpretive framework that encompasses exams at different levels, over different exam administrations and test versions, making it possible to generate tests with very similar measurement characteristics and to grade them to constant standards. In practice, exploiting an item-banking methodology also implies having technical systems in place to streamline the construction and storage of tasks and tests, the conduct of analysis, and updating the statistical information held on items. That is, doing item banking requires an item bank. Cambridge English Language Assessment uses an in-house development called LIBS (Local Item Banking System), which supports an end-to-end item production process, building in scheduling and quality control procedures. Figure 2.10 gives a schematic view of item banking as a methodology for test construction.

Figure 2.10 Item banking approach to scale construction and use (after Jones and Saville 2007)

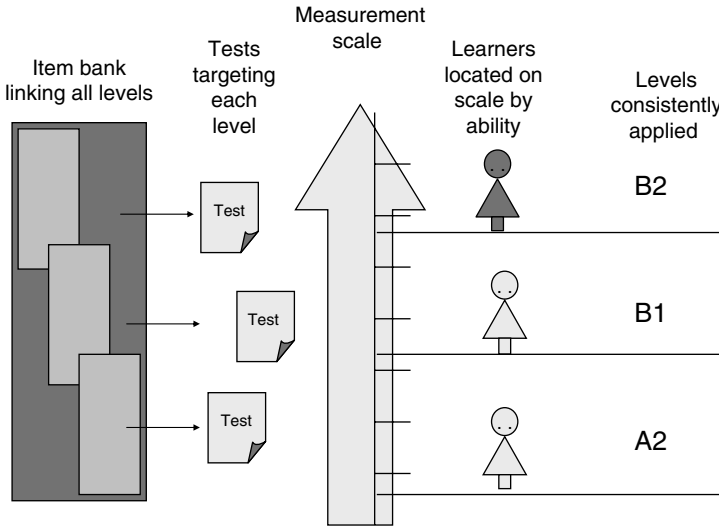


Figure 2.10 shows on the left an item bank containing tasks ready for use in a test. The difficulty of the items in each task is known, that is, they have been calibrated, using an IRT model to process test responses. The data needed to calibrate these tasks has come from some form of pretesting. Cambridge English Language Assessment practice is to pretest on volunteer groups of learners, who treat the pretest as familiarisation with the exam for which they are studying. It is vital that calibration is done in such a way that there is linking in the data to connect all the levels into a single scale. Operationally this is most often achieved by using *anchor tests* – short tests consisting of already calibrated items, which are administered to pretest candidates together with the pretests themselves. Another approach to pretesting is to seed new items into live online tests, so that each candidate may be exposed to a few items which do not actually contribute to their final results. Imaginative approaches to doing pretesting may be needed. Perceived problems, above all concerning risks to security, can in many contexts make it impossible to introduce an item-banking model.

With the item bank stocked, Figure 2.10 shows tests being assembled by selecting tasks in an appropriate range of difficulty for the target levels. Candidates' scores on tests locate them on the measurement scale according to their ability. Figure 2.10 shows tests at three levels, and three candidates. Although they might all have the same score – say, 70% – on the test which they took, 70% on the easiest test indicates a lower ability than 70% on the hardest test.

Finally, the standards are applied. These are fixed points on the scale reflecting our current best understanding of the standard. They can be directly applied to find the grade a candidate has achieved in the test. Even if different test versions at the same level may differ slightly in difficulty, the standard can be applied quite precisely. Over time, of course, the collection of more information may enable the standards to be modified to reflect our better understanding. The change will then impact all tests in the same way.

Figure 2.10 thus illustrates the power of a fully functional item-banking system. In such a system *ad hoc* subjective standard setting is neither necessary nor possible. The great benefit of an item-banking approach is not simply that it facilitates the construction of a stable measurement scale, but that in consequence it facilitates the construction of meanings which explain in various ways what it is that the scale measures. The items in the bank provide a concrete, detailed description of progression in terms of test content. Moreover, the fact that standards can be precisely maintained from exam session to exam session, and from level to level, means that it is much easier to do the research to develop stable interpretations of learners' performance in the world beyond the test. This volume offers several examples of how this can be done. The ALTE Can Do project described in Section 3.4.2 illustrates one way of linking test results to real-world competence, through self-report questionnaires. The CEFR descriptive scales were developed by a similar process involving teachers' judgements of student performances. Another procedure described in Section 3.2.3 involved supervisors in a work context judging the ability of office staff to deal with language-related tasks.

The above examples deal with descriptions of language proficiency level in functional terms. The EPP (http://www.englishprofile.org) is a large-scale study which has produced a description of CEFR levels in linguistic terms, identifying salient features of each level based on an extensive corpus of learner performance data (Hawkins and Filipović 2012). The data includes each learner's CEFR level as indicated by a Cambridge English language exam result.

2.3 Alignment across skills, languages and contexts

The fundamental framework-building activity is *alignment*: bringing different scales, perhaps independently developed, into a relationship with each other, enabling comparison between them. Comparability is valuable because it adds meaning and interpretability to the notion of language proficiency levels. Comparison, in fact, is at the heart of meaning: 'There is no absolute judgement. All judgements are comparisons of one thing with another' (Laming 2004). Our understanding of the world depends substantially on comparative judgements, based on quantitative or qualitative evidence.

The frameworks described in this volume might be seen as formalised instances of a universal heuristic based on classification and comparison as tools for organising experience. Given that our habitual use of this heuristic is largely unconscious, it is important that in formalising comparative frameworks we attempt to be explicit about the bases of comparison.

Frameworks extend the inferences that can be drawn from a test result in a specific context to something more context neutral – their usefulness lies precisely in enabling some comparison across different contexts. Jones and Saville (2007) conclude that interpretations of test performance in relation to a wider framework might therefore – probably should – be less specific than interpretations in relation to the original context for which the test was developed. However, the huge impact of the CEFR would lead us to reconsider this. Once we might have assumed that labelling performance as simply 'B1' or 'B2' necessarily under-represents a more complex construct, which targets specific aspects of the testing context: the learners' age, their purpose in studying the language and seeking accreditation, and so on. But over time more meaning has accrued around the CEFR levels themselves. Reference level descriptions, such as the English Profile (Hawkins and Filipović 2012) provide linguistic descriptions of progression. The ESLC (Chapter 6) is a major study which has provided authoritative comparative data to feed back into the evaluation of particular contexts. Most significantly, many examinations – older and newer – have integrated CEFR reference into various aspects of their construct definitions and test construction or administration procedures (Council of Europe 2011), so that the distinction between the specific context and the encompassing framework is blurred. So is a B1 exam with a business focus *primarily* a B1 exam, or primarily about business? I would now tend to say the former.

It would still seem reasonable that an interpretation of an assessment linked to a framework should combine specific features of the assessment context with general features of the framework. But how should this look in practice? In the case of the CEFR – explicitly a framework of *reference* – the task would seem to be to take *from* the framework what is relevant to the specific context. But at the same time, if the CEFR is to continue to develop and demonstrate its relevance to a range of contexts, then it would also seem necessary to contemplate continually re-interpreting or extending the descriptive apparatus of the CEFR itself to better reflect the contexts where it is put to use. We will come back to this in Section 4.5

2.3.1 Alignment across skills

In discussing alignment across skills let us focus on the case of a single language assessment consisting of several different parts (or 'papers'), each one testing a different skill. This seems relatively straightforward compared with

the situation of, for example the ESLC, with its requirement to align standards across both languages and skills; at the same time, it allows us to address a particular issue concerning how individual skills and composite 'exam levels' align to the CEFR.

The core Cambridge English exams test Reading and Writing (sometimes combined in one paper), Speaking and Listening, and at higher levels a Use of English paper tests grammar and usage. The candidate's performance on these different parts of the exam is combined to provide a single result. A graphical profile included on the candidate result slip reports performance in each skill in relative terms; but the exam certificate mentions only the global level.

In pursuing the goal of efficient measurement, assessments tend over time to adapt themselves to the profile of abilities displayed by candidates. They are always, to an extent, norm referenced. Of course, the basic premise of the multilingual framework is that it is criterion referenced: that is how it provides meaningful interpretations of exam performance. However, as Angoff (1974) has pointed out, any criterion-referenced assessment is underpinned by a set of norm-referenced assumptions. Tests are designed with a particular population in mind: a test at B2 is designed in the expectation that it will be taken by cohorts of candidates who seek accreditation at that level. Some will achieve the level, others fail. The quality of the test depends largely on how effectively it distinguishes these two groups. Thus it will probably seek to gather most information around the pass/fail point on the ability continuum (although it may also be important to measure well over a wider range, if the exam reports a range of grades).

With performance skills (writing and speaking) it is sufficient to set tasks at an appropriate level of challenge to elicit rateable performances. This can readily be done by reference to descriptions of expected performance at the level. With objectively marked tests such as Listening or Reading efficient measurement requires items to be pitched at an optimum level of difficulty. That level depends on the skills profile of the typical candidates: how good are they at reading, relative to listening?

If we accept the above analysis it is to be expected that Cambridge English exams should reflect a skills profile representing their traditional candidature. It is this norm which we have in mind if we state that groups of candidates from particular backgrounds, are, for example, 'bad at listening'. This implicit norm-referencing is relevant to understanding how exams such as those offered by Cambridge English interpret or define CEFR levels. In the original conception of the CEFR, based on functional definitions of language ability over a range of skills and contexts, there is no expectation that the typical learner would progress in all skills at the same rate. Asset Languages (Chapter 5) was designed explicitly as a modular qualification (that is, with each language skill separately tested and accredited) because

for many community languages it was expected that learners would show just such uneven development, having perhaps good oracy but weak literacy skills. But the logic of an assessment system which reports a single global result is that predictable differences in the difficulty of skills for the expected candidature should be minimised, in order to maximise the reliability of the global measure.

In practice it is the performance skills of writing and speaking which, being observable, determine most clearly whether a candidate has achieved a CEFR level or not. The objectively marked skills of reading and listening are less directly interpretable in terms of CEFR levels, being measured on continuous scales relating to indirectly observed mental processes, and with levels of difficulty approximately normed on a particular cohort. It is very difficult to say what, if any, impact these assessment issues have had on the meaning of the CEFR levels as intended by its authors. After all, the illustrative descriptor scales included in the CEFR reflect the same processes at work, with the Can Do statements concerning each skill normed on the (European) subjects of the calibration study (North 1996).

Following the above arguments, Cambridge English exams report a single CEFR-linked exam result, although CEFR advocates often insist on the importance of profiled reporting (for example, Beacco, Byram, Cavalli, Coste, Cuenat, Gouiller and Panthier (2010:9) state: 'In general, "levels" should be dropped in favour of "competence profiles", which provide a more accurate picture of learners' actual skills in their languages'). However, in practice students generally work towards a particular CEFR level, take an exam at that level and wish to have evidence of having achieved the level. Profiled reporting by skill may be useful additional information, but the assessment approach described above is coherent with the notion of measuring global achievement of a level, rather than skill by skill.

2.3.2 Alignment across test contexts and levels

The introduction to the Rasch model in Section 2.2.2 defined proficiency as something that does not exist until someone measures it. This distinguishes it from other terms such as 'ability' or 'competence' which describe properties of learners which exist whether they are measured or not. This use of the term underlines that what is measured in a proficiency test depends critically on the items in the test. But this at once raises questions about the limits on comparability across tests and contexts. Exams are frequently designed for specific groups, distinguished by age or professional interest, and this enhances their validity for that group, because the test items can be made wholly relevant and appropriate. But again, does this not call into question the possibility of comparison?

At first sight, it may seem quite unrealistic to seek to measure language

proficiency in terms of a single, universal progression. Is not every progression context specific? This is what Spolsky suggests:

> A functional set of goals exists in a social context . . . Where this is consistent and common as in the Foreign Service, or in the Council of Europe notion of the Threshold Level for tourists and occasional visitors, it is not unreasonable to develop a scale that proceeds through the skills . . . If it cannot be based on a single social goal, a single set of guidelines, a single scale could only be justified if there were evidence of an empirically provable necessary learning order, and we have clearly had difficulty in showing this to be so even for structural items (Spolsky 1986:154).

Spolsky suggests that communicative language ability may develop along a path strongly determined by the content of the curriculum – in this case, emerging from a needs analysis approach. Were this in fact the case, then the achievement element of the test would outweigh the proficiency element, and it would be difficult to compare this context with any other. Contemplating the construction of a measurement framework which will link different learning contexts across a range of levels, and yet enable useful comparisons between them, we are faced with an apparent paradox: the more we tailor a test to be relevant to a particular context, the less comparability there can be with other contexts.

But to go back to Spolsky's example: what if learners' progression in fact reflects more general cognitive processes which underlie language use, whatever the details of the functional curriculum, and enable the learner to satisfy a range of needs in a range of contexts? This is where the value of thinking in terms of *constructs* of proficiency asserts itself. We can look beyond the superficial aspects of test performance – that is, the performance of specific communicative acts – to what we believe lies beneath them. As presented in Section 2.1, construct definition is the process of building models of language ability, based on the best evidence available from psycholinguistics and related fields, and validated against test response data.

Section 2.1 introduced a socio-cognitive model which informs the development of test tasks that we can posit to measure underlying cognitive processes, while being adapted to be relevant to a specific context, through choice of topic, of lexis etc. In this way it is possible to develop a high-level conception of language ability which is applicable across a range of contexts. The proficiency frameworks which are the subject of this volume all measure proficiency in terms of an ability to use language in purposeful communication. This is what the ESLC set out to test (see Chapter 6), and for this reason it made no accommodation for students whose learning experience, notwithstanding any other outcomes it had, did not equip them to deal with the communicative tasks designed to serve the purpose of the survey. Other conceptions of language ability were excluded by the survey's

purpose – to measure communicative language ability as described by the CEFR. Similarly, the approaches to framework construction treated in this volume seek or assume a substantial degree of commonality across the range of relevant learning contexts. There may be other contexts which are simply out of scope. There are also limits to the comparisons we may make between the learner groups within a framework, but those limits are largely a matter of common sense: one does not often find a practical reason for comparing young children and university graduates, even though both groups might be represented in an inclusive framework such as the CEFR.

As noted above, curricula differ in details, and this can complicate comparison of groups following different curricula. This can be problematic, particularly at low proficiency levels, where the overlap between what two groups of learners have learned may be very small. The Asset Languages project provides an illustration of this problem (see Section 5.2.2), where no specific curriculum was provided, and teachers (particularly at low levels) found the tests difficult to prepare children for.

Learners' first language, their age, their profession, their purpose in studying the language, and the prominence of the studied language in learners' lives outside school are all factors which play a role in determining *what* is taught and learned in a specific context, and thus the nature of the link to a more general framework. Construct definitions can accommodate these differences, but naturally, some contexts will be more comparable than others. At the least, contexts should be comparable in terms of a general notion of level, as in the inclusive CEFR.

After all, there is something intuitively meaningful about the notion of being a beginner or an advanced learner, or at some level in between these extremes. The six standard levels of the CEFR seem to be distinct enough in the minds of most people working in language education that they can operate with them happily enough without the confirmatory evidence of tests, even if their understanding of the levels may differ in absolute terms. It does not strike us as paradoxical that the CEFR serves two purposes: firstly to provide a conceptual framework, describing all the ways in which contexts of learning differ, and secondly, to provide a levels framework which proposes that they can all be compared with each other. This seems reasonable because the CEFR is presented by its authors as a point of *reference*. But if we accept the idea of a set of levels serving as a valid reference point for widely differing contexts, this still leaves us with the question of *what* it is that learners considered to be at the same level actually have in common.

2.3.3 Language

Comparison across languages seems a simple notion to the extent that what is compared is communicative language ability. With the focus squarely on

the exchange of meanings it should be possible, one might think, to leave to one side language-specific features. However, there is evidently a strong link between mastery of a language system and ability to communicate through that language. We may try to reduce language use to its component aspects, as is done for example in Table 3 of the CEFR (Council of Europe 2001:28–29) entitled 'Common Reference Levels: qualitative aspects of spoken language use' (see Appendix A); however, there is clearly much interdependence between the five components of competence identified: Range, Accuracy, Fluency, Interaction and Coherence. Of these, Accuracy and Coherence seem to link more directly to lexico-grammatical features, while Fluency and Interaction seem to relate more to ease and effectiveness of communication.

However, the aspect Range seems to summarise the salient aspects of each level in a way which points up the interdependence: for example, Range at C1 describes a learner having 'a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics [. . .]'. One cannot imagine this being achieved through, say, inaccurate or incoherent language use: a high degree of accuracy and coherence is understood in the description.

Experience with CEFR Table 3 (e.g. at the Sèvres 2008 benchmarking event described in Section 2.3.4.3) confirms that raters can identify globally better or worse performances, but tend *not* to use the different aspects to identify marked profiles of performance (e.g. accurate but lacking fluency, or fluent but inaccurate). This demonstrates the well-known *halo effect* originally identified by Thorndike (1920). However, they characteristically mark performances relatively lower on the linguistic aspects and higher on the communicative aspects, that is, the former are marked negatively and the latter positively, but they do it rather uniformly for all performances. In this way raters apparently satisfy an inherent need to balance punishment of error with reward of communicative effort, thus coming at a conclusion from two directions.

Two projects discussed in this volume, Asset Languages and the ESLC, used a scheme based on the two aspects Language and Communication as an approach to the marking of writing (and speaking in the case of Asset Languages). A useful study, but one that has not yet been conducted, to my knowledge, would be to compare how these two aspects are treated across languages, given that languages clearly differ in the degree to which concepts to be communicated relate explicitly to lexico-grammatical features of the language. In English, for example, the sentence 'the cat sat on the mat' seems grammatically simple, but in Polish a fully grammatical rendering would require the speaker to identify the gender of 'cat', to select between the perfective and imperfective aspect of 'sat', and dependent on that selection, to use either the locative or accusative case of 'mat'. For each level of Asset

Languages, for example, the guidelines provided to item writers and teachers for these two languages, and the expectations of performance, would need to be quite different. Thus the nature of written or spoken production can be expected to differ across languages in terms of the salience of grammatical features and their impact on the effectiveness of communication. This may be expected to complicate comparison.

English doubtless presents specific problems of comparability. The presence of English in social life outside the language classroom, particularly its visibility in the media, means not only that it tends to be learned to a higher level, but also that learners – particularly young people – are likely to acquire a more natural and fluent style at an early stage. As the Sèvres 2008 event illustrated, this is readily interpreted by raters as evidence of a high level of competence, although what is actually communicated may well be fairly minimal.

One further, specific problem of comparability for the 25-language Asset Languages framework arose in the case of languages with a difficult, non-Latin script (see Section 5.2.2.3). Acquiring literacy skills in such languages is an extended learning process even for native speakers. This reveals a problem with the implementation of a CEFR-linked assessment system which can pass unnoticed in the context of European languages, where it does not really arise. In the European context it seems reasonable to treat all languages as comparable in terms of the learning effort required to reach a certain level. There appears to be no tension between the two functions of the CEFR levels – to provide a learning ladder of accessible targets, and a set of significant goals worthy of accreditation. But the learning effort required to achieve A1 in a literacy skill in a non-Latin-script language is considerably greater than in other languages, so that A1 as a first learning target may be judged too far away in terms of learning effort. Intermediate learning targets can of course be inserted, as recommended and illustrated in the text of the CEFR, but at the expense of abandoning the symmetry of the assessment framework across languages (which in the case of Asset Languages was not technically possible).

This is a significant issue, because for many users of assessment frameworks in an educational context comparability is chiefly conceived of in terms of learning effort. In practice, this is how expected levels of performance in dissimilar subjects – say, mathematics and geography – are determined in schools examinations. Attempts to introduce criterion-referencing against the CEFR into school examinations are likely to run into difficulties over this issue, as illustrated by the experience of Asset Languages (Chapter 5).

To conclude this discussion of language, and the extent to which proficiency in communicating is influenced by grammatical features of a given language, it is worth considering the role of the interlocutor in ensuring successful communication. In a discussion of the social conditions which

impact on communication, the CEFR identifies the number and familiarity of interlocutors, relative status, and social relationships between participants (e.g. friendliness/hostility, co-operativeness) (Council of Europe 2001:47). The CEFR's Can Do descriptors also make some reference to an implied interlocutor, e.g. in the A1 listening descriptor 'can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning' (Council of Europe 2001:66). Despite these references I suspect that the interlocutor, and cultural norms of behaviour in particular settings, may be an important missing parameter in attempts to define equivalent levels of functional language proficiency across countries and languages. Generally, there are contexts where one might expect more accommodation and patience on the part of interlocutors – shops and hotels, for example, and others where one might expect much less. To some extent this must also encompass issues of grammatical accuracy – to what extent do errors actually impede communication, or strain the tolerance of the interlocutor? While we must recognise that differences between language systems will impact on clarity of expression, probably to a different degree at different levels, the impact on the success of attempted communication cannot be straightforwardly predicted.

The current Cambridge English exam suite offers an example of a complex multi-level, multi-exam system. It covers every level of the CEFR, with one or more exams available at each level, for different contexts – that is, candidate groups and purposes. Understanding and controlling the relative difficulty of these exams, and the standards applied in grading them, is a crucially important task. Since the early 1990s this task has been supported by statistical scaling methods (see IRT in Section 2.2.2.2, and the historical account of the construction of the Cambridge ESOL Common Scale in Section 3.3). Alignment is thus both horizontal, between two exams at the same level, and vertical, between exams at different levels. Specific approaches to alignment, based on a variety of different kinds of data, are described in the following chapters. Here we will present in general terms two basic approaches to alignment in Cambridge English practice: quantitative, as used for objectively marked tests such as Reading and Listening, and qualitative, as used for performance skills of writing and speaking.

2.3.4 Practical alignment issues

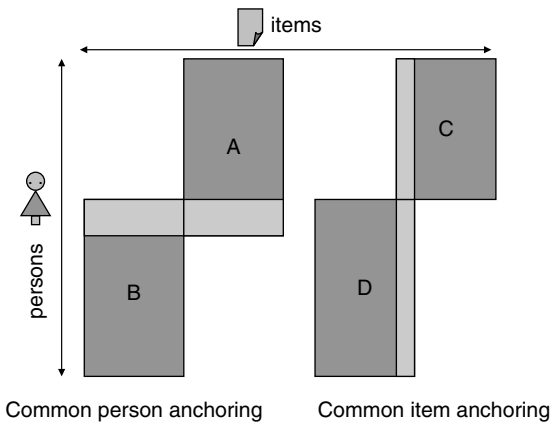
2.3.4.1 Alignment of objectively-marked tests

Sections 2.2.2 to 2.2.4 introduced IRT, and its implementation using an item-banking methodology. This technology provides an effective and practical way of linking examination systems and individual test versions into an encompassing framework, thus aligning them horizontally (linking two tests at the same level) and vertically (linking tests at different levels). A classic text

on test equating is Kolen and Brennan (2004). This section takes forward the presentation above to consider some specific practical issues in the vertical alignment of tests in a framework.

Where tests or exams are to be linked using IRT, some form of anchoring must be implemented. Generally two approaches are distinguished: *common person* anchoring, where a group of testees are recruited to respond to two different tests; and *common item* anchoring, where a group of items are inserted into two physical test forms, or administered concurrently with them. Either method produces a similar result: a set of responses which link two datasets and enable them to be calibrated to a common scale. Figure 2.11 illustrates. On the left, response data has been collected for two tests, A and B. A proportion of the testees have taken both tests. On the right, two tests C and D contain a proportion of identical items (or have been administered alongside a separate anchor test).

Figure 2.11 Two approaches to anchoring



Calibration can also be done in two different ways, using *separate* or *concurrent* estimation. In the first, the two datasets are analysed separately, producing two independent sets of difficulty/ability values for the common anchor elements. Alignment of the two datasets is done by taking the difference between the two sets of values – perhaps a simple average, or perhaps evaluating the individual data points, such that data judged as invalid might be excluded from the anchor. Typically, one dataset retains its original values, and the other has its difficulty/ability values adjusted by the estimated difference, thus aligning it to the first dataset. In concurrent estimation all the responses are included in a single analysis, providing a single best-fit calibration of all the items.

The concurrent approach is preferable if all the datasets (and there might

be many more than two) constitute the core of a new bank of material. Operationally this is most often not the case: item banks are progressively extended, session by session, and it is desirable that items already calibrated retain their values over time. In this case separate estimation of new material, anchored to existing material, is preferred. Operationally, also, common-item anchoring tends to be used far more than common person anchoring, because it is simpler to administer anchor tests at the time of pretesting than it is to find groups of learners willing to complete two whole tests. Common person anchoring is more characteristic of experimental research designs than operational procedures.

The IRT approach described above offers powerful ways of constructing scales and aligning tests within them, horizontally and vertically. However, it is not magical and is not guaranteed to provide sensible and interpretable results.

Firstly, such quantitative approaches to alignment cannot automatically reconcile qualitative differences between tests or groups of learners. Exams designed for and used with particular groups of learners are difficult to compare, to the extent that what they test may differ. Statistical anchoring may or may not reveal a problem. A study linking two Cambridge English exams –the Key English Test (KET) and Young Learners (Flyers) used a specially developed test consisting of material from the two exams, and collected data from learners preparing for either KET or Flyers (Flux 2001). Predictably, perhaps, the two learner groups responded to the test material differently, depending on which exam it was taken from. This example shows how qualitative differences can invalidate or complicate quantitative comparisons; but even in cases where an anchor appears to work well, caution is needed in interpreting a statistical alignment between very different exams or learner groups.

Secondly, the IRT models which we use are just that – models – which make theoretical assumptions that are in reality always violated, to an extent. Actually, the technical term 'assumption' is misleading as it suggests a condition that is somehow taken for granted. It would be better called 'requirement', that is, a condition that is acknowledged but may not be fully achievable. An explicitly identified requirement of many IRT models is that of unidimensionality: all items measure the same trait. But as Andrich (1988:9) explains, 'unidimensionality is a relative matter – every human performance, action, or belief is complex and involves a multitude of component abilities, interests and so on. Nevertheless, there are circumstances in which it is considered useful to think of concepts in unidimensional terms'.

Some assumptions can be addressed by choice of model (there are multidimensional IRT models, for example). Some assumptions can be sidestepped: the assumption of *local item independence* states that responses to any one item should not depend on responses to any other items; it is frequently violated in the case of text-based tasks, or matching tasks where a correct match

increases the chance of further matches also being correct. But a solution is to treat the task using a partial credit model, so that instead of single item scores there is one summed score for the whole task.

More troublesome in practice is the fact that all IRT models have practical limitations on the response data they can sensibly deal with, and there are no absolute rules for determining where those limitations begin. As Figure 2.4 illustrated, the curve that links test scores to abilities reflects the fact that as scores in a test approach extreme values (0 or 100%) a learner's estimated ability approaches minus or plus infinity. Of course, extreme scores are not strictly interpretable, because we cannot know how much higher or lower the learner's ability might be. But according to the logistic transformation of scores which IRT models use, the learner's ability at an extreme score is not so much unknowable as infinitely biased. And if a score of 100% has infinite positive bias, then a score of 99% must be strongly biased in the same direction. But how extreme can a score be before we reject it as a measure of ability? Extremely high and low scores (item facilities) are thus problematic. Particularly in the case of vertical scaling, where linking across levels implies covering a wide range of abilities, it is important to keep item facilities within as narrow a range as possible – say, between about 20 and 80% facility.

It is also worth checking the discrimination of items used as anchors. Discrimination concerns the steepness of the curve: a highly discriminating item has a steep curve, showing that it distinguishes strongly between learners at different abilities along the horizontal axis; a poorly discriminating item has a flatter curve. Some IRT models include a discrimination parameter in modelling how an item functions. The Rasch model used in the scaling projects reported here does not explicitly model discrimination (the curve for all items has the same slope), but the analysis software reports it.

When using the Rasch model a set of anchor items for use in vertical scaling should be selected to be of *average* discrimination. Generally, items which discriminate highly are considered to be better items; but higher-than-average discrimination in a set of anchor items will cause them to separate two datasets more than they should; lower-than-average discrimination will have the opposite effect.

2.3.4.2 Alignment of subjectively marked tests

Alongside the construction of the Common Scale for the objectively marked skills of reading and listening, using the quantitative methodology of IRT, considerable effort at Cambridge English Language Assessment has gone into developing common scales for writing and speaking.

Galaczi, ffrench, Hubbard and Green (2011) offer a review of rating scale construction methodologies, referring to the distinction made between intuitively and empirically developed rating scales. Intuitive methods primarily rely on expert judgement and the 'principled interpretation of experience'

(Council of Europe 2001:208). Empirical methods, by contrast, are data-driven and based on actual learner performances. Empirical scale development methods may be subdivided into 'quantitative' and 'qualitative', based on the type of data they draw on.

The tests in the Cambridge English suite have developed at different times to meet specific needs, and accordingly, the assessment scales for speaking and writing at each level were developed independently of each other. Although scales were based on the same criteria and approach to assessment they did not link up to form a common scale, and could not fully reflect the progression in levels found in the CEFR.

Concerning speaking, Galaczi et al (2011) report on a project initiated to review and revise the core Cambridge English and Business English Certificate (BEC) assessment scales for Speaking through a series of interrelated development activities. There were requests from examiners for closer reference to the CEFR. It was even proposed that CEFR descriptors might be adopted for the Cambridge English tests with minimal modification. However, due to issues such as vagueness in the criteria for judging performance and inconsistencies in wording of the CEFR descriptors (see Section 4.4), this suggestion was rejected.

New draft descriptors were submitted by experts, then refined in an iterative process. An important presentational change was to integrate bands across levels into a continuous, overlapping scale, so that, for example, the descriptors at A2 Band 5 would correspond to the descriptors at B1 Band 3 and B2 Band 1. This 'stacking up' on a common scale would better reflect the continuum of language proficiency in the CEFR and the intended overlap between scoring bands in the tests at different levels.

Significant changes were made in the formulation of descriptors, which as Galaczi et al (2011) report, were well received by examiners: descriptors at all levels were positively oriented, and vague terms such as 'adequate', 'mainly', 'may', 'might', 'usually' were eliminated. The new assessment scales were thoroughly validated using a range of empirical approaches.

Concerning writing, an early study initiated by Nick Saville and reported by Hawkey and Barker (2004) undertook a qualitative analysis of a corpus of scripts elicited from candidates covering three Cambridge English examination levels, but all in response to the same writing task. The intention was to use both intuitive and computer-assisted (corpus linguistic) approaches to identify key language features distinguishing performance in writing at four pre-assessed proficiency levels, according to the perceptions of expert markers, and to suggest how these features might be incorporated in a common scale for writing. The study was also informed by the criteria for writing used in a number of existing band scales, including those of the CEFR.

The scale was intended to help test users interpret levels of performance

across exams and locate the level of one examination in relation to another: in Alderson's (1991) terms, a 'user-oriented' scale.

The approach chosen, of using a single writing task, led to descriptions of levels in relative terms which for the lower levels were inevitably negative. A table in Hawkey and Barker (2004) describes the lowest level thus:

- very limited command of the written language
- produces short texts but does not operate beyond paragraph level
- has a limited knowledge of basic structures and vocabulary but is unlikely to produce these accurately
- may be able to produce a simple message, though elementary errors will sometimes impede communication.

These are effectively 'can't do' rather than Can Do statements. Although such statements might be useful in some purely summative assessment contexts, statements used in some learning context would do well to heed North's recommendation (Council of Europe 2001:Appendix A) that effective descriptors should be, among other things, positively formulated and definite.

Shaw and Weir (2007) is the book-length study of the construct of writing as assessed at each Cambridge English exam level, as described in the levels of the CEFR, and as understood by researchers in psycholinguistics. This triangulated approach has proved productive as a way of understanding the nature of progression through levels, and imposing a common and coherent approach to marking across the different exams.

Lim (2012) reports on a major revision of the writing mark schemes for the core Cambridge English and BEC examinations. In an extended two-year study a new mark scheme was developed and validated, drawing upon the recent experience of revising the assessment scales for speaking for the same group of exams (Galaczi et al 2011).

2.3.4.3 Alignment across languages: multilingual comparative approaches

We have argued that in a multilingual framework cross-language alignment should be treated as a step distinct from and prior to standard setting (Jones 2009b). It seems logical first to align tests in different languages to the same scale, and only then to develop interpretations – i.e. set standards. Those interpretations will then apply equally to all the aligned languages. By analogy with the development of thermometers as instruments for measuring temperature: first, invent the instrument, next, agree on a single scale to use (nowadays most agree to use Celsius) and lastly, begin using it to develop a common understanding of temperature, for example in relation to medicine. Common understanding develops from the common scale.

Alignment across languages requires comparative judgements, or rankings, of one scale against another, where those judgements might focus on

tasks in language tests, or exemplars of performance in speaking or writing. Making comparative judgements is something that human beings appear to be significantly better at than making absolute judgements.

A comparative approach cannot remove the need for standard setting at some stage, but by deferring it until after the alignment of languages to a common scale it substantially reduces the scope of standard setting and gives it a specific focus. The standard is set once but applies equally to all aligned languages. Subsequent languages can be aligned to the same framework by a further, relatively simple, comparative exercise. There is no need – in fact it is not possible – to do standard setting separately for each such language, because the act of alignment applies the standard already set.

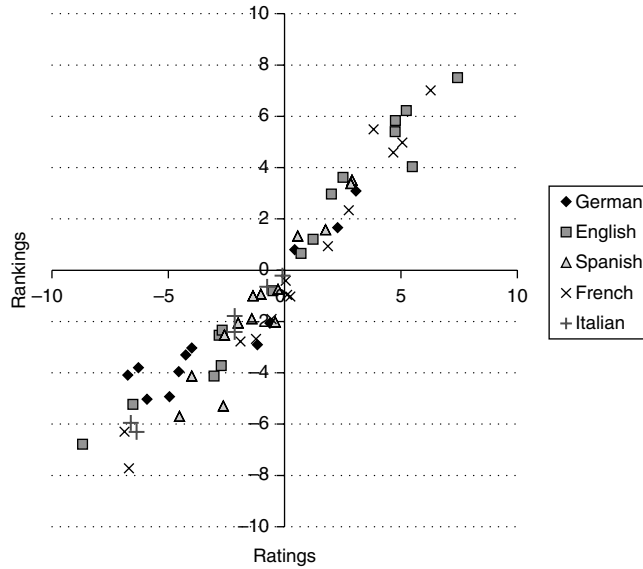
Jones (2009b) reviews comparative approaches (Bramley 2005, Thurstone 1927). Linacre (2006) reviews different methods of analysing rank-ordered data.

An encouraging case study of using comparative approaches in this way is the multilingual benchmarking conference organised by le Centre international d'études pédagogiques (CIEP) at Sèvres in June 2008. This focused on the performance skill of speaking (Breton 2008). Two kinds of data were collected. At the conference itself judges rated video performances against the CEFR, in multilingual sessions. Prior to the conference ranking data was collected from the same judges, using a specially developed web-based platform which allowed them to view samples and record their ranking by dragging samples to re-order them in a list. The allocation of samples for the ranking exercise was such as to ensure that each judge rated in two languages with which they were familiar, and that there was linkage in the data across all samples and languages.

Figure 2.12 shows the high correlation between the two approaches – rating and ranking. In this case the ratings constituted the standard setting, while the ranking exercise was an exploratory study which succeeded in demonstrating very good agreement with the ratings in terms of ordering the performances. It thus showed the practicality of ranking performances across languages, as a procedure which could henceforth be undertaken prior to a unified, multilingual standard-setting event.

The cross-language ranking approach has several advantages. It uses a direct comparison of performances, in contrast to an indirect comparison via the performance descriptors of the CEFR. It does not require judges to have a shared understanding of the CEFR levels in absolute terms, although it is of course essential that judges should share an understanding of what it is they are comparing, i.e. the construct of communicative language ability at the heart of the CEFR. Thus reference to the CEFR is useful, indeed, indispensable. Most importantly, ranking allows new languages to be added to an existing framework already aligned to CEFR standards, without the need for a further standard-setting event.

Figure 2.12 Ranking and rating compared for speaking (Breton 2008)



A similar alignment study for writing was undertaken for the ESLC, the results of which are presented in Section 6.7.4.

Thus, standard setting must include the verification of standards across languages. In this context it is worth noting that to date applications of the pilot Manual (Council of Europe 2003) have largely involved single languages. The ESLC provides a significant opportunity to make progress in establishing comparability of standards within Europe. The approach taken to aligning standards across languages in the ESLC is presented in Section 6.7.

While a comparative approach has been shown to work well for the performance skills of speaking and writing, it is more difficult to apply it to task-centred judgements on reading or listening, as measured by objectively marked items. This is taken up in the following discussion of standard setting.

2.4 Standard setting

Standard setting is the judgemental though evidence-based process whereby learners are assigned to levels. For example, the standard-setting event for the ESLC (see Chapter 6) determined the proportion of candidates in each country deemed to have achieved Pre-A1, A1, A2, B1 or B2 level. In the IRT model used in this volume, standard setting entails setting cut-offs (thresholds) on a measurement scale: once set, the same standards can be applied from exam session to exam session, until they are found to need revision.

Standard setting within a multilingual setting should ideally take place subsequent to the step of alignment described above, because where this is possible a single set of standards can be set which apply across languages. But this implies that standard setting itself must take place in an explicitly multilingual context.

Standard setting is a topic within assessment which has developed rapidly in recent years (for an overview see Cizek and Bunch 2007, also Council of Europe 2009). Figueras and Noijons (Eds) (2009) and Martyniuk (Ed) (2010) provide case studies and research perspectives specifically on linking language exams to the CEFR. Standard setting has spawned a large number of methods, which can be seen as variants on a few basic approaches. These can be grouped in terms of where they focus judgement. *Task-centred* approaches focus on features of test tasks, while *learner-centred* approaches focus on the performance of learners in a test; task-based approaches being the more commonly used (Cizek 2001, Cizek and Bunch 2007).

Approaches are sufficiently standardised to constitute a body of orthodox practice. This is problematic if it is taken to mean that a standard-setting outcome might claim to be valid simply because it followed officially sanctioned procedures. The idea that standard setting is essentially an exercise in showing due diligence is encouraged by certain premises or assumptions which, although applicable to standard setting in its classical context, seem questionable in the case of setting standards within a multilingual framework.

This argument is developed in Jones (2009a, 2009b) in the context of setting standards in relation to the CEFR levels, and is summarised below.

A classical standard-setting context is the professional licensure exam – for example, a 100-item multiple-choice question (MCQ) test for nurses. In this context the following premises hold:

1. The judges and candidates are members, or prospective members, of a specific professional community.
2. The test tasks relate to discrete items of professional knowledge.
3. The judges are qualified to say which items a practitioner should master.
4. Hence the notion of 'minimal competence' has substantive meaning.
5. The judges must balance the interests of the candidate nurses and the public whom they will serve. Judgements are not 'correct', only defensible in terms of the interested parties.
6. The frame of reference is the profession and its stakeholders, and no judgements have implications outside this frame.
7. The judges' professional and cultural background (for their practice is culturally embedded) impacts on their decisions and actually reinforces their validity (within that culture).

The CEFR context clearly differs in several respects. Listening and reading are skills: tests do not simply measure discrete atoms of knowledge, but attempt to tap hidden mental processes. Listening and reading are indirectly observable continua of ability: the notion of minimal competence, or any discrete level of competence, is hard to pin down. The frame of reference is languages across Europe, and so all judgements have implications which extend beyond the immediate context of a particular test or language. Judgements can and must aspire to be 'correct' in the sense of *consistent* with other judgements being made within the wider frame of reference. Therefore the culturally determined nature of judgements, far from reinforcing their validity, becomes a threat to it. This last point in particular presents the major challenge for aligning standards across languages. The intention of the CEFR is to provide a practical point of reference that enables a common understanding of levels. But level descriptors are not wholly concrete or definitive. They require interpretation, and we must expect that interpretations across countries and languages will reflect existing cultural expectations and may differ.

The reality of this problem was recognised when exam boards began to provide exemplars of test material as illustrations of CEFR levels on the Council of Europe website (Council of Europe 2012). It appeared to many who attempted to use these that the standards were not wholly comparable. The multilingual benchmarking event held by CIEP in Paris in June 2008 (Figure 2.12 above) represented a serious attempt to address the issue for the case of speaking. The ESLC (Chapter 6) also provided strong evidence that understanding of levels varies by context, being normed on local expectations of achievement in language learning (Section 6.7.5).

In the context of the CEFR a widely referenced resource is *Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR): A Manual* (Council of Europe 2009), hereinafter referred to just as the Manual. A preliminary pilot version was made available in 2003 (Council of Europe 2003), and the proceedings of a conference on case studies of using the pilot version are reported in Martyniuk (Ed) (2010). The treatment of standard setting in the Manual is thorough. Standard setting of written or spoken performance is referred to as *benchmarking* in Chapter 5 of the Manual; Chapter 6 is titled Standard Setting, and includes examples of both task-centred and learner-centred methods. Chapter 7, titled Validation, discusses various kinds of evidence that might help to confirm or disconfirm a standard-setting outcome. The disadvantage of this organisation is that it suggests a priority between those activities described as standard setting and those described as validation, as if the former were essential and the latter were additional options. A chapter in the volume on case studies (Jones, Ashton and Walker 2010) evaluates the Manual in relation to the standard-setting approaches adopted for Asset Languages (see

Chapter 5), which did not respect the priority suggested by the Manual, but rather accommodated the particular constraints and requirements of the Asset Languages development.

In fact, a common theme of the Cambridge English Language Assessment approach to standard setting in relation to the CEFR is that it requires to be integrated into the whole cyclical process of test design, construction, administration and evaluation, rather than, as the Manual seems to suggest, constitute a one-off event which can demonstrate a once-and-for-all 'linking' to the framework. Milanovic (2009:4–5) states:

> The recommendations found in the Manual on how to use the CEFR and other resources supplied by the Council of Europe for alignment purposes (e.g. familiarisation activities with stakeholders and standard setting exercises of different types whether task-based or person-based), need to be integrated within the standard procedures of the assessment provider and should not be seen as 'one-off events'. This is particularly true for an examination board like Cambridge ESOL which works with (literally) thousands of stakeholders in developing, administering, marking and validating many different types of examination within a consistent but evolving frame of reference. For example, in 2010 over 400 administrations of different Cambridge examinations will take place, all of which include the assessment of four skills (including face-to-face speaking tests). Given the complexity of this operation, the arguments for alignment to external reference points need to be developed on a case-by-case basis and must be one part of the broader validity argument which is needed to support the appropriate uses of each examination.

A second manual, a *Manual for Language Test Development and Examining for Use with the CEFR* has been produced by ALTE on behalf of the Language Policy Division (Council of Europe 2011). It presents an operational model which among other things locates issues relevant to standard setting at various points in the development and administration cycle. The different projects described in this volume provide illustrations of such an approach in action.

It is significant that the prescriptions of the Manual (Council of Europe 2009), and the uses most frequently made of them, relate to the case of setting standards for a single language. This fact alone limits the Manual's relevance to the multilingual frameworks treated in this volume. The CEFR remains of course the single target of standard setting: essentially, all languages are made comparable via the points of reference provided by the scales of the CEFR. However, as pointed out above, the levels of the CEFR may be understood differently in different countries, and an explicit cross-language comparison is more likely to avoid such culturally determined bias.

Standard setting is treated at length in the case studies which make up the

second half of this volume. They show standard setting approached in different ways, located at different stages in the test development and administration cycle. Together they depict standard setting as a complex and difficult area, and indicate that current orthodox approaches are not wholly satisfactory, particularly for the case of a multilingual framework. The pursuit of new techniques specifically targeted at this case should be considered a research priority.

The studies presented in this volume support the following general conclusions:

- performance skills are a more practical target for standard-setting judgement than indirectly observable, objectively marked skills
- comparative judgements are easier than absolute judgements, and therefore ranking may offer more than rating
- in a multilingual framework it is essential to minimise the role of subjective judgement.

These conclusions are further developed in the discussion of validity and validation in the next section.

2.5 Validity and validation

This section attempts to define in general terms the meaning of 'validity', and the possible scope of validation activities, in the context of a multilingual framework and of assessments which lay claim to a location within it.

So far we have considered validity in terms of Weir's (2005a) framework (see Section 2.1). However, in order to link back to fundamental issues in the discussion of validity it will be useful to refer to the model of validity presented in the 1999 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association and National Council on Measurement in Education 1999:9), which is essentially that of Messick (1989). This defines validity as the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests; in Messick's words:

> Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other modes of assessment (Messick 1989:20).

It is what Newton (2012) calls the consensus definition of validity, referring to its broad acceptance by the educational and psychological measurement and assessment (EPMA) community. Newton identifies four elements of the consensus definition, although he proceeds to question or seek clarification of each of these:

> First, it is bad practice to talk about validity as though it were a property of a test. Second, it is good practice to describe validity as though it were a property of an interpretation. Third, it is good practice to describe validity as a unitary concept. Fourth, it is good practice to define construct validity as the unifying essence of all validity (Newton 2012:5).

Messick (1989) presents four aspects of validity as a linked series of steps – a progressive matrix – relating to the interpretation and use of tests. *Test interpretation* and *test use* have both an *evidential* and a *consequential* basis. In terms of evidence, test interpretation appeals to *construct validity*. Test use appeals to construct validity but also adds considerations of *relevance* and *utility*. In terms of consequences, test interpretation justifies these through construct validity, but additionally the *values* which the construct implies. Test use takes both of these into account while adding consideration of *social consequences*.

Thus the central, most basic concept is that of *construct validity* – that is, evidence that the test measures what it purports to measure. Newton offers the following useful gloss on construct validity:

> Finally, at the heart of construct validity theory, although not always recognised as such, has remained a distinction between how we think and talk about the world and how the world 'really is.' This is to draw a distinction between the phenomena of EPMA (e.g., patterns of human behavior and interaction) and how EPMA professionals think and talk about those phenomena (e.g., in terms of competences, proficiencies, abilities, aptitudes, disorders, etc.). Or, in other words, it is to distinguish attributes (the evidential level: how the world is) from constructs (the theoretical level: how we represent the world) (Newton 2012:5).

Messick's progressive matrix has been critiqued and interpreted in different ways, particularly regarding how test use and social consequences relate to validity. In a narrow view, only uses justified by the meaning of test scores are relevant to judging the validity of a test. Messick, certainly, is concerned with score meaning. But as Bennett (2012:31) points out, the *Standards* cited above include consequences related to wider claims about how the test can be used, e.g.:

> Validity Standard 1.23: When a test use or score interpretation is recommended on the grounds that testing or the testing program *per se* will result in some indirect benefit in addition to the utility of the information from the test scores themselves, the rationale for anticipating the indirect benefit should be made explicit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided (American Educational Research Association et al 1999:23).

Thus, for example, if by linking a test to a framework we claim not only that this makes the scores more meaningful, but also that this in turn will improve student motivation and lead to more positive learning outcomes, this becomes a validity claim which it should be possible to substantiate in some way.

An even wider treatment of test use treats all good or bad consequences, intended or unintended, within or beyond the power of the test developer to control, as relevant to judging validity. Newton (2012:6) warns that extending the definition of validity too far into other fields such as programme evaluation or social policy analysis, means that in justifying test use:

> We would not simply be making a measurement claim, but also a legal claim, an economic claim, many ethical claims, and so on. The claim to validity would involve a judgement on the overall legitimacy or defensibility of the procedure. That is definitely a judgement for someone to make. But it is a judgement that few EPMA professionals would ever be likely to be in a position to make.

2.5.1 Making valid use of a framework

As the above presentation indicates, it is how tests are used which is the focus of a validity claim. How would we extend this to the use of a multilingual framework?

It is possible to state conditions which a framework would have to satisfy for it to be capable of valid use. Assuming that each assessment in the framework is linked to a particular context of learning, then, construct validity requires that:

- different contexts of learning within the framework should have enough in common to make comparison practical and worthwhile
- progression across levels of the framework should follow reasonably similar paths
- measurement scales should be reasonably comparable, so that scores can be interpreted in the same way.

Additionally it is possible to state enhanced benefits to expect in terms of interpretation and use:

- linking an assessment to the framework should provide increased evidence to support interpretation and use
- linking an assessment to the framework should enhance its value and promote positive social consequences.

Thus we can propose that linking an assessment into a framework should enhance its validity (that is, its fitness for purpose, utility), to the extent

that the construct represented by the test is sufficiently consistent with the construct represented by the framework.

These conditions and benefits can all be imagined in the abstract: a framework can be defined without there being any assessments ready to put in it. This describes to an extent the origins of the CEFR: it was conceived to serve as a point of reference which could be relevant to – i.e. valid – for a wide range of contexts. Its underspecified and extensible nature serves this purpose.

Furthermore, we can explain a validity claim for a framework in terms of the interpretation of score meaning – that is, within a strict interpretation of validity, in Messick's (1989) terms. If meaning inheres in comparability – comparability with a performance criterion, with other learner groups, with learners of different languages – then the value-added of linking an exam to a framework can be demonstrated in terms of greater comparability, that is, enhanced score meaning.

In practice, the extent to which we can guarantee the valid use of a framework doubtless depends on the amount of control we have over the assessments which are incorporated within it. The testing frameworks of Asset Languages, or the ESLC, were tightly controlled within dedicated developments. With the CEFR on the other hand we might assent to the validity of the overall model, pointing in particular to its values – a positive focus on proficiency, a common language for educationists – and the positive social consequences it aims at – better educational outcomes, more focus on plurilingual competence, etc. However, our attention would naturally focus more on the wide range of more or less valid and invalid uses made of it, for example in the way particular exams lay claim to an alignment, or in the kind of comparisons actually made by users. This is where practical validation efforts will be directed.

2.5.2 The validity of linking an assessment to a framework

Validity as invoked in this section may seem like an esoteric concept, but working with it is at heart a practical issue. As Kane (2012:66) states, on the question of how to define and how then to demonstrate validity: 'Of course, one needs at least a rough and ready answer to the first question before tackling the second, but I would argue that one needs some experience with the second question before tackling the first.' This captures nicely the practical and experiential dimension to validity, and it agrees well with the critical realist view of the assessment enterprise which we appealed to in the introduction. In the introduction to this volume frameworks were presented as heuristic constructs which seek to provide adequate models for describing and predicting language behaviour. Cronbach (1971:464) asserts, similarly, that 'a construct is an intellectual device by means of which one construes events. It is a means of organising experience into categories.'

Validity for a given assessment will have two aspects: the validity of the assessment in relation to its specific testing context, and the validity of further interpretations or uses based on its alignment to a framework. It may be, and increasingly is, the case that an assessment is developed from the outset with the intention of claiming linkage to a framework such as the CEFR; in which case these two aspects will be combined in the design and development stages, and validation activities may be seen as an integral aspect of test construction which should be built into each stage of the process.

2.5.3 Validation: The approach of the Manual

However, where a decision is made to relate an existing test to a framework *post-hoc*, then additional, specific validation work is called for. This is the situation which is presupposed in the treatment of validation in the Manual (Council of Europe 2009). Although it specifically focuses on linking to the CEFR, it is convenient to locate the Manual's contribution to the discussion of validity and validation in this general methodological chapter. The Manual presents the process as follows:

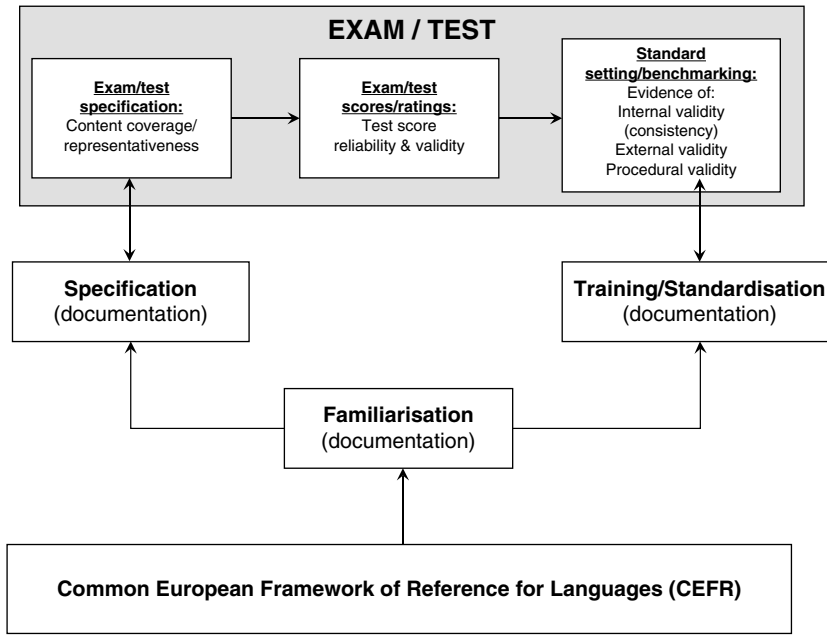
Relating an examination or test to the CEFR can best be seen as a process of 'building an argument' based on a theoretical rationale. The central concept within this process is 'validity'. The Manual presents five interrelated sets of procedures that users are advised to follow in order to design a linking scheme in terms of self-contained, manageable activities:

- Familiarisation
- Specification
- Standardisation training/benchmarking
- Standard setting
- Validation (Council of Europe 2009:9).

The Manual thus envisages linking as a process which includes consideration both of the content of tests (specification) and their level (standard setting), and which must be quality assured through appropriate training (familiarisation, standardisation). These are summarised in the schema reproduced as Figure 2.13 below.

Chapter 7 of the Manual is devoted to validation. It discusses validation directly linked to a standard-setting event under the heading *Internal validity of the standard setting*, and other, confirmatory approaches as *External validation*. Procedures connected with the first of these include checks on the accuracy and consistency of judgements, including intra-judge and inter-judge consistency, and ways of evaluating indices of agreement, or standard errors of cut-scores. It is pointed out that the indices found for a single event cannot be interpreted as indices of its absolute accuracy, given that the outcome depends critically on procedures carried out by

Figure 2.13 Validity evidence of linkage of examination/test results to the CEFR (Council of Europe 2009:8)



the same person or group of persons and on test data usually collected on a single occasion on a single group of students and using a single test or examination.

This leads into the presentation of external validation, which aims at providing evidence from independent sources to corroborate the results and conclusions of the original procedures. Not all evidence provided, however, is independent from the information used in the standard setting to the same degree, and not all evidence is necessarily equally convincing. Claims of generality are quite difficult to support. *Cross validation* is introduced as a procedure for verifying cut-off scores found from a standard-setting procedure by applying them to an independent sample. Comparison of indices of quality on the original sample and the cross validation sample give an indication of the generalisability of the results. External validation may also involve comparison of the results of the decision rule underlying the standard-setting procedure with the results of another decision rule (for example, teacher ratings of students). As the Manual points out, a statistical test may detect systematic differences in how these two rules allocate students to CEFR levels, but will not explain why these differences are there, or which rule is more to be believed.

Further approaches described under external validation include taking advantage of IRT calibration of items, and using 'Can Do' statements in various ways. Brief consideration is given to explicitly cross-language standard-setting approaches requiring the participation of plurilingual panel members. The cross-language benchmarking seminar held at Sèvres in June 2008 (see Section 2.3.4.3) is mentioned.

The Manual (Council of Europe 2009:117) is modest in its own appraisal of the methodology presented in the validation chapter, conceding that 'it does not make a clear distinction between good and bad, and it does not give clear prescriptions on what to do in every conceivable situation'. The reasons given for this are twofold:

> Firstly, there is no authority that owns the truth but is refusing to reveal it. Language testers are urged to discover this real but unknown truth by an appropriate choice of methodological and/or psychometric methods and to report their work so that in the (hopefully not so distant) future, we will reach a point where we have approximated the 'real truth' so closely that we can consider the problem as solved.
>
> Secondly, even in the case of a widely agreed frame of reference, the determinants of performances on a language test or examination are so varied (and imperfectly understood) that any attempt to categorise studies to link performances to the CEFR either as clearly good or clearly bad must be considered as simplistic and categorical. In reality, we are attempting to develop a system that gives insight into the strong and weak points of any such attempt (Council of Europe 2009:117).

In the first of these claims the reference to a 'real but unknown truth' may be seen to represent a realist philosophical position (rather than a naïve positivist one). The 'real truth' is that alignment which would best reconcile all potentially comparable assessments and contexts. The second claim points out the inevitably approximate nature of any alignment, for both theoretical, construct reasons and practical, measurement ones. These reasons are why the 'real truth' will be difficult to establish with any great certainty.

2.5.4 Validation: Some practical principles

While the later chapters will give an account of the validation undertaken for specific framework projects, below are proposed a number of principles which might constitute useful advice for validating a link into a multilingual assessment framework. As explained above, validation is seen here not as a *post-hoc* evaluation but as a continuous process embedded within the cycle of test development, administration and revision. The principles discussed below are:

- treat each context on its own terms
- focus on purposeful language use

- focus on salient features of levels
- focus on the learner
- use human judgement appropriately
- focus on linguistic features.

2.5.4.1 Treat each context on its own terms

As discussed at greater length in Chapter 4 on the CEFR, every context of learning is different and should be treated on its own terms. The argument that relates one context of learning to a level of the CEFR may be rather different from the argument presented for a different context. The CEFR being a framework of reference, it is intentionally underspecified so that differing contexts can find points of attachment. As the Manual states:

> There is no need for there to be a conflict between on the one hand a common framework desirable to organise education and facilitate such comparisons, and on the other hand the local strategies and decisions necessary to facilitate successful learning and set appropriate examinations in any given context (Council of Europe 2009:3).

The decision to link a learning context or an examination to the CEFR is finally a practical one reflecting acknowledgement of the fundamental value of communicative language ability, and aimed at enhancing the usefulness and interpretability of learning outcomes. To support the claim that a given course or exam targets a specific CEFR level it is necessary to identify relevant features of the course or exam which demonstrate the link sufficiently clearly. By agreeing to align differing contexts of learning within a framework, while treating each context on its own terms, we recognise that a claim of alignment is relative and qualified. Comparability between contexts is, equally, a relative and qualified matter.

We must also recognise and find ways of accommodating the fact that each new context linked to the framework enriches and extends the representation of the levels. We will return to the nature of a complex, multidimensional framework in the final chapter (see Section 7.2).

Clearly it is important to identify the significant dimensions that may distinguish one context from another. While the CEFR text provides a wealth of description in its Can Do scales and taxonomical lists, homing in on the significant distinguishing features of contexts requires some work. Jones (2013a) presents the CEFR as an instance of a more general framework, parameterised and illustrated for a particular foreign language learning context. Even for this narrow purpose it can be found lacking. For example, two foreign language contexts not best covered by the CEFR are young learners, because there is no explicit treatment of cognitive stage, and Content and Language Integrated Learning (CLIL) because language

for learning is not clearly distinguished from language for social use – the familiar distinction between Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP) (Cummins 1979, 1984. Cummins (1984) refines this distinction in terms of two intersecting dimensions. The first is a cline between *embedded* and *reduced* contexts of language use, and the second distinguishes lesser and greater cognitive demand. Social interaction tends to refer to the here-and-now (context embedded) and to be less cognitively demanding; academic use is one example of a context-reduced kind of discourse which tends to be more abstract as well as cognitively more demanding. But the two aspects are distinct. A CLIL context, for example, might well be cognitively demanding but also context embedded in practical experiments.

While the wealth of descriptor scales in the CEFR certainly include references to BICS- and CALP-like contexts of use, they do not raise this fundamental distinction to a level of explicitness. Such under-representation of important constructs makes it more difficult to describe and compare contexts, as well as to inform action, for example, aimed at helping children to acquire academic aspects of competence. An observation made at the Sèvres 2008 standard-setting conference for speaking (see Section 2.3.4.3) was that judges found it particularly difficult to compare the performance of adults and younger students, given their sometimes very different character: younger students might impress by their fluency and naturalness, while older students expressed themselves more slowly and carefully. Closer analysis revealed that in many cases the younger students were actually communicating a great deal less than the adults. Awareness of the distinctions made by Cummins would have helped to characterise these different profiles of competence.

2.5.4.2 Focus on purposeful language use

The assessment frameworks described in this volume share a focus on the ability to use language purposefully. As explained in Section 2.2.2, tests provide a score which may be interpreted as a *measure* of proficiency, where what we intend by 'proficiency' is determined by the test tasks. To measure the ability to use language purposefully we need to adopt a socio-cognitive approach, addressing the questions:

- Do the cognitive processes required to complete test tasks sufficiently resemble the cognitive processes a candidate would normally employ in non-test conditions, i.e. are they construct relevant (Messick 1989)?
- Are the range of processes elicited by test items sufficiently comprehensive to be considered representative of real-world behaviour i.e., not just a small subset of those which might then give rise to fears about construct under-representation?

- Are the characteristics of the test task an adequate and comprehensive representation of those that would be normally encountered in the real-life context?

We acknowledge that the relationship of language assessments to the real world is necessarily an indirect one, because it is not possible to recreate authentically within a test all aspects of target language use. Performance on test tasks (e.g. answering multiple-choice questions) may not directly relate to real-world Can Dos. However, we may appeal to the notion of *interactional authenticity*, 'a function of the extent and type of involvement of task takers' language ability in accomplishing a test task' Bachman (1991:691), citing Widdowson (1978)), such that they engage the learners' cognition in the same way as tasks in the real world. Through validation studies we can attempt to provide the evidence that supports that interpretation.

The socio-cognitive model of language use and learning presented in Section 2.1 above addresses in particular a perception that the CEFR provides too little specification or detail concerning the nature of cognition, as it relates to different language skills, and as it develops over levels of competence. Completed studies – notably those reported in four volumes of the SiLT series on the skills of writing, reading, speaking, and listening – Shaw and Weir (2007), Khalifa and Weir (2009), Taylor (Ed) (2011), Geranpayeh and Taylor (Eds) (2013) – provide useful implementations of the model. In seeking to align assessments to each other within a framework such as the CEFR these descriptions of developing competence may offer clearer articulation of constructs and more practical points of reference than the outcome-focused level descriptors in the body of the CEFR itself. To the extent that we can focus on socio-cognition, rather than on specific instances of language use, high-level comparison across contexts should be facilitated.

2.5.4.3 Focus on salient features of levels

Levels of a framework need not be described extensively and symmetrically: at each level only a subset of descriptive categories may be salient, and therefore relevant to describing the level. Figure 2.14 is a redacted version of CEFR Table 3: Qualitative aspects of spoken language use. It is a subjective reflection on an observation I made at the multilingual rating event held in Sèvres in 2008 (see Section 2.3.4.3 and Breton 2008), where Table 3 was used as the basis of all judgements. According to the procedure adopted, for every speaking sample participants provided ratings on all five of the categories – Range, Fluency, Interaction, Coherence and Accuracy. But in fact about half the cells in the table did not appear to contribute to identifying the level of a spoken performance.

The EPP has addressed two aspects of salience: lexico-grammatical, through analysis of extensive corpus data (Hawkins and Filipović 2012) and

Figure 2.14 The salient aspects of CEFR Table 3: Common Reference Levels: Qualitative aspects of spoken language use

	Range	Accuracy	Fluency	Interaction	Coherence
C2	Virtually everything	Precise	Appropriate		Coherent
C1	Broad range		Fluent, spontaneous	Flexible	Controlled
B2	Complex, own specialisation		Comfortable	Effective	Presents arguments clearly
B1	Transactions for living and working		Can keep going	Can deal with transactions	
A2	Familiar, routine				
A1	Personal				

functional, with reference to how levels are perceived in 'Can Do' descriptive terms (Green 2012). Focusing on salient features is not simply a question of simplifying the description of levels. As explained in Section 2.2.2.4, salience is related to test scores: it is whatever discriminates between learners at a particular level. What judges are most aware of is likely to be what accounts for the marks awarded.

2.5.4.4 Focus on the learner

Jones (2005a:18) criticises an apparent emphasis on task-centred methods in the standard-setting literature, and particularly in relation to the CEFR, asking: 'where have all the learners gone?' As argued in that paper, the largely North American standard-setting literature distinguishes between task-centred and examinee-centred standard setting, but focuses heavily on the former. Cizek's (2001) guide to standard setting contains just a few references to examinee-centred approaches, and while such approaches are said to be growing in popularity (2001:3) they are also criticised for threatening to hand authority over standards back to teachers, and thus failing the requirement to remove important educational decisions to a more objective level. Scepticism as to the competence of teachers to rate their own students in an accountable manner may be one reason for the relative absence of the learner in standard setting, also reflected in CEFR-related standard setting.

The authors of the Final Report of the Dutch CEF Construct Project (Alderson et al 2004:20) point out the significance of the move towards standards-based assessment, and the issues it poses for assessment:

> Rather than rating test takers as 'advanced', 'intermediate' or 'beginner', or as 'good', 'satisfactory' or 'fail' or in some other similar way, test takers receive a level rating, which describes what a person at that particular level can do with the language. This has been a challenge to psychometrics and test development and rating alike.

While there may be something in this, it neglects the fact that even in the days before the CEFR, and before the introduction of the measurement approaches presented in Section 2.2.2 above, there were contexts where levels of proficiency were quite well understood, albeit in intuitive and subjective ways. The publishers, teachers, test providers and perhaps to some extent the users of test results who worked with the Cambridge English exam levels could do so effectively because they knew and understood the learners. In fact, they had quite a good idea 'what a person at a particular level can do with the language'. That these levels constituted a set of practical targets for organising language learning was captured in the idea of 'natural levels' espoused by the director of Cambridge ESOL, Peter Hargreaves.

The above argument as presented in Jones (2005a) is in fact somewhat two-edged, because an understanding of levels based on familiar cohorts of learners, though it may work well within a particular context, is not guaranteed to hold good across different contexts: the 'natural levels' are not as natural as all that. Practical approaches to targeting standard-setting judgement on familiar cohorts of learners would differ according to whether the intended standard were to be local or to encompass a range of contexts. In the latter context self-assessment or teacher assessment might indeed not be trusted, as the experience of the ESLC confirms (see Section 6.7.5). However, using multilingual samples of performance in Speaking or Writing can be effectively used in cross-language alignment. In fact, triangulating observations of learners in different contexts against descriptions of levels such as those of the CEFR should be the most grounded and effective way of ascertaining if those levels are understood in the same way across contexts.

2.5.4.5 Use human judgement appropriately

As argued in the above section on standard setting, working with a multilingual framework requires human judgement to be used as sparingly as possible, for purposes to which it is suited.

Of the many standard-setting methods presented in the Manual (Council of Europe 2009), several do *not* require judges to determine the relative difficulty of objective test tasks, because those difficulties are provided from empirical data. This relieves judges of the necessity of doing something which, experience generally shows, they are not good at. Such methods are preferable to those which *do* require the judges to determine (i.e. guess) the absolute difficulty of test tasks, and thus their relative rank order. Thus for

example, scale construction for Asset Languages explicitly set out to minimise the role of human judgement in setting standards: level cut-offs on the measurement scale were defined in proportional terms using the Cambridge ESOL Common Scale as a template, a heuristic considered preferable to allowing different teams to set standards for each level of 25 languages (see Section 5.3.3).

Human judges can still mark productive skills (writing, speaking) better than a computer. The notion of *subjectivity*, though generally associated with problems of reliability of marking, is nonetheless central to the interpretation of language use: it concerns the impression made on the interlocutor in spoken interaction or in reading a written text. The problem is to ensure that subjectivity is exercised with reference to shared criteria. In practice it is the performance skills of writing and speaking which exemplify a CEFR level most clearly. The objectively marked skills of reading and listening are less directly interpretable, being measured on continuous scales which relate to indirectly observed mental processes. Using the performance skills to anchor the receptive skills is in practice almost inescapable in the way composite language examinations work (see Section 2.3.1), and good arguments in its defence can be found.

Human judgement generally works better where the judge has no personal interest in the outcome. In a summative, high-stakes context a teacher might rate their students too highly, while in a no-stakes context no such temptation exists. Thus for example the learner-centred approach to developing standards for Asset Languages exploited pretesting as an occasion for linking test performance to teacher judgements of level (see Section 5.3.3).

2.5.4.6 Focus on linguistic features

It may well be that for a single language the most context-neutral link will be a linguistic one. The EPP provides a detailed description of progression in linguistic terms which may prove to be reasonably stable across contexts of learning and of language use. The idea that the acquisition of a language system follows a roughly similar route, irrespective of teaching, is an attractive one, if doubtless disconfirmed in detail due to the impact of first language transfer and other contextual variables. To the extent that similarity can be demonstrated, learning a language to a particular level is essentially the same journey for everyone, the actual topics and communicative functions encountered being of secondary importance.

To the extent that linguistic progression is comparable across contexts it is likely that general notions (of time, place, existence, possibility, necessity etc.), which tend to be closely linked to a limited number of lexico-grammatical realisations, would also provide a stable point of comparison. Least stable across contexts are likely to be topic-specific features of levels, as well as the priority with which functional, communicative acts

are introduced. The treatment of topic range within the CEFR (Figure 2.14) assumes a progression from the personal, through the familiar and routine, to the transactional, and on to increasingly complex and abstract topics at the highest levels. This may reflect a typical approach to sequencing in mainstream language teaching, but it might well not be followed in a CLIL context or in language courses for professional adults, for example.

Concerning functional progression, we have already noted Spolsky's (1986:154) point that a needs-analysis approach might well identify different priorities for each specific learning context (Section 2.3.2 above); therefore, one cannot expect to find any natural order in a communicatively-defined course. But this assumes that needs analysis alone provides an adequate principle for organising an efficient programme of learning, something that nowadays we are perhaps less likely to accept.

2.6 Summary

In this chapter we have presented a general outline of scale construction as a series of stages, which is intended to be practical, in the sense that it is distilled from practical experience of the projects described in this volume.

Construct definition determines how each tested skill is conceptualised, for example in terms of a cognitive model, and thus underlies the whole approach to item writing and test development. It is the necessary first stage, at least, for a closed framework entirely in the control of its authors. An open framework such as the CEFR must also reflect a construct, even if expressed in more inclusive and general terms, for example the CEFR's 'action-oriented' conception of language as a tool for communication. The intentional under-specification of the CEFR enables a wide range of different assessments to seek alignment within it.

Scale construction is the next stage. The purpose is to develop scales which provide accurate measurement and criterion-referenced interpretation. The account in this chapter focused on the Rasch model as the IRT model used by Cambridge English Language Assessment, and objective testing approaches, partly because these offer a clear illustration of some fundamental scaling conceptions: the notion of the trait as a useful measurement dimension, the mutually defining relationship of person ability to task difficulty, as the basis on which traits are empirically constructed, and the composite nature of a performance level as a blend of task difficulty and performance quality. However, these concepts apply equally to subjectively assessed skills. The discussion of alignment across skills, languages and contexts raised a range of practical measurement issues and rather philosophical issues on the nature of comparability. In fact, everything comes back to comparability – of measures or of judgements. Standard setting emerged as a critical area and something of a work in progress.

Validity and validation were considered in relation to a multilingual framework. If validity concerns the interpretation of score meaning, then asserting the alignment of an exam to a level within the CEFR is clearly an important validity claim. The approach of the Council of Europe Manual, which envisages a one-off standard-setting event, was contrasted with that advanced by Cambridge English Language Assessment, which sees the pursuit of validity as a continuous process, located at a number of points in the exam development and administration cycle.

Finally, eight practical principles for addressing validation within a framework were proposed. These principles are distilled from a considerable amount of practice, and it is to the more detailed historical account of that practice that we now turn. The following chapters follow the development of related thematic strands or particular projects, starting with the earliest work on scale construction at Cambridge, from the beginning of the 1990s.

3 Scaling comes to Cambridge ESOL: The 1990s

3.1 Early scaling developments

The introduction to this volume identified 1998 as marking the beginning of the modern era in the history of Cambridge English Language Assessment. A dedicated EFL Evaluation Unit created in 1989 set out to establish a validation programme and research agenda specifically focusing on the EFL examinations, which led to many new projects in the early 90s. Milanovic (1996) identified seven key areas in which major advances had been made over a period of five or six years:

- major revisions of existing exams and development projects for new exams
- increased pretesting of materials and the setting up of an electronic item-banking system
- the rationalisation of data capture to allow for routine analysis of examinations
- research on the triangulation of test content, candidate background and test performance
- the rationalisation of systems to support oral examiners throughout the world
- research projects into the direct assessment of spoken and written skills
- item writer training programmes and increased investment in the development of key personnel (Milanovic 1996; also reported in UCLES 2000).

The scope of this research agenda led to a rapid expansion of capacity, from the original three members of the Evaluation Unit to more than 50 in the Research and Validation Group today. The development of an item-banking approach based on IRT (see Section 2.2.2), and the consequent focus on scaling and measurement, was central to much of this activity. Topics covered in this chapter include:

- Computer-adaptive tests
- The Cambridge ESOL Common Scale
- The ALTE Can Do project.

The Common Scale project was a long-term endeavour to place all Cambridge English assessments on a single proficiency scale based on an IRT measurement model. The origins of the approach were projects undertaken by Mike Milanovic and Alistair Pollitt to calibrate the International English Language Testing System (IELTS) test, and also coincided with the wider introduction of pretesting. The aim was to provide tractable and reliable ways of working with the existing suite of exams, supplementing traditional human judgement of standards with a new source of empirical evidence. In terms of Weir's (2005a) socio-cognitive validity model (see Section 2.1) this directly addressed scoring validity, making scores more interpretable and reliable. It also marked the beginning of a long process through which Cambridge English Language Assessment came to reconceptualise its exam suite as an integrated system rather than as a set of separate products.

As the Common Scale was beginning to take shape, an initiative was launched with partners in the recently founded Association of Language Testers in Europe (ALTE). This had the aim of aligning the partners' exam systems to levels within a single framework. It began with a content analysis of test material, and continued with an empirical project aimed at providing a description of the levels in Can Do terms. This project ran in parallel with the development of the CEFR, and shared several of its objectives. The CEFR development was also supported by Cambridge and ALTE partners, and accordingly, a later phase of the ALTE Can Do project incorporated an empirical link to the CEFR level descriptors, providing evidence which eventually led ALTE to align to and adopt the CEFR levels.

3.2 Computer-adaptive testing

Computer-adaptive testing (CAT) sounds like a rather ambitious area for an organisation just getting to grips with the use of Item Response Theory, but in fact it provided quite a practical low-stakes context for doing basic research. It was a completely new area, so that it could not be seen as competing with existing exams, and it was attractive as a possible way of recycling existing test material. Algorithms could be taken from the CAT routines included in the item-banking system I had developed for my PhD (Jones 1992).

3.2.1 Singapore Telecom

In 1993 an opportunity arose to develop a CAT system in conjunction with Singapore Telecom. It would offer subscribers an adaptive test of proficiency in English. It would be, or so it seemed, a stunning application of cutting-edge technology. An internal document (Jones 1993) explained:

> Singapore Telecom's Teleview system is an advanced photo-videotex system (akin to the UK's Prestel or French Minitel systems) that provides a wide range of information and other services. Users log on by personal computer via telephone modem or FM radio links. Educational services are popular, with a wide range of tutorial-style material already available, but this project, which brings true computer-adaptive testing into the user's home, is believed to be an altogether new development, not only for Singapore but on a world scale too.

Two kinds of test would be offered: a simple and cheaper practice test consisting of multiple-choice discrete items, and a full test including a wider range of items: cloze passages, multiple-choice questions on reading passages, and a constructed-response task type. The full test was expected to be sufficiently challenging to allow prediction of performance in one of the Cambridge ESOL mainstream EFL examinations: First Certificate in English (FCE), Certificate in Advanced English (CAE) or Certificate of Proficiency in English (CPE). The test would provide an appropriate recommendation.

The 1993 document presents the project as 'one application of the continuing work on constructing a "common scale" to describe the mainstream EFL exams, work which is proceeding within the IRT paradigm'. This was true, because the recycled test tasks used in the Singapore Telecom project were described at best by classical analysis data, and coming from exams at several levels required vertical scaling (see Section 2.3.4.1). Cambridge provided the calibrated items for the bank, with calibrations based on local trialling in Singapore, using a common items anchoring design. Cambridge also provided the algorithms, which were implemented by Singapore Telecom. In 1993 the practice bank was successfully in operation, with the launch of the full service planned for June 1994. The paper acknowledged some usability issues, due to technical constraints of Teleview's VAX minicomputer, although downloading to a user's own PC was seen as a possible future option. The Singapore project would provide us with valuable experience in what the document called 'a key technology for the future of testing'.

Unfortunately the project provided valuable experience of a different kind: that technology continually moves forward. This was the year that the World Wide Web burst upon the scene. Overnight, the cutting-edge photo-videotex system was obsolete, and the full test was never implemented. Still, this was, we believe, the world's first online computer-adaptive language test, and it laid the groundwork for a series of subsequent developments.

3.2.2 Linguaskill

The employment agency Manpower approached Cambridge in 1994 with a request for a simple language proficiency test to be used for placement of staff

in temporary employment. Requirements were that the test could be administered at any Manpower office, on demand, and without the need for specially qualified staff to mark it. It had to be short (1 hour the absolute limit) and yet as reliable as possible. Above all it had to accurately identify clients' practical language ability, in a general business setting. While tests for specific work areas, such as banking or accounting, might also be useful in future, the first requirement was for a general proficiency test, which would be equally applicable to a variety of settings. The first product was a test of English, which was soon followed by projects to develop tests in French, German, and Spanish, in collaboration with partners in ALTE: Alliance Française, Goethe-Institut and Universidad de Salamanca. Dutch was a later addition, in collaboration with the University of Louvain.

CAT was the ideal approach: it satisfied the need for a relatively short test which would nonetheless measure accurately over a wide range of levels. It would also satisfy all the requirements for administrative convenience. The test was conceived of as a general proficiency test, given business relevance by using texts and topics of general interest to people working in business, and extracting from the item banks anything which appeared inappropriately literary or academic. The original version of Linguaskill ran under DOS and was used in Manpower offices for a year. By 1996 it had been replaced by a Windows multimedia version.

From the multilingual point of view the Linguaskill tests broke new ground, requiring tests in four languages which would assess language proficiency to the same functional level. The validation project which preceded release included Manpower offices in a number of countries: mostly in Europe, but also in Mexico and Japan. The data collected included:

- actual test responses and results, including repeated sittings, enabling test retest analysis of reliability
- a paper-based (PB) anchor test to assist in final calibration of the items (the items having been only provisionally calibrated)
- questionnaires with demographic and soft feedback from test takers
- supervisor ratings: a set of rating scales describing functional abilities, which were completed by supervisors who were familiar with the test takers in a work context.

The supervisor ratings were key. They were used to set standards for each language, and thus to equate standards across languages. It was certainly a direct approach, and subject to a number of sources of error. It did appear, for example, that supervisors' perceptions of what constituted adequate performance in the office depended in part on local standards and the supply of job applicants. But by and large the approach was seen as valid by the client, Manpower, and the system was seen to work well by its users. It is striking that this early multilingual project thus applied a quite radical and authentic

criterion-referenced interpretation of test performance to the purpose of standard setting.

3.2.3 CommuniCAT

Having successfully demonstrated with Linguaskill the deployment of a low-stakes computer-adaptive test on CD-ROM, a number of further adaptive testing projects were undertaken, in collaboration with members of ALTE. The ALTE Framework (see Section 3.4) brought the ALTE members' examinations into an alignment based initially on an analysis of content. Computer-adaptive testing projects provided the focus for attempting to develop a common measurement dimension to the ALTE Framework, and thus represented a significant early step in efforts to implement a multilingual assessment framework.

ALTE News (Association of Language Testers in Europe 1998) reported that 'a smaller group of six ALTE members has been working on the *Communi*CAT project. They are the members representing Dutch (Certificaat Nederlands als Vreemde Taal), English (UCLES), French (Alliance Française), German (Goethe-Institut), Italian (Università per Stranieri di Perugia) and Spanish (Universidad de Salamanca).' This group went by the name KoBaLT, standing for Komputer Based Language Testing (curiously, as the only languages in which 'computer' is thus spelled appear to be Polish and Malay). The report described how a computer-adaptive test works, and proposed that *Communi*CAT could be used 'for diagnostic, work and general assessment purposes. It can be used for placement to decide a person's level at the beginning of a course or to assess their ability at the end of a course. *Communi*CAT can also be used to screen candidates to find the appropriate level of examination they should take.' The low-stakes role of CAT was thus emphasised.

In 2000 *Communi*CAT was the winner of a European Academic Software Award (EASA), as *Research Notes* (University of Cambridge Local Examinations Syndicate 2000) was pleased to announce, describing *Communi*CAT as 'the multilingual, computer adaptive language testing engine that drives such UCLES EFL products as CB BULATS [computer-based Business Language Testing Service], the British Council Placement Test and the UCLES/OUP [Oxford University Press] Quick Placement Test'. The report indicates that at this time computer-based testing (CBT) was a focus of significant effort:

> The CBT Team, co-ordinated by Michael Milanovic (Deputy Director EFL), has drawn on the skills and contributions from many in UCLES EFL over the past five years during which *Communi*CAT has been developed. The work on item banking and the calibration of items has been particularly important and the research of Neil Jones from the Research and Validation Group into computer adaptive testing was particularly commended by the EASA Jury (University of Cambridge Local Examinations Syndicate 2000:14).

3.2.4 BULATS – The Business Language Testing Service

BULATS is one of the *Communi*CAT family of adaptive tests which is of interest for two reasons:

- it was a CAT version of an existing PB test, and so raised issues as to the comparability of measures from different modes of testing
- it was the context of an interesting attempt to use multilingual candidates as a method for anchoring across the four tested languages (English, French, German and Spanish).

3.2.4.1 Comparability of PB and CB BULATS

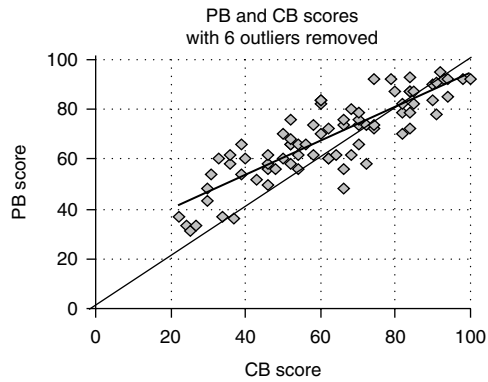
The first issue is the subject of a study reported in *Research Notes* (Jones 2000a). It identifies a number of general issues in comparing CB and formats, but taking BULATS as a case study focuses on specific issues arising in comparing an adaptive CAT test with a linear PB test:

- the effect of an adaptive mode of administration on test reliability, discrimination and the effective scale length of the CB CAT mode
- the effect of guessing in the PB mode (which is considerably less significant in a CAT).

The research project involved a basic test–retest design where volunteers took the test in two modes, in a random sequence. Comparison could not be based directly on raw scores, given the different modes of administration, but on scores representing abilities estimated using the Rasch model. The reliability of each test was high: 0.93 for the PB and 0.94 for CAT.

Figure 3.1 shows that there is quite good correlation between scores on the two modes. Statistical investigation suggested that this was about as good as would be expected from any two sittings of the test, irrespective of

Figure 3.1 CB and PB abilities compared



mode. Taking into account the experimental conditions, where the two tests were completed one after the other, producing variations in performance due to fatigue or inattention, it was concluded that the effect of test mode on the correlation of test results was minimal for this group of respondents.

None the less, Figure 3.1 shows that the spread of scores is narrower for the PB mode, as indicated by the slope of the trend line. In other words the CAT test mode is slightly more discriminating. This narrower spread of scores on the PB version of a test is characteristic. The adaptive CB test selects the most appropriate items for each candidate, according to their estimated level. It gives each candidate a chance to show just how high or low their level is. The PB test gives slightly less information, because each item is not at the optimal level for each candidate.

The study also included a simulation exercise which demonstrated that guessing in the PB mode was able to explain the generally higher scores on PB, particularly at the lower end of the ability continuum. The study concluded that the differences between the two modes of administration were systematic enough to enable a compensatory linear scaling to be applied. However, this was considered operationally problematic, and given the low-stakes nature of the tests, the issue was not followed up until some years later, when a new adoption of PB BULATS led to its higher-stakes use to make decisions at a low (CEFR A2) level, below the target level of the PB test. Here correction for guessing had to be estimated and introduced into the operational administration of the test.

3.2.4.2 Aligning standards across languages using multilingual candidates

Jones and Thighe (2005) is an internal validation report on a study in which students admitting to some degree of plurilingual competence took the BULATS test in two of the four BULATS languages, and also provided self-ratings of their ability in each language. The self-assessments were interpreted as estimates of students' *relative* ability in each language, which it was expected might be more accurate than absolute judgements of level. Thus it was hoped that this would support an approach to aligning the standard of the tests across languages. The overall correlations between self-assessment and test performance are shown in Figure 3.2. The relationship appears quite weak, because it includes effects due to misalignment of the tests.

Table 3.1 shows correlations by language, which makes the English data, with the highest correlation and the lowest score standard deviation (SD), look more coherent than the other languages.

Data was prepared for a multi-faceted Rasch analysis using FACETS (Linacre 2011). Each pair of self-assessment and test score was transformed into a difference score, represented by a positive integer between 1 and 10. The higher the difference score, the higher the test score relative to the self-assessment. The FACETS analysis produced a set of scaling adjustments,

Figure 3.2 Scatterplot showing relation of self-assessment to BULATS score

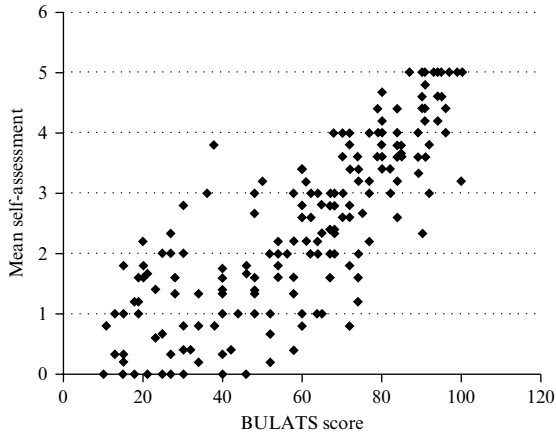


Table 3.1 Correlation of BULATS scores and self-assessment by language

Language	Correlation	SD BULATS scores	Sample size*
English	0.92	16.23	47
French	0.78	25.60	67
German	0.84	28.86	36
Spanish	0.70	22.87	46
All	0.82		196

** omitting candidates with incomplete self-assessments*

Table 3.2 Estimated scale adjustments by language (logits)

English	0
French	−0.5
German	−0.7
Spanish	−1.1

shown relative to English in Table 3.2, indicating a practically substantive amount of misalignment.

The report concluded that given the limited number of candidates in the dataset there was a need for further data collection and analysis before such adjustments could be implemented. The report further recommended that plurilingual data should continue to be collected, presumably in the operational operation of the test, and periodically analysed. Given the logistical problems of such data collection it is perhaps not surprising that this did not happen.

This study remains intriguing, as it produced a seemingly interpretable result, but was for logistical reasons not repeated. The idea was returned to in the tender document for the ESLC, where such a study was proposed, though not finally delivered. The practical difficulty with the design is of course to identify a sufficient number of informants with a plurilingual competence in the pairs of languages to be aligned. There is no ready-made identifiable population – each individual must be somehow identified. Still, it remains a variant on the use of comparative judgements which I find conceptually pleasing, particularly in the case of objectively measured skills such as reading or listening, where the learner is the person best placed to say just how much more or less competent they feel in one language compared with another.

3.3 The Cambridge ESOL Common Scale

The current suite of Cambridge English exams has developed over an extended period, with each new exam added in response to the recognition of a need. Historically, this process began in 1913 with the introduction of Cambridge Proficiency. Although now associated with Level C2 in the CEFR (the highest level), in 1913 this was seen as the lowest significant level of language competence worthy of certification. Over the years the notion of 'significant' has been progressively amended, so that now every level of the CEFR is supported by one or more Cambridge English exams. This large shift in perception reflects the changing role of assessment, from accrediting one or two valuable final outcomes to scaffolding the learning process from its earliest stages – assessment articulating a 'learning ladder'.

Weir (2013a:86), gives a historical account of the development of language assessment at Cambridge which points to concerns about the approach to standards in the early years.

> Roach, the Assistant Secretary at UCLES, had obviously been troubled by the absence of any specifications of performance levels, described in terms of agreed criterion descriptors. 'The Syndicate did not define standards of attainment to the examiners for the LCE when the examination was started in 1939', notes Roach. 'Candidates,' Roach worried, 'tend to set the standard in any test which has no absolute criterion'.

The critical importance of criterion reference, so perceptively identified by Roach, is a major theme of this book. As Weir concludes, 'the importance of standards of attainment was clearly understood, but the conceptual framework and criteria for their definition were not yet established. In 1944 the standards of attainment were those based on candidate performance without any recourse to external levels other than the remembered performances of candidates in previous years.'

Thus for much of their history Cambridge ESOL exams managed without the support of any scaling or measurement. Levels were understood by reference to the kind of learners placed in classes, the preparation courses and materials used to teach them, and the perceptions of teachers as to expected levels of performance. Exams were constructed and administered (then as now) by people familiar with and recruited from these contexts. There were also considerable differences between exams, not only in terms of level but in the way that they defined and tested language competence. This situation changed, though not overnight, when IRT scaling procedures were introduced to Cambridge exams in the early 1990s. The construction of a common scale to link exams at different levels to a single measurement scale was to provide a statistical underpinning for a system of exam levels which had grown up based on an experiential, professional understanding of levels, and which for some time continued to trust in it to a certain extent. Central to this shared understanding of levels were the large and relatively stable groups of learners who constituted the candidature for each level.

However, the stability of the candidature could not be taken for granted. In the early 1990s Cambridge ESOL began to collect information about candidates, using machine-readable forms that enabled the capture of a large amount of data. Candidates' age, the length of time spent learning English, and whether or not they had followed a preparation course were among the data captured. This contributed significantly to understanding the makeup of the candidature, and to detecting possible changes over time. Traditional sources of information which continued to be referred to at grading events, even long after the introduction of the Rasch model, included reports on performance in specific countries, with particular reference to countries believed to be stable in terms of the size of the candidature and its makeup. The adoption of the Rasch model and an item-banking approach to test construction created the possibility of constructing a measurement scale to link the existing exams.

3.3.1 Pretesting and calibration

When I joined Cambridge in 1992, having completed a PhD in applying the Rasch model to language testing, extensive pretesting had recently started. Classical analysis was done using the ITEMAN program (Assessment Systems Corporation 2013), and some Rasch analysis was beginning, using BIGSTEPS, the DOS forerunner of WINSTEPS (Winsteps 2013). An initial requirement was to develop bespoke software in order to streamline analysis, for example, enabling a single file format to be shared across software packages, and to provide simple uniform approaches to quality control, for example by applying standardised checks on the quality of anchor items.

Pretesting was the necessary first step towards implementing an item

banking approach, although it was to take some time before all examinations could adopt it. In fact, the construction of the Cambridge ESOL Common Scale began not only with pretesting but also with two long-term research projects: the Monitoring of Exam Difficulty project, and analyses of candidates taking more than one exam.

3.3.2 Monitoring of Exam Difficulty

The Monitoring of Exam Difficulty project used an experimental design to construct a link between live exams, using a series of anchor tests. Centres which agreed to participate had students who were enrolled to take a Cambridge exam at one of the two levels to be linked. Centres administered the anchor tests to these students at about the same time as their live exam. This created additional work for the participating exam centres, and so the design had to be kept simple: the anchor tests comprised short tests of reading. In one exam session several different anchor tests were used, chosen to be of appropriate difficulty, that is, somewhere between the levels of the two exams to be linked. The first two exams to be included in this project were in fact FCE and CPE, nowadays associated with CEFR levels B2 and C2. This was a large gap to bridge with a single anchor test. With the introduction of CAE at what is now C1 level the anchoring design became more complex, but also more tractable. As new Cambridge ESOL exams appeared the scope of the project was extended to link these: first the Preliminary English Test (PET) (at CEFR B1) and then KET (at A2).

By 2003 the Monitoring of Exam Difficulty project encompassed 14 separate assessments: KET, PET, FCE, CAE and CPE, BEC at three levels, and the now discontinued Certificates in English Language Skills (CELS) suite of modular exams for Reading and Listening, also at three levels. The 2003 programme also set out to create a link between the anchors used in the live exam context and those used at pretesting. Following this the focus of anchoring the levels transferred fully to pretesting and the Monitoring Of Exam Difficulty project came to an end.

3.3.3 Candidates taking two exams

It may be surprising to learn that significant numbers of candidates for Cambridge English exams actually choose to take two different exams in the same session. This fact enabled an approach to understanding exam levels which was in use by 1991 and continued until about 2002. It depended on identifying candidates taking two exams, a somewhat painstaking task as it involved matching candidates at a given centre by name and date of birth.

Table 3.3 shows the basic form in which results were presented: a contingency table of the number of candidates achieving each combination of exam

Table 3.3 Grades of candidates taking CPE and FCE exams, December 1992

FCE grades	CPE grades					
	A	**B**	**C**	**D**	**E**	**Totals**
A	20	66	121	10	14	**231**
B	0	4	72	24	48	**148**
C	2	3	7	4	51	**67**
D	0	1	0	0	7	**8**
E	0	0	0	0	12	**12**
Totals	**22**	**74**	**200**	**38**	**132**	**466**

grades in two exams. The table shows, for example, that most candidates achieving grade C in CPE achieved grade A in FCE.

An early internal report (Royal-Dawson 1994) reviews sessions between 1991 and 1993. It is worth quoting at some length for the insight it gives into the difficulty of establishing standards in a multi-level exam framework before the introduction of statistically based scaling using IRT. The exams referred to – FCE, CAE and CPE – are nowadays associated with CEFR Levels B2 to C2 respectively.

> CAE was first administered in December 1991. The number of candidates taking it has increased as it became more widely known. A new EFL examination on the market targeted between FCE and CPE would have attracted students from diverse backgrounds. By dredging the names of candidates, 10% of the CAE candidates were found to be also registered to take FCE, indicating a certain unease about the new examination.
>
> Added to the uncertainty about the level and background of candidates taking the new examination was the task of setting a pass mark and concomitant grade boundaries for a new examination. In June 1992, 73% of the candidates passed the examination. This was felt to be too high because the placement of CAE between FCE and CPE was too close to FCE as shown by the candidates taking FCE and CAE. Candidates passing FCE with a C or above still stood a good chance of passing CAE prior to June 1993.
>
> It is supposed that the true level of CAE has not yet settled because it is not known how CPE candidates perform at CAE. If a candidate passing CPE at grade C is able to pass CAE at grade C, it would be clear that CAE is not very distinct from CPE either. This however is not very likely. In December 1994, CAE will be set on a different day to CPE, so for the first time, candidates will be able to take CAE and CPE in the same session.
>
> In June and December 1993, careful efforts were made to ensure that there was a clearer distinction between FCE and CAE. This was shown by the shift between the grades obtained by the candidates who took

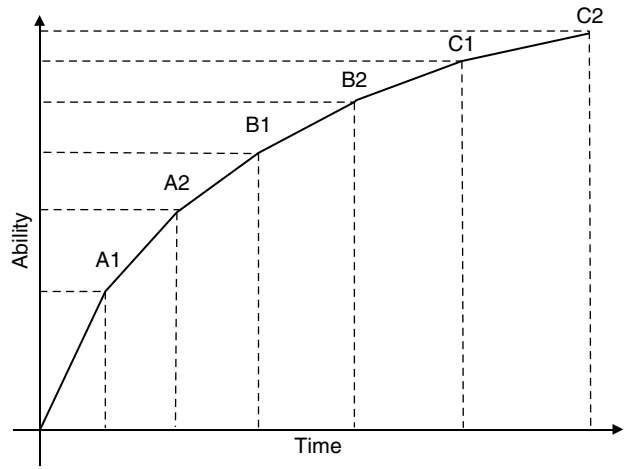
> both FCE and CAE. In June and December 1993, only 25% of the candidates passing FCE with a grade C passed CAE compared to nearer 50% in the previous administrations. In addition, a smaller proportion, 61%, of the overall population passed CAE in December 1993 than in the previous sessions indicating again the conscious effort on the part of the grade boundary setters to make a clearer distinction between the levels of FCE and CAE.

A later internal report (Banks 1999) shows many more exams included in the study: KET, PET, BEC 1, 2 and 3. It also reports Rasch analysis being used to calibrate PET, using the anchor to FCE provided by the common candidates, and to specify PET grade boundaries on the Common Scale. Thus by this stage the pragmatic aim of locating exams within a multi-level framework had merged with the psychometric ambition to construct a stable measurement scale to underpin the different exams.

3.3.4 The shape of the Common Scale

Figure 3.3 offers a sketch of the Common Scale, showing the relationship between levels of ability on the vertical axis, and the time required to achieve each level. It is intentionally schematic, in order to illustrate clearly the classic learning curve shape. In reality the ability dimension has been constructed empirically by scaling each exam, and then anchoring these scales together using pretesting and experimentally collected data, as described above. The time dimension reflects common wisdom based on the experience of exam users.

Figure 3.3 Schematic view of the Common Scale



Although the core Cambridge ESOL exam levels evolved without statistical support over an extended period it demonstrates a certain regularity. The story it tells is as follows. What we can measure is proportional learning gains. At the earliest learning stages a relatively small amount of effort produces a substantial change in observable behaviour – enough to warrant identifying a level and offering accreditation of it. Subsequently it takes progressively longer to make a substantial difference, and indeed, many learners plateau or drop out on the way. The higher levels are separated by smaller measurable differences, but each level is needed because it accredits a final learning achievement or provides an interim target for those who wish to go further (Cambridge ESOL's CAE exam was introduced at C1 explicitly to bridge a perceived gap between FCE and CPE). The Common Scale illustrates the learning curve, first identified by Ebbinghaus (1885), a concept widely used in industry and economics to predict the efficiency of processes involving training or experience. It illustrates a general law on diminishing returns of effort, investment etc.

The characteristics of the Common Scale are discussed further in Section 4.3, where it is compared with the CEFR Can Do scales. Its use as a template learning curve for defining the Asset Languages levels is presented in 5.3.3 below.

3.3.5 The Common Scale evaluated

By the first decade of the 21st century the Common Scale became embedded in Cambridge ESOL's operational practice, as test construction moved to an item-banking model supported by the in-house development of the Local Item Banking System (LIBS). Test construction guidelines were established for each exam and level, specifying the target mean and the range of difficulty of test items in common scale units. Common scale grade thresholds for each exam were also specified, with the intention of ensuring the application of constant standards. With the transfer of the Common Scale into the realm of operational exam production the non-operational research effort around it decreased. Alongside statistical estimates of common scale difficulty, grading events continued to involve subjective judgement, particularly in respect of the performance skills of writing and speaking. Reference to normative information on pass rates over sessions and in particular countries played as important a role as the statistical link to the Common Scale. There was no clean hand-over of grading authority from the subject experts to the statisticians, which is understandable to the extent that the statistics were derived from processes which were known to be imperfect (for example, issues with the methods available for pretesting).

An internal paper from 2000 looked backwards at the Common Scale development to date, while proposing a research programme to carry it forward. The paper finds compelling reasons for developing a measurement

framework and promoting Cambridge EFL exams as components of a five-level 'system':

- Users of test results (employers, admissions bodies, the test takers themselves, and so on) increasingly require that exam results should relate to an inclusive framework.
- There is a move away from traditional 'academic' exams towards life-long, continuing education. Continuous re-training and updating of skills is becoming the norm. In this more utilitarian atmosphere the meaning of language qualifications needs to be clearly spelled out.
- There is a need for European standards, enabling cross-language comparison of levels. This has prompted the extension of the Cambridge five levels into the ALTE Framework.
- New tests (placement tests) and testing methods (CB tests) have an explicit requirement to measure over all levels and report in terms of the 5-level scale (Jones 2000b:1).

The paper acknowledges that vertical scaling involves simplification of a complicated reality, subject to multiple dimensions of variability, but continues:

> However, it is possible to conceive of the Common Scale as an exact, if ideal, construct – a reference line in relation to which we can accurately characterise the complex of observable features for any given exam or individual. The usefulness of this would be that in addition to providing the simple scheme it would also support more accurate characterisations of the complex reality.

This suggestion appears to point forward to what has become a current field of research interest: how diverse sources of information – qualitative and quantitative – can be used in complementary ways, not simply to scale but to characterise contexts of learning. This is at the heart of two current research areas: studies concerning the impact of Cambridge exams in particular contexts, and work to define a model for LOA. There is more on these in the concluding chapter.

Let us conclude by bringing the Common Scale up to date. It still stands at the centre of the complex systems which support the conduct of the Cambridge English assessment business. It continues to contribute to the validation and quality assurance of the exams, sometimes in new ways. For example, work within the EPP to develop a detailed linguistic description of English across the levels of the CEFR depends on information on the levels of the exam candidates whose responses make up the corpus on which the findings are based.

So far the role of the Common Scale in supporting the exams has remained

largely invisible to end users. It has not figured explicitly as a communicative device, even though in recent years the exams have been increasingly presented as a coherent, integrated system of levels. Systematic reference to the CEFR has greatly emphasised the unity of the system. Thus there is an opportunity to articulate this more clearly by making the Common Scale, or a reporting scale derived from it, a visible element in how the exams are presented and their results interpreted.

Writing in 2013, work is proceeding to provide such a single reporting scale, covering the CEFR range of levels, as information additional to the exam grade achieved. This raises issues of presentation and interpretation which require careful consideration, not least to distinguish the purpose and value of different kinds of exam within the Cambridge English offering: multi-level tests, like BULATS or IELTS, serve different purposes to those of the level-based tests constituting the Cambridge English core exams. It is the strength of these exams that by focusing on a single level they provide an excellent focus for a programme of study, as well as accurate accreditation of learning outcomes. The purely psychometric perspective potentially offered by the Common Scale should not be allowed to obscure the broader educational perspective. There are no simple answers to how such additional levels of interpretation can be usefully provided.

3.4 The ALTE Framework and Can Do project

3.4.1 The ALTE Framework

When ALTE was set up in 1990, one of its stated aims was 'to establish common levels of proficiency in order to promote the transnational recognition of certification in Europe'. To this end, a long-term project was envisaged with the final aim of establishing a framework of levels within which meaningful comparisons between qualifications in different languages, gained in various states of the European Union, could be made. The ALTE Framework located members' exams according to a set of levels (later aligned to the CEFR) using judgemental methods based on close analysis of test task features. A range of projects were undertaken to enable meaningful comparison of the exams: content analysis checklists, item writer guidelines, and even the compilation of a glossary of testing terms (all developed in all the languages of the ALTE members. The ALTE Can Do project set out to validate the framework by comparing test takers' self-ratings of their ability with exam grades achieved. The ALTE 'Can Do' project ran in parallel to the finalisation of the CEFR and its descriptor scales, and finally included studies to establish a link between these two frameworks. As well as providing a useful interpretative framework for exam users, it informed several revisions to the placement of exams in the ALTE Framework, and thus served a useful heuristic function.

The construction of the ALTE Framework served a useful objective – to enhance the meaning of the different members' exam levels by aligning them to a framework which could achieve wider currency and understanding. The basis for comparison and alignment across languages was the intuitively simple notion of equivalent levels of functional language ability, and undertaking a Can Do study was thus an appropriate approach to validation.

ALTE's aims were thus somewhat narrower than those of the nascent CEFR project, being confined to the realm of assessment, but responded to the same social developments: the increasing mobility of European citizens, workers and students within an expanding European Union. Students, teachers, employers and employees need to know what language qualifications gained in a variety of countries mean in practical terms, and how to make meaningful comparisons between qualifications gained from different awarding bodies situated in different states.

In fact, the work of ALTE illustrates many of the social and educational priorities which the measurement frameworks described in this volume set out to address, and their coherence with the goals of European language policy. In the 1990s the European Union's Lingua Programme echoed the priorities of the Council of Europe's Modern Languages Project. The Lingua Programme's three main objectives were:

- to encourage and support linguistic diversity throughout the EU
- to contribute to an improvement in the quality of language teaching and learning
- to promote access to lifelong language learning opportunities appropriate to each individual's needs.

More recently language requirements have also been playing increasingly important roles in domains other than mainstream education. Decisions on immigration or the granting of citizenship through naturalisation processes often depend on language test results. This has raised the stakes for many language learners – and raises the question of which proficiency levels should be set for which purposes (Saville 2012).

Reliable, valid, and above all comparable and meaningful standards are essential here. Measurement frameworks which take the CEFR as their point of reference have important roles to play. It is significant that the event used to launch the CEFR in its published form was the first European Year of Languages, jointly organised by the European Union and the Council of Europe in 2001.

ALTE set itself three goals: the first, to establish common levels of proficiency across languages in order to promote the transnational recognition of certification in Europe, is clearly coherent with European language policy centred on the CEFR.

ALTE's second stated aim, to address issues of quality and fairness in

examinations and to establish professional standards for all stages of the process, underlines the importance of the quality of the assessment system that leads to fair outcomes.

The third ALTE objective is a condition for achieving the other two. It acknowledges the importance of collaboration on joint projects and the exchange of ideas, know-how and best practice in the field of assessment – in other words, to form a community of practice to bring about improvements to professional standards. The work on implementing measurement frameworks described herein has provided a major focus for such collaboration.

To return to the ALTE Framework project: Saville (1995) outlines its major phases. Phase one of the project ran from 1994 to 1995, supported by funding from the European Lingua Programme, and was entitled *The description and comparison of foreign language qualifications in the EC*. It had two aims:

1. To develop a means of analysing the content of test materials so that they can be compared across a range of languages.
2. To develop a set of materials for the guidance of test item writers.

These materials were developed in the range of European languages represented by the project partners, and were intended to enable comparison of the levels of exams in different languages. On this analytical basis the exams of ALTE members were aligned within a single system: the ALTE Framework.

Phase 2 of the project ran from July 1994 to June 1995, and was also supported by European Lingua Programme funding. Its title was *An instrument for the provision of activity-based curricula, linguistic audits and diagnostic test tasks*. This phase produced a series of Can Do statements, expressed in clear, everyday language intended to be comprehensible to learners, employers or other interested parties. The statements were written and grouped within three categories of experience: 'Social and tourist', 'Work' and 'Study', which were seen as the three main areas of interest for adults learning a foreign language. Within each category a number of more particular concerns were identified, such as 'health' and 'travel' within the Social and tourist category. These were further broken down into 'activities' and the skills of listening/speaking, reading and writing. Figure 3.4 illustrates. Further illustrative scales are provided in Appendix B.

Note that the ALTE Framework identified five levels, which were subsequently taken to correspond to CEFR Levels A2–C2. That is, the ALTE Framework did not at that time have a level corresponding to A1, although work on Breakthrough level was undertaken by the so-called FINGS group within ALTE (Finnish, Irish, Norwegian, Greek and Swedish), for whom an A1 level was required.

An example Can Do scale is illustrated in Figure 3.5.

Figure 3.4 Structure of the ALTE Can Do statements

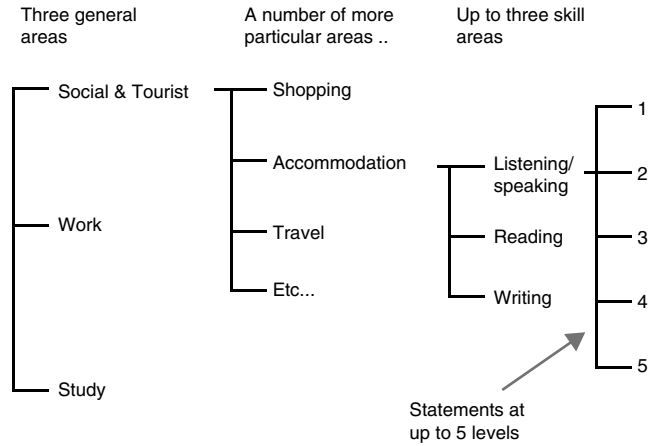


Figure 3.5 Selected statements at Levels 1–5 from an example Can Do scale

Area	Work
Activity	Requesting work-related services
Environment	Workplace (Office, factory etc.)
Language skill	Listening/Speaking
1	CAN state simple requirements within own job area, for example 'I want to order 25 of …'.
2	CAN ask questions of a fact-finding nature, for example establishing what is wrong with a machine, and understand simple replies.
3	CAN put her/his point across persuasively when talking, for example about a familiar product.
4	CAN give detailed information and state detailed requirements within familiar area of work.
5	CAN argue his/her case effectively, justifying, if necessary, a need for service and specifying needs precisely.

3.4.2 The Can Do project

Phase three of the ALTE Framework project focused on the empirical calibration of the Can Do statements, to enable them to 'form the intuitively derived backbone of the framework' (Saville 1995:3). The statements already existed in translations into 10 different European languages (the final number was 13: Catalan, Danish, Dutch, English, Finnish, French, German, Greek, Italian, Norwegian, Portuguese, Spanish, and Swedish).

It was important to establish whether users understood the statements in the same way and ranked them in the same order. Finally this would allow the statements to be placed at a series of levels. An experimental design was proposed involving language teachers, students and employers. In the final event however the major source of data for calibrating the Can Do statements was the self-ratings of large numbers of language students, as reported in an appendix to the text of the CEFR: 'Appendix D: The ALTE 'Can Do' statements' (Council of Europe 2001:244). See also Jones (2002).

Each respondent completed a questionnaire on one of the three categories: Social and Tourist, Work or Study. In order to link these categories and enable comparison a common section was needed as an anchor. Initially a subset of the Social and Tourist statements was included for this purpose, making the assumption that these statements would call upon a common core of language proficiency and serve as a valid point of reference for linking all three domains of language use. In a later phase these were replaced by descriptors taken from the Council of Europe Framework document (1996 edition), which provided an additional anchor to the CEFR levels, as described further below.

The questionnaires consisted of Can Do statements grouped by situational area and skill, and ranked in expected order of difficulty from low to high. A yes/no response was elicited to each statement:

- Put ONE tick next to each statement.
- Tick 'YES' if the statement describes your level, or if you 'Can Do' BETTER than this.
- Tick 'No' if you CAN'T do what is described because it is TOO DIFFICULT for you.

Nearly 10,000 respondents completed questionnaires, and for many of these respondents additional data was available in the form of language examination results. Evaluation of this rich dataset focused initially on how the individual statements functioned within each Can Do scale.

3.4.2.1 Initial analysis

Comparison of the intended and observed ranking identified several systematic and explicable problems with the wording of statements. To take one example, that of negative phrasing:

- Negatively phrased statements functioned incorrectly with higher level students. Thus the statement 'CAN make simple complaints, for example, "the food is cold". CANNOT argue/complain effectively, for example about the service' required the negative qualification to be removed: 'CAN make simple complaints, for example, "the food is cold".'

- However, positively phrased qualifications functioned well, e.g. 'CAN take part in "small talk" with peers, but MAY have problems with people of a different age-group or background'.
- Other negative statements required positive rephrasing, e.g. 'CANNOT ask more than simple questions for further information' as 'CAN ask simple questions for further information'.

Such problems with the text of the Can Do statements were identified during the first pilot study and corrected in time for the major data collection.

The analysis approach was to bring the response data together into a single dataset, including all the languages in which the questionnaire was completed (L1, first language), and all the target languages which the respondent was describing (L2, second language). The analysis finds the line of best fit through all the responses, so that the difficulty of any statement reflects an averaging across all L1s and L2s in the data. Having found this, then analysis of *fit* can be used to identify those statements that show differential item function, or *bias* – i.e. that particular groups respond to in a significantly different way.

There was evidence that some statements demonstrated bias for particular groups. Concerning groups of learners of particular languages one interesting effect was noted: comparing learners of French and English, it was found that whatever their overall level, learners of French were likely to be relatively more confident of their receptive language skills; learners of English on the other hand to be relatively more confident of their productive communication skills. One might hypothesise that this reflected a substantive difference in how these two languages are commonly taught.

There was some evidence that specific groups of language users might understand Can Do statements differently. For example, concerning different professional groups it was found that employees at middle or junior level judged it relatively harder than more senior staff to deal with routine letters or understand the fax machine and the photocopier. This suggests that these two groups differed in how they habitually engaged with these activities, and hence in their understanding of the meaning of the Can Do statements.

3.4.2.2 Linking to the CEFR descriptor scales

The ALTE Can Do project succeeded in establishing a reasonably firm link to the descriptor scales of the CEFR. The link was implemented by including in the ALTE'Can Do questionnaires two sets of statements from the 1996 version of the CEFR. One study used the descriptors in the self-assessment grid (Table 2 in the 2001 version). A second study used 16 descriptors relating to communicative aspects of Fluency. The latter provided a particularly stable anchor across groups, and the difficulties found in the analysis correlated very highly ($r = 0.97$) with those originally reported by North (1996).

The linking was not without issues. As reported in the ALTE appendix to the CEFR:

> Table 2 [i.e. the descriptors of the CEFR self-assessment grid] produced a longer scale, distinguishing finer levels than the ALTE 'Can Do' statements. The likely reason for this is that Table 2 represents the end product of an extended process of selection, analysis and refinement. The result of this process is that each level description is a composite of carefully selected typical elements, making it easier for respondents . . . to recognise the level which best describes them. This produces a more coherent pattern of responses, which in turn produces a longer scale. This is in contrast to the present form of the 'Can Dos' which are still short, atomic statements which have not yet been grouped into such rounded, holistic descriptions of levels (Council of Europe 2001:248–249).

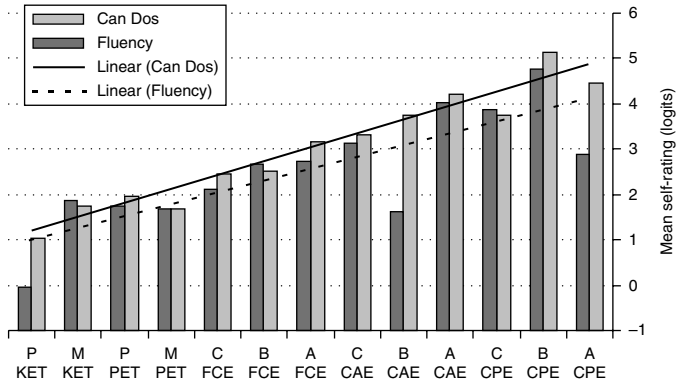
The Can Do difficulties were therefore scaled, using the ratio of the spread of person abilities as estimated separately from the Can Do and CEFR statements. This spread out the Can Do statements, approximating more closely the original assignation of Can Do statements to ALTE levels. Jones (2002) describes in more detail the approach to linking the two scales.

3.4.2.3 Linking Can Do statements to exam levels

Figure 3.6 shows the degree of correspondence between self-ratings and exam performance, on the example of Cambridge ESOL exams. Similar comparisons were possible for the exam systems of several other ALTE members. The figure shows the mean self-rated ability of candidates grouped by the exam grade which they achieved). The exams are ordered by level (KET = ALTE Level 1, CPE = ALTE Level 5).

Figure 3.6 shows self-ratings on the Can Do statements and on the CEFR

Figure 3.6 Mean self-ratings (Can Do statements, Fluency) by exam grade



'Fluency' statements separately estimated. A clear relationship is evident between self-rating and exam grade achieved (the odder values are due to very small numbers of candidates in particular groups).

Grouping on exam grade, a high correlation was found between mean self-ratings and exam grade achieved. Table 3.4 shows that the Can Do ratings bear a slightly closer relation to exam grade achieved than do the 'Fluency' statements.

Table 3.4 Correlations between exam level, Can Do and Fluency self-ratings, grouping by exam level achieved

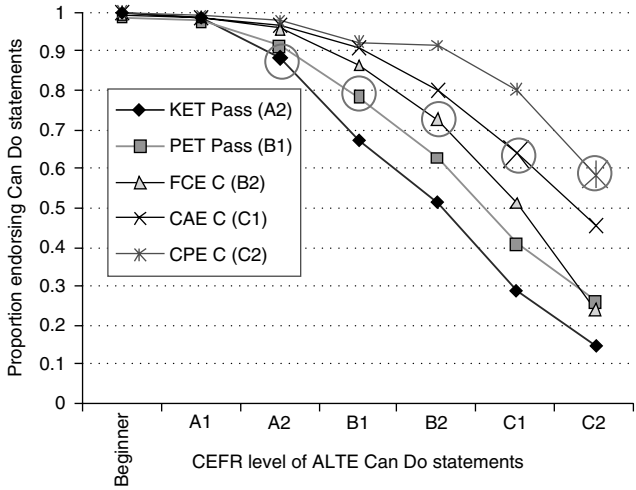
	Can Do	Fluency
Fluency	0.86	
Exam level	0.91	0.79

Summarised by exam group, the strength of the relationship between exam grade and self-rating of ability is clear. None the less, there is considerable variability in self-rating at the level of the individual respondent, which weakens the relationship between individual self-rating and exam grade.

Figure 3.6 suggests what other analysis indicated, that lower level learners appeared to overestimate and higher level learners to underestimate their ability. This finding has been noted elsewhere and folk-wisdom explanations of it have been proposed: that lower level learners do not realise how much they still have to learn, while advanced learners recognise that learning a language is a never-ending process. In fact it is not necessary to invoke such psychological explanations, as the phenomenon is easily explained as an example of regression to the mean. All measures including self-ratings are subject to error. Learners classified as beginners by a test have *de facto* not benefited from error in that test, and the error in their self-ratings can only make their mean score on these higher. Conversely learners classified as advanced by a test have *de facto* not been penalised by error in that test, and the error in their self-ratings can only make their mean lower. This predictable effect made it difficult to use this self-report data in a straightforward way to assign Can Do statements to CEFR or ALTE levels. It was felt that a definition of 'mastery' was needed, such that Can Do would refer to a fixed probability of a learner endorsing a particular statement, and hence, it was reasoned, succeeding on the task described in the statement. The ALTE appendix to the CEFR suggests a value of 80%, as one 'frequently used in domain- or criterion-referenced testing as an indication of mastery in a given domain'. If self-ratings accurately reflected the true difficulty of tasks, then the probability of respondents at a given level endorsing Can Do statements which describe that level should be constant (e.g. 80%) across all levels.

However, analysis showed that rather than remaining constant across levels, the probability of endorsing statements identified by the analysis to be at the level of the exam dropped from 90% at Level 1 to less than 60% at Level 5 (Figure 3.7).

Figure 3.7 Probability of a candidate endorsing Can Do statements at the level of the exam taken



Assuming that the problem lay with the responses of learners and not with the placement of the statements, the data showed a mismatch between the notion of a fixed criterion mastery level and the final assignment to levels of the Can Do statements. This clearly reflected the regression effect with the self-report data described above, and it complicated its use in assigning Can Do statements to exam levels.

3.5 Conclusions: A pioneering age

For Cambridge ESOL the early period of scale construction described in this chapter has all the marks of a pioneering age: new territories encroached on, if not fully conquered; rapid innovation and progress; exploring new frontiers and breaking with conservative tradition. Clearly, the software would be an issue.

The systems that were put in place within the Research and Validation Group worked initially alongside the traditional mainframe computer, which continued to produce the same, standard reports upon which all UCLES exams depended. In time a PC-based system replaced the mainframe, via an extended development process which gave each of the three

business streams (as Cambridge ESOL formally separated from Oxford, Cambridge and RSA Examinations (OCR) and Cambridge International Examinations (CIE)) most of what they needed for standard exam processing. It was possible to pursue additional specific developments with the assistance of the more flexible desktop systems branch of IT. The most significant of these was LIBS, which enabled the more efficient production of test tasks and the application of IRT scaling procedures to item calibration and test construction.

LIBS was described in a two-part contribution to *Research Notes* (Beeston 2000:5). Beeston could report that:

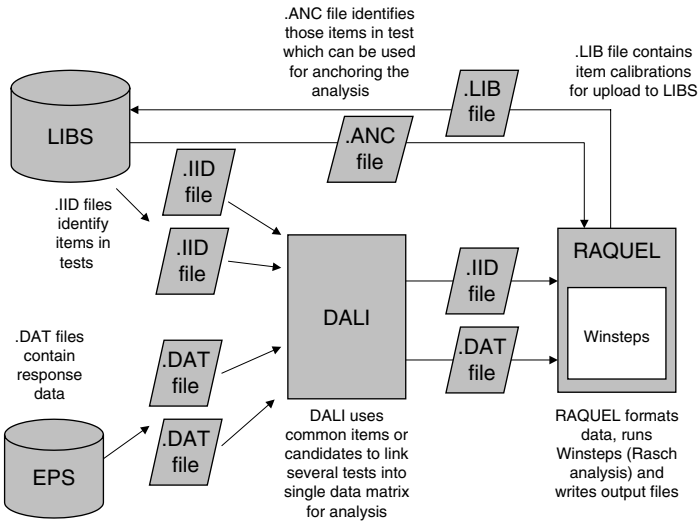
> Since August 1998, the EFL Test Development and Validation Group have been entering exam material onto the Item Bank and training staff in order to support the introduction of LIBS, which will house all of the material used for the Cambridge EFL [core] examinations, as well as IELTS, the Business English Certificates and a number of other examinations.

The benefits emphasised included the imposition of standard procedures, with many clerical tasks being automated, such as loading of test statistics, formatting, part numbering, and the generation of routine test construction reports, answer keys, comparison of pretest and live test statistics, and so on. Automatically calculated question paper preparation schedules and email alerts ensured timely delivery ('late papers' being a constant concern in those days). The improved security of test material was also stressed. LIBS was very successful in terms of its foremost purpose, and with re-developments continues to support the business up to the present. The development benefited from a close working relationship between EFL staff and the small team of in-house developers, which favoured a collaborative, interactive approach to development.

Alongside these two levels – the heavyweight Exam Processing System (EPS) and the more agile collaborative developments – there was a third more informal level: the suite of programs I initially produced to facilitate operational statistical analysis within the Research and Validation Group. There were many of them, each built separately, but growing into a system of interconnected operations: data formatting and merging, marking and classical analysis reports for non-standard item types, item calibration and anchoring, and more besides. The LIBS development added the requirement to link into the analysis cycle, by extracting from LIBS information on items and tests, and then uploading the analysis results back into LIBS, having preserved the item IDs through each stage of the analysis process. This linking was accomplished in the same way as every other step, by writing output data into files which could be used as input to the next stage.

Figure 3.8 depicts this cycle, showing only the central processes. It shows the files through which the analysis steps were linked. Each step required manual intervention on the part of the analyst, which became increasingly burdensome as the exam suite developed and the number of sessions grew. Asset Languages (Chapter 5) nearly broke the system, with its tests in 25 languages, two objectively marked skills, and multiple permutations of levels, versions, learner groups and sessions. In fact the final phase in this development averted an Asset Languages crisis by engineering a transformation which allowed a list of tests for analysis to be read into an Access database, which then processed them automatically, executing all the steps in sequence without human intervention, and writing a useful summary report on each step as it went.

Figure 3.8 Data transfer through steps in analysis



'C'est magnifique, mais ce n'est pas la guerre!', as Marshal Bosquet evaluated the Battle of Balaclava. Apart from anything else, to have so many critical processes depending on essentially unsupported software represented a substantive threat to the business. Finally, a training and development programme was set in hand to develop the programming expertise within the Research and Validation Group to produce an integrated, supportable analysis system, running in SAS, to facilitate linking to other SAS-based enterprise software.

4 A universal standard? The Common European Framework of Reference

4.1 Origins of the CEFR

The Common European Framework of Reference (CEFR) (Council of Europe 2001) has become the most important and influential multilingual framework, not only in Europe but worldwide. The CEFR is not only a book, now translated into many languages, which provides a comprehensive discussion of learning, teaching and assessment, and which is complemented by a range of further guides and resources. It also constitutes the focus of a large amount of work, and continues to exert a considerable impact on all aspects of language education in Europe, and beyond.

The CEFR is the result of developments in language education that date back to the 1970s and beyond, and its publication in 2001 was the direct outcome of discussions, meetings and consultation processes which had taken place over the previous 10 years. As outlined by Cambridge ESOL (2011), the development of the CEFR coincided with fundamental changes in language teaching, with the move away from the grammar-translation method to the functional/notional approach and the communicative approach. The CEFR reflects these later approaches.

4.2 The CEFR and Cambridge English

Thus the origins of the CEFR go back a long way, and its development intersects in a number of ways with that of the Cambridge English exam suite, and with other scaling and framework projects that Cambridge ESOL and partners in ALTE embarked on in the 1990s. Looking backwards one might see the CEFR as the end point towards which all of these projects were inevitably directed, although this was not, of course, evident at the time.

Taylor and Jones (2006) discuss the link between Cambridge English exams and the CEFR from a number of perspectives: historical, conceptual, empirical and evolutionary. Historically, they explain, the origins of the CEFR date back to the early 1970s when the Council of Europe sponsored work to develop the Waystage and Threshold levels: specified learning objectives which reflected achievable and meaningful levels of

Table 4.1 Summary of the development of the CEFR (from Cambridge ESOL 2011:8)

1960s and 1970s: Emergence of the functional/ notional approach	The Council of Europe's Modern Languages projects start in the 1960s and (following 1971 intergovernmental Symposium in Rüschlikon) include a European unit/credit scheme for adult education. It is in the context of this project that the concept of a 'threshold' level first arises (Bung 1973).
	Publication of the Threshold Level (now Level B1 of the CEFR) (van Ek 1975) and the Waystage level (van Ek and Alexander 1977) (now Level A2 of the CEFR).
	Publication of *Un Niveau-seuil* (Coste 1976), the French version of the Threshold model.
	1977 Ludwigshafen Symposium: David Wilkins speaks of a possible set of seven 'Council of Europe Levels' to be used as part of the European unit/credit scheme.
1980s: The communicative approach	Communicative approach becomes established. Attitudes to language learning and assessment begin to change. Greater emphasis placed on productive skills and innovative assessment models. The concept of levels is extended in practice.
1990s: The development of the Framework and a period of convergence	1991 Rüschlikon intergovernmental Symposium 'Transparency and Coherence in Language Learning in Europe', outcome of which was the setting up of an authoring group and an international working party. Authoring group comprises head of the Language Policy Division, Joe Shiels plus John Trim, Brian North and Daniel Coste. Key aims are:
	To establish a useful tool for communication that will enable practitioners in many diverse contexts to talk about objectives and language levels in a more coherent way.
	To encourage practitioners to reflect on their current practice in the setting of objectives and in tracking the progress of learners with a view to improving language teaching and assessment across the continent.
	Publication of revised and extended Waystage and Threshold and first publication of the Vantage Level which sits above these at Level B2 of the CEFR (van Ek and Trim, 1990a/1998a, 1990b/1998b, 2001).
	Pre-Waystage level called Breakthrough developed by John Trim.
2000s: Using the Framework and the emergence of the toolkit	2001 final draft published simultaneously in English and French (Cambridge University Press and Didier).
	2001 European Languages Portfolio launched.
	CEFR translated into at least 37 languages.
	'CEFR toolkit' developed including manuals, reference supplements, content analysis grids and illustrative samples of writing and speaking. Council of Europe encourages development of Reference Level Descriptions for specific languages.

language competence at a relatively low proficiency level. In 1990 the revised Waystage and Threshold specifications (which had been partly sponsored by Cambridge ESOL) formed the basis for specifying the new KET and updated PET exams. Vantage level, published in 1999 (van Ek and Trim 2001), reflected input from ALTE, and took account of Cambridge's FCE exam. In this way the concept of a framework of reference levels began to emerge from interaction between systems of exams, new and existing, and the Council of Europe's work on learning objectives.

Conceptually, the emerging framework formalised levels which were already familiar to English language teaching professionals and learners. As North (1996:8) states:

> The CEFR levels did not suddenly appear from nowhere. They have emerged in a gradual, collective recognition of what the late Peter Hargreaves (Cambridge ESOL) described during the 1991 Rüschlikon Symposium as 'natural levels' in the sense of useful curriculum and examination levels. . . . The first time all these concepts were described as a possible set of 'Council of Europe levels' was in a presentation by David Wilkins (author of 'The Functional Approach') at the 1977 Ludwigshafen Symposium . . .'

The development of the Cambridge ESOL exam suite in the 1990s extended the tested proficiency range downwards, adding the PET and KET exams at what would become CEFR Levels A2 and B1. This reflected a fundamental re-orientation of the function of assessment, no longer to accredit one or two advanced levels of proficiency, but rather to provide a supportive framework of learning targets covering the entire range from near-beginner upwards. The addition of the CAE exam as a stepping-stone between FCE and CPE addressed the same need: to support learning across the whole range of levels. This view of assessment as a 'learning ladder' was wholly consistent with and complementary to the aims of the CEFR.

In terms of the empirical relationship, Taylor and Jones (2006) could point to a range of studies relevant to linking the CEFR levels to Cambridge English exam levels, referring to the ALTE Can Do Project (Jones 2001, 2002), the item-banking methodology underpinning test development and validation (Weir and Milanovic (Eds) 2003, Weir et al 2013), the origin of PET and KET test specifications in Threshold and Waystage levels, the provision via the Council of Europe of exemplar benchmarking materials, and research studies such as the Common Scale for Writing Project (Hawkey and Barker 2004). Cambridge ESOL had also provided exemplar speaking performances (Galaczi and Khalifa 2009).

The Common European Framework project, conducted between 1993 and 1996 by the Council of Europe, brought together the separately defined Waystage, Threshold and Vantage learning objectives, along with other

levels associated with existing exams, to construct a common framework in the European context. It is worth recalling that the CEFR is a framework in two different senses. Its original purpose was to provide a conceptual framework: a common meta-language to talk about learning objectives and teaching methodologies, based on an action-oriented conception of language as a tool for communication, which would enable and encourage practitioners to reflect on and share their practice. Between the first draft of the CEFR (Council of Europe 1998) and its final 2001 form the descriptive system of levels was more fully elaborated and given greater prominence. The CEFR as a conceptual framework emphasises the many ways in which contexts of learning may differ, while its framework of levels, illustrated through sets of 'Can Do' descriptors, amounts to a claim that despite the differences between contexts of learning they can be usefully compared in terms of a notion of functional language proficiency – the 'action-oriented' approach.

Valuable insights into the original conception and development of the CEFR are given in a series of interviews with John Trim, the leader of the project (Saville 2011).

4.3 The CEFR as a measurement construct

From the above account it is clear that there is an empirical underpinning to the CEFR levels. They have their origin in real and familiar cohorts of language learners: an understanding of progression shared between teachers, language school course designers, publishers and language assessment experts. There is also a statistical dimension: a much-cited feature of the CEFR Can Do descriptor scales is that they are derived from a number of studies based on empirical data (North 1996, 2000). To what extent then can the CEFR be seen as a measurement construct, comparable for example to the Cambridge ESOL Common Scale? The reader might wish to review the presentation of the Common Scale in Section 3.3.4 above before proceeding.

Jones (2005a) considered this question in the context of the Asset Languages development (see Section 5.3.3), in an effort to identify a rational basis for defining a framework of proficiency levels. This is particularly an issue where levels serve two functions:

- as learning objectives: a series of steps forming an accessible 'learning ladder', each worthy of accreditation
- as distinct levels of proficiency, representing substantively different degrees of communicative language ability.

As noted in Section 4.2, Brian North, the author of the CEFR's Can Do scales confirms that 'the CEFR levels did not suddenly appear from nowhere'.

In defining the scales he explicitly sought to find a best fit to an existing notion of levels. This was done:

- by referring to logit values in an attempt to create a scale of more or less equal intervals
- by looking for patterns and clusters, and apparently natural gaps on the vertical scale of descriptors which might indicate 'thresholds' between levels
- by comparing such patterns to the intentions of the authors of the source scales from which descriptors had been taken or edited, and to the posited conventional or 'natural levels' (North 2000:272).

Thus both the Cambridge ESOL Common Scale and the CEFR's Can Do scales refer their levels to familiar cohorts of language learners. However, the two approaches to scale construction are quite different, the former being based on vertical linking of the objectively marked components of a suite of exams, and the latter on the judgement of groups of teachers as to the ability of their students relative to Can Do statements. Both use IRT, and it will be useful to consider in a little more detail the nature of an IRT scale.

The measurement unit of an IRT scale is called a *logit*. A distance of one logit between a task and a person represents a specific probability of the person responding correctly to the task – that is, makes a specific prediction about performance. Thus a difference of one logit entails a particular *observable difference* in performance, which we might expect to be constant across the ability continuum. Would we then expect the 'natural levels' to be separated by roughly the same observable difference? As shown above, this is how North's original 9-level scale was defined. However, this 9-level scale was adapted for the 6-level CEFR by combining levels, so that A2, B1 and B2 contain 'plus levels' and are about twice as wide in logit terms as the other levels.

The Cambridge ESOL Common Scale covered the range from KET to CPE (A2 to C2), and Jones (2005a) used data from the Young Learner tests to add a provisional A1 level. The lower Cambridge levels are significantly wider than the higher: the logit interval defined below A2 is as wide as that between the B1 and C2 thresholds.

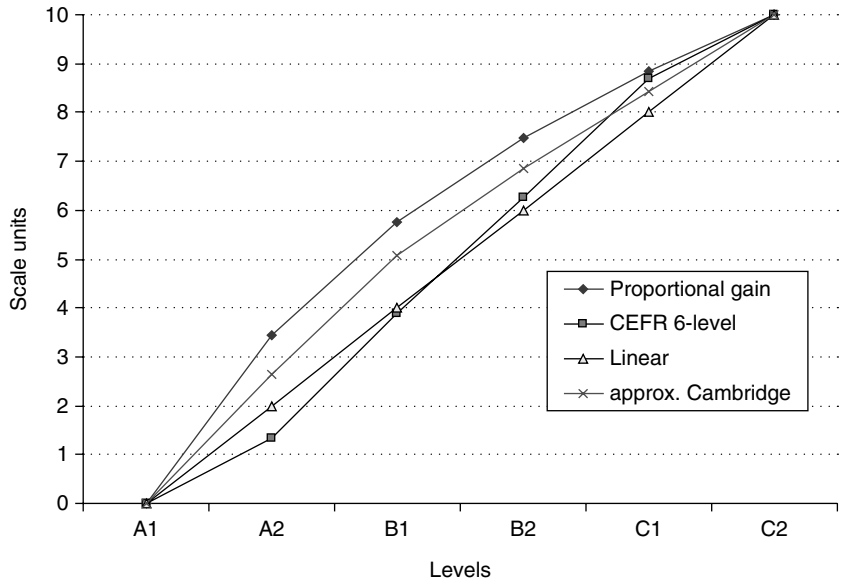
There is agreement that higher levels generally take progressively more effort or 'learning hours' than lower ones. Thus a given learning effort makes a bigger observable difference at lower levels than at higher. Consistent with our experience, this suggests that what is observable, and hence measurable, in language proficiency is *proportional gain*. The difference in the Cambridge logit bandwidths between lowest (Young Learner) and highest (CPE) levels clearly reflects this (see too Figure 3.3 on page 73).

Figure 4.1 illustrates four scales defined according to different principles, anchored at the A1 and C2 thresholds (against an arbitrary 10-unit scale).

The linear scale is the simplest: all six thresholds are equidistant. This implements the 'constant observable difference' principle. North's PhD study (1996) defined a 9-level framework on this basis.

The highest curve in Figure 4.1 implements the 'proportional gain' model, a hypothetical framework where the levels are separated by a constant quantity of learning effort.

Figure 4.1 Different approaches to defining scales (from Jones 2005a)



The Cambridge Common Scale can be seen to represent a compromise between the 'proportional gain' and the linear 'constant observable difference' models. This is pleasing, because it suggests that the system of exam levels has evolved over time to serve as well as possible the two functions of a framework identified above: to provide an accessible 'learning ladder', and to distinguish substantively different levels of communicative language ability. If the notion of 'natural levels' as intended by Hargreaves makes any sense then it is surely in terms such as these.

To return to Figure 4.1, it is clear that the CEFR scale does not fit very well when presented as a 6-level system. Restoring the 'plus levels' and plotting it as nine equidistant levels would aid comparison. But the important point that emerges from this discussion is that as measurement scales the Cambridge Common Scale and the CEFR's calibrated Can Do scales are not directly comparable, having been developed in completely different ways. This is worth pointing out, because researchers who wish to treat the CEFR

as a measurement scale must be careful to link like with like. The attempt to link the ALTE Can Do project to the CEFR (Section 3.4) was valid to the extent that the scales were constructed in similar ways; but to link calibrated objective test data directly to the CEFR's Can Do scales would not be meaningful.

4.4 The contribution of assessment to the CEFR

A survey on the use of the CEFR (Council of Europe 2006) found that although it was well known, widely consulted and positively rated, there were a number of issues with its adoption in schools. Though quite well known to teachers, the text was found to be difficult and inaccessible. Many were only familiar with the descriptor scales. Approval of its philosophical approach was matched by concerns that this was being subverted by simplistic applications to testing. Users paid too little attention to the differentiation within the scales (profiling), with the result that it was used too prescriptively. A need was expressed to bridge a gap between teachers and testers. Generally, the survey suggested that the CEFR had exerted a disproportionate influence on assessment, going beyond the capacity of teachers to keep up.

And yet at the same time a number of assessment specialists were reporting problems with using the CEFR as a basis for test construction. Many of the issues raised related to the categories and content of the illustrative descriptor scales. Some criticisms of the descriptors relate to perceived inconsistencies in terminology and the vagueness of terms used, e.g. 'short', 'familiar' etc. (Alderson, Figueras, Kuijper, Nold, Takala and Tardieu 2006: 9–13). Perhaps more serious criticisms related to the theoretical status of the illustrative descriptors. The descriptors are designed to communicate with and be useful to a range of audiences including teachers, learners, testers and other users of tests. These functional, user-oriented Can Do descriptors are then organised in a hierarchy which provides 'an operational definition of knowing a language' (Shohamy 1996:145). They reflect the concern of the Council of Europe in the 1970s to address learners' concrete communicative needs: what 'learners should most usefully be able to communicate in the foreign language' (Wilkins 1976:19).

This needs-analysis, outcomes-targeted feel to the descriptors has been criticised for neglecting cognition, and theoretical models of language proficiency based on second language acquisition research. Fulcher (2008), Weir (2005b) and Alderson et al (2006:3) discuss the unclear relationship between language testing rating scales and second language acquisition theories, particularly commenting on their failure to offer a view of how language develops across these proficiency levels in terms of cognitive processing, and alleging that the scales present 'a taxonomy of behaviour rather than a development' in reading abilities (Alderson et al 2006:3). In part this perception

may be due to the very success of the illustrative descriptors, which as the above-mentioned survey demonstrates, remain the best known component of the CEFR. Intentionally the descriptors focus on observable outcomes of learners' competence – that is, on measurable Can Do terms – rather than focusing on the processes underlying performance.

In developing the descriptive scales for the CEFR, North (1996, 2000) selected Can Do statements from a wide range of existing sources, then amended them through workshops with teachers. They were then used by teachers to rate a total of 2,500 students, providing data for a Rasch analysis to calibrate the statements. To this extent, the scales constructed from the Can Do descriptors in the CEFR can be seen as empirical constructions rather than armchair conjectures as to the nature of progression. However, it has been suggested (Alderson 2007, Hulstijn 2007), that because of the procedures adopted the descriptors reflect teacher perceptions of how learners progress rather than empirical observation of the learners themselves. Shohamy (1996:146) argues that the scales only give the 'illusion that they were based on something scientific, on theory'.

These criticisms are perhaps insufficiently sympathetic to issues which North (1996, 2000:2) is well aware of. The CEFR scaling project:

> involved attempts to relate statements about learner achievement in terms of communicative language proficiency to theoretical models of communicative language competence. The relationship between competence and proficiency is complex, and is related to the distinction between theoretical models and operational models. Scales of language proficiency can be seen as a branch of Behavioural Scaling, the classic methodology for which was developed by Smith and Kendall (1963).

Thus North presents the CEFR descriptor scales as an exercise in behavioural scaling, with a consequent emphasis on observable performance, but as an indicator of an underlying competence.

The CEFR's authors present the descriptor scales as context free but context relevant (Council of Europe 2001:21): that is, applicable to all contexts but specific enough to be interpretable for practical purposes. The issue perhaps is not so much with the scales' treatment of competence, but with the inevitable tension between the generality of the levels and the context-specificity of the descriptors. The empirical research underlying the development and calibration of the scales gives them substance, but reflects the specific contexts in which the research was conducted, and thus limits their generalisation to different contexts.

Offered as illustrations of CEFR levels, the scales seem to function more as *definitions*, for example in the way they are selected from to compile the global Common Reference Levels tables (Table 1: Global scales, Table 2: Self-assessment grid), or in Section 3.6: Content coherence in Common

Reference Levels, where each level is epitomised by identifying the salient features of selected descriptors (Council of Europe 2001:33). These compilations seem to an extent to undermine the stress placed elsewhere in the CEFR on using the scales selectively, to profile contexts and learners – that is, to point up what makes them *different*.

Of the three purposes identified in the title of the CEFR – learning, teaching and assessment – it is assessment which has most need of theory. Theory is the basis for defining testable constructs explicitly. North (1996, 2000:39) is not unaware of how theory and description need to relate within a common framework scale: 'Put briefly, it should be possible to relate the development of the scale to both descriptive theory . . . and to measurement theory . . . It should relate to a competence model, yet it should develop a metalanguage and descriptor style which is accessible and relevant to practitioners.'

Initially the CEFR's authors could point out in their defence that the fields of applied linguistic study which might be expected to provide theoretical underpinning for a language proficiency framework had not yet done so. While recognising this they defended the CEFR's practical orientation. North (1995:447), states: 'although the available theory and research is inadequate to provide a basis for [the Can Do scale], and whilst relating to theory, it must remain user-friendly – accessible to practitioners.'

Similarly, John Trim (Saville 2005:284–285), a key figure in the development of specifications of levels which formed the basis of the CEFR, states:

> I don't see, at the moment, any sign that much of the work which is currently being done in theoretical linguistics is in fact leading to our getting greater control over the problems in the individual and in society; [. . .] I become rather impatient with theoretical criticism which doesn't actually provide anything to put in its place.

As described in Section 2.1, work done at Cambridge ESOL applying Weir's socio-cognitive model (2005a) to the definition of constructs for reading, listening, speaking and writing has filled out the socio-cognitive model presented in the text of the CEFR itself and provided significant practical assistance to other developers of assessments. Writing in 2014 it is clear that assessment focused scholarship has already contributed significantly to carrying forward the CEFR enterprise, and will continue to do so, strengthening both the theoretical framework and the practical guidance available to users who are implementing communicatively oriented language educational reforms.

Practical guidance is needed, because adopting the goals of the CEFR in language education necessarily involves major changes at a number of levels, from the development of language policy through to the micro level of classroom practice and pedagogy. The CEFR is a point of reference, and a key principle for using it is that each context of learning must be related to the

CEFR on its own terms – it is not a case of applying the CEFR prescriptively to every context. The CEFR is not presented as a bible, and it is ironic that its severest critics often seem to perceive it as one. It is an area of ongoing work. Thus Trim states:

> What [the CEFR] 'Can Do' is stand as a central point of reference, itself always open to amendment and further development, in an interactive and international system of cooperating institutions . . . whose cumulative experience and expertise produces a solid structure of knowledge, understanding and practice shared by all (Trim in Green 2012:xli).

The CEFR states at the outset: 'We have not set out to tell people what to do or how to do it' (Council of Europe 2001:1). However, we can imagine many readers feeling disappointment at encountering this often-cited statement, because explicit instruction is just what they would have appreciated. And certainly, the CEFR is not neutral with respect to how it views the purpose and value of language education. A conference held by the Council of Europe in 2007, prompted by concerns that the CEFR text was in many contexts being misunderstood or misused, was found to have served usefully to 'clarify the status and the purpose of the CEFR – as a descriptive rather than a standard-setting document it allows all users to analyse their own situation and to make the choices which they deem most appropriate to their circumstances, while adhering to certain key values' (Council of Europe 2007:7). Byram and Parmenter (Eds) (2012:5) cite this extract to stress the 'value-bearing' nature of the CEFR text.

The need to provide support in understanding and using the CEFR has been recognised by Cambridge English. While consistently stressing the open and extensible nature of the CEFR, Cambridge has offered practical guidelines in the form of *Using the CEFR: Principles of Good Practice* (Cambridge ESOL 2011).

4.5 Conclusions: Taking the CEFR forward

What the CEFR's authors intended is that it should be possible to build a broad community of practice concerning goals, methods and outcomes of language study, across contexts which may differ in terms of the learner groups catered for, curricular objectives, educational traditions and so on. Let us consider the dynamics of such a process. To an extent it is one of convergence around a central notion: the importance of learning languages for communication. Despite the authors' claim that they are not telling people what to do, that much is clearly stated in the CEFR, which refers to recommendations of the Council for Cultural Co-operation of the Council of Europe to member governments:

> To ensure, as far as possible, that all sections of their populations have access to effective means of acquiring a knowledge of the languages of other member states (or of other communities within their own country) as well as the skills in the use of those languages that will enable them to satisfy their communicative needs and in particular:
> 1.1 To deal with the business of everyday life in another country, and to help foreigners staying in their own country to do so;
> 1.2 To exchange information and ideas with young people and adults who speak a different language and to communicate their thoughts and feelings to them;
> 1.3 To achieve a wider and deeper understanding of the way of life and forms of thought of other peoples and of their cultural heritage (Council of Europe 2001:3).

The cited recommendations are also reasonably clear on telling people how to do it:

> 2. To promote, encourage and support the efforts of teachers and learners at all levels to apply in their own situation the principles of the construction of language-learning systems (as these are progressively developed within the Council of Europe 'Modern languages' programme):
> 2.1 By basing language teaching and learning on the needs, motivations, characteristics and resources of learners;
> 2.2 By defining worthwhile and realistic objectives as explicitly as possible;
> 2.3 By developing appropriate methods and materials;
> 2.4 By developing suitable forms and instruments for the evaluating of learning programmes (Council of Europe 2001:3).

The reference to 'needs, motivations, characteristics and resources of learners' is particularly interesting and betrays methodological trends influential at the time: as already noted above, the CEFR text frequently adopts the language of needs analysis.

Writing in 2014, one might assume that the above recommendations, which may have appeared fairly radical when the CEFR was new, are now universally accepted; and yet with the ESLC fresh in our memory the importance of focusing on language for communication is as clear as ever. Failure in language learning is widespread in many European countries, and can be linked, I personally believe, to failure to treat languages as communication tools. So convergence around the CEFR's main goal, even if achieved in theory, is far from being achieved in practice. There is still much to be done.

At the same time the world has moved on, so that some of the concepts visible in the text of the CEFR, such as needs analysis, may have been left behind. Trim's idea of a system 'open to amendment and further development' envisages a process not only of convergence but also of evolution

(Saville 2011). That evolution might yet take us much further. We can understand why the *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education* (Council of Europe 2010:29) might assert that 'language teaching in schools must go beyond the communication competences specified on the various levels of the CEFR'. We can envisage more complex conceptions of the goals of language education; in fact, each country already has its own more complex conception, within which the promotion of communicative competence may be only one priority. Certainly, when we set out to advise a country on how to adopt the CEFR we should be sensitive to such issues and ready with ideas on how to exploit the positive synergies between developing communicative competence and other goals of language education.

Thus the dynamic development of a broad community of practice concerning the purposes of language study may naturally lead to convergence on how to extend goals and outcomes beyond the CEFR's narrower focus on communicative language competence. That would represent a challenge for language assessment, but one consistent with the spirit of the age. The frameworks described in this volume all focus on assessment, specifically of language ability. Generally, they stand outside educational processes. As exam providers move to involve themselves more closely in learning, and we begin to define our business as that of 'language education' rather than 'proficiency testing', we will need different, broader frameworks. The concluding chapter of this volume explores such possibilities in more detail.

5 Asset Languages: A formative framework for language learning

The multilingual framework enterprise which is the subject of this chapter is certainly the largest in which Cambridge English Language Assessment has been involved, in terms of the number of languages and tests included, and in the number of external and internal collaborators on the development of the test material.

The Asset Languages scheme emerged from the National Languages Strategy undertaken in 2002, which in turn emerged from an independent inquiry into the state of language education in the UK (Nuffield Foundation 2000). As a case study of a multilingual framework which set out to promote communicative ability as the primary goal of language education it is instructive in many ways, not least when we come to consider the final failure of the National Languages Strategy, despite the boldness of its conception, to deliver the reforms which it set out to achieve.

Asset Languages is the only project to date which has involved Cambridge English Language Assessment directly in schools examinations in the UK, working in partnership with its sister organisation OCR. The project provides a rare, possibly unique instance of the use of an item-banking, IRT-supported approach to setting standards for certificated exams in the UK schools sector. The Asset Languages development presented several new challenges for multilingual framework construction:

- it was a large and complex exam system, imposing a need for practical, robust methods of framework construction
- it referred a large number of different languages to the CEFR, raising conceptual and practical issues
- it promoted criterion-referenced values within a norm-referenced assessment tradition which found them hard to deal with
- it attempted to fulfil an expressly formative function, developing the notion of levels constituting a 'learning ladder'.

This is the longest chapter in the book, and it goes into a greater level of detail than other chapters in discussing the context of the development and the practical constraints that were encountered. The justification for providing this extra level of detail is that the constraints and compromises which

are inevitable in all such complex developments are part of the story, and certainly offer instructive lessons, which are summarised in Section 5.6.

5.1 The origins of Asset Languages

5.1.1 The assessment environment in 2000

Let us pick up the story in 2000, when several forces for change, and for the introduction of a new type of qualification, began to appear.

Giving teachers a greater role in assessment was proposed as a way of lessening the burden which external assessment placed on learners and teachers alike (Department for Education and Skills 2004).

Developing the use of e-learning and e-assessment seemed to promise great benefits, not only in terms of decreased cost, but also to drive out the frailties imposed by human markers using subjective marking schemes (NESTA Futurelab 2002).

The need to increase learner motivation was recognised, with the new exam-driven curriculum criticised for stifling creativity and autonomy in learning. By the early 2000s there had already been a substantial decline in numbers of secondary school learners taking language qualifications.

At the same time, ministers and Ofsted (the Office for Standards in Education, i.e. schools inspectorate) had expressed a desire to expand provision for languages at primary level.

The need for continuity between primary and secondary schools was also evident, as pupils started secondary school with widely differing levels of language learning experience, possibly in different languages, making continuity of study difficult to organise.

5.1.2 The Nuffield Languages Inquiry

All the issues described above contributed to a general feeling that changes were needed in language learning and assessment. The outcome was an inquiry into the state of language learning, teaching and assessment in England by the Nuffield Foundation. The findings of this inquiry were to feed into the National Languages Strategy of 2002, which in turn led to the introduction of Asset Languages.

The Nuffield Languages Inquiry (Nuffield Foundation 2000) reported on the current state of language learning, teaching and qualifications in England. With its long history of involvement in language teaching the Nuffield Foundation was well positioned to establish a national inquiry with the aim of providing an independent view of the UK's future needs for capability in languages and the nation's readiness to meet them. The report was very critical, concluding that 'at the moment, by any reliable measure,

we are doing badly' (2000:5). Key conclusions of the report (2000:5–9) were that:

- English is not enough – The UK needed competence in many languages and was currently disadvantaged in the recruitment market.
- Government strategy – The Government had 'no coherent approach to languages' and a change of policy and practice was needed (2000:5).
- Motivation – Motivation was seriously lacking and although 'more pupils now learn a language to age 16 than ever before, too few leave school with an adequate level of operational competence'. The current provision 'does not motivate and too many pupils, also lacking positive messages about languages from outside the classroom, see language learning as irrelevant'.
- Qualifications – The range of current qualifications was deemed to be confusing with many finding it difficult 'to predict from qualifications what the standards being achieved actually represent' (2000:8).

The inquiry stated that 'one way or another we must give our children a better start with languages and equip them to go on learning them through life' (Nuffield Foundation 2000:5). It recommended that the 'government should arrange for a sustained campaign to promote positive attitudes towards languages, raise awareness of their potential and foster a culture where using more than one language is seen as an attainable goal for the majority in the UK'.

It also found that language learning should be seen as a life-long skill and that in particular at secondary school, provision for learning a language 'should be uprated to provide a wider range of languages, a more flexible menu to cater better for different needs, abilities and interests and more use of information technology' and that 'all pupils should leave secondary education equipped with foundation language skills and the skills for further learning in later life' (Nuffield Foundation 2000:8).

The report made specific recommendations for how this could be achieved. For example, it stated that 'the government should establish a national strategy for developing capability in languages in the UK and a system capable of supporting such a strategy' (2008:8). In order to resolve the issues with qualifications, the report recommended that a new framework be developed which should 'embrace the Council of Europe Framework and existing UK qualifications both in education and the world of employment. It should be clear, transparent and couched in terms which are intelligible to non-specialist users'. The aim of this framework would be to provide coherence, transparency and consistency to language education across sectors, providing continuity from primary school through to higher education and beyond. There should be priority in ensuring adequate provision to all across a range of languages, including the UK's community languages. It stated that qualifications should be motivational, provided as 'small steps' and be available

as modular units, thereby allowing learners to be assessed separately by skill (that is, in reading, writing, listening and speaking), thus also allowing learners to progress at different rates in each skill. Greater use should be made of information and communication technology (ICT): it was necessary to 'move the use of ICT from the margins to the centre of course planning and materials development' and to 'ensure that lack of access to appropriate ICT facilities does not limit opportunity'.

The key features of the new qualifications framework proposed by the Nuffield Foundation are summarised below. It should:

- be linked to both the CEFR and existing UK qualification frameworks
- be inclusive for primary, secondary and adult learners
- be motivational and represent small steps of positive achievement
- be expressed in functional Can Do terms, e.g. 'I can read a short story'
- be inclusive and widen the range of languages available, both taught and assessed
- be modular, i.e. reading, writing, listening and speaking skills should be assessed separately
- reward partial competences, i.e. allow for learners to progress in some skills but not others, and/or in some skills more rapidly than others
- promote self and peer-assessment.

The Nuffield Inquiry's vision recalled the small steps of the 'graded objectives' system which had operated in the 1980s (Nuffield Foundation 2000:76). The Inquiry commented that 'the UK has failed to build on the successful initiatives of the 1980s to develop a flexible step-by-step approach to rewarding learners as they take their first steps in a new language. The benefits of such an approach in these early stages remain as important as ever, both in schools, where the idea was born, and also in the world of adult learning, where the positive and unthreatening nature of these "graded objectives" schemes is appreciated.'

5.1.2.1 Feasibility study

Following the publication of the Inquiry's findings, a study was conducted into the feasibility of implementing the recommendations (Nuffield Foundation 2002).

The study concluded that (2002:2) 'A Learning Ladder for Languages (LLL hereafter) could motivate, inform and reward learners of all ages and backgrounds'. It was argued that the 'potential benefits to be derived from "step by step" approaches are as valid today as when they were recognised by the graded objectives schemes developed in the UK twenty years ago'. These benefits were highlighted as particularly important for those in the early stages of language learning.

None of the current schemes of assessment for languages were found to meet the requirements as outlined by the Inquiry. Therefore the new LLL should be implemented not only as a motivational framework for learning, but also provide some form of certification. Key aspects of the LLL as conceived by the feasibility study included (2002:2):

- a single generic LLL, expressed in functional Can Do terms
- encouragement of self-assessment in relation to the LLL, but also with the option of taking a test and receiving certification
- modularity, i.e. reading, writing, listening and speaking skills should be assessed separately
- the scheme should provide for all languages for which there is a demand.

Despite its criticisms of current schemes of assessment, the certification proposed in the feasibility study was relatively informal. The feasibility report concluded that the central responsibility lay with Government. A long-term vision was needed for the UK to have a 'national system that provides a common structure for languages at all levels' and all educational contexts (2002:3).

5.1.3 The National Languages Strategy

In July 2001, following the Nuffield Languages Inquiry and in parallel with the feasibility study being carried out, the Department for Education and Skills (DfES), now the Department for Education, set up a Languages National Steering Group with the remit to 'develop a strategy to change perceptions and raise awareness amongst young people and the wider public of language competence as a key contemporary life skill' (Department for Education and Skills 2002:2). The Group's report, *Languages for All: Languages for Life – A Strategy for England*, also known as the National Languages Strategy, was published in December 2002 (Department for Education and Skills 2002:10). The strategy had three overarching objectives (2002:6):

1. To improve teaching and learning of languages. This included an entitlement to language learning for pupils at primary level, as well as ensuring the continuation of opportunities to learn languages in secondary schools.
2. To introduce a recognition system. This was to complement existing qualification frameworks and give people credit for their language skills.
3. To increase the number of people studying languages. This was stated particularly in relation to further and higher education and was aimed at encouraging employers to play their part in supporting language learning.

In addition to these overarching objectives the strategy asserted the importance of understanding and communicating in other languages 'in our society and in the global economy' (2002:4). It made clear references to European Council policies, stating for example that the 'strategy demonstrates our commitment to making progress towards fulfilling the conclusions of the Barcelona European Council with regard to language teaching and learning in schools' (2002:5).

The strategy stressed the central importance of motivation. It 'will be geared towards both motivating individuals to learn – the push factor – while at the same time ensuring that high quality and appropriate opportunities are available – the pull factor' (2002:6).

The entitlement for primary learners was a significant aspect of the National Languages Strategy. 'If a child's talent and natural interest in languages is to flourish, early language learning opportunities need to be provided, and their aptitude needs to be tapped into at the earliest opportunity when they are most receptive' (Department for Education and Skills 2002:10). The strategy had ambitious aims for primary learners, stating for example that they 'should have access to high quality teaching and learning opportunities, making use of native speakers and e-learning. By age 11 they should have the opportunity to reach a recognised level of competence on the CEFR and for that achievement to be recognised through a national scheme.' At the same time the strategy noted that this could pose challenges for continuity of learning from primary to secondary school, pointing out that transition must be carefully managed if learner achievement were to be recognised and learner motivation and enthusiasm sustained.

The primary entitlement promised in the strategy was well received. However, at the same time the Government controversially removed the statutory requirement for learners to study a modern foreign language at Key Stage 4 (upper secondary school learners, of 14–19 years of age). Far from promoting language learning, many argued, this would deter learners from studying languages at secondary school, and thus undermine the whole basis of the strategy's aspirations to raise motivation for language learning and teaching. Subsequent developments showed this fear to be well founded.

Asset Languages originates in the second of the National Strategy's three overarching objectives: the national 'recognition system' for languages. It is important to understand the features of the recognition system which impacted not only on the design of Asset Languages, but also on its adoption by schools as an alternative or addition to existing assessments.

The recognition scheme was voluntary – there was no statutory requirement for schools, teachers or learners to participate in the scheme. The DfES described the Scheme at its launch as follows:

> . . . designed to endorse competence in foreign language learning, it will allow learners to progress in one or more of the 4 skills (listening,

> speaking, reading, writing) in one or more languages and also offers the opportunity for people to assess their own levels of language competence. Each stage is externally assessed, the 'Can Do' statements within each stage can be used for formative assessment and can be endorsed by the teacher/tutor' (Department for Education and Skills 2004).

Thus the scheme had two distinct elements. The first was the Languages Ladder, a framework document owned by the Department for Children, Schools and Families (DCSF) as the DfES was renamed in 2007. It consisted of four sets of Can Do statements (for reading, writing, listening and speaking) against which students and teachers could informally assess progression in learning over 17 grades. Table 5.1 illustrates how a Languages Ladder stage is articulated. For further illustrations of Languages Ladder stages see Appendix C.

Table 5.1 A Languages Ladder stage (Listening, Breakthrough – Grades 1–3)

Breakthrough	Grade 1	I can understand a few familiar spoken words and phrases.
	Grade 2	I can understand a range of familiar spoken phrases.
	Grade 3	I can understand the main points from a short spoken passage made up of familiar language.

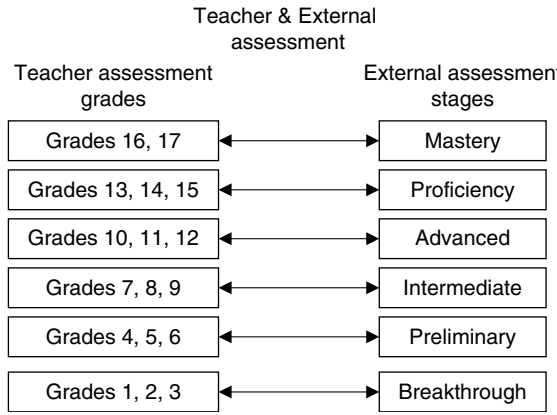
ON COMPLETING THIS STAGE:
You should be able to understand a basic range of everyday expressions relating to personal details and needs. You may need to listen several times to get the information you need, depending on how fast and clearly the speaker talks. You should have some understanding of a few simple grammatical structures and sentence patterns. You should be familiar with the sound system of the language. You should be aware of how to address people both formally and informally as appropriate.

The Languages Ladder is made up of six *stages*: Breakthrough, Preliminary, Intermediate, Advanced, Proficiency and Mastery. Each of the first four stages – Breakthrough to Proficiency – is made up of three smaller 'steps' or *grades*. Mastery is made up of two grades. Each stage and skill is described by Can Do statements at each of three grades (where grade 3 is the highest). A summary paragraph at the end of each stage also outlines what learners could be expected to do having achieved this stage.

The second element of the scheme was Asset Languages, the formal accreditation system providing tests linked to the Languages Ladder. Figure 5.1 illustrates the relationship between the Asset Languages external assessment stages and the Teacher Assessment grades.

Many features of the new recognition scheme directly reflected recommendations of the Nuffield Inquiry. The vision was of a scheme which (Nuffield Foundation 2002:7):

Figure 5.1 Relation of teacher assessed grades and external assessment stages



- can recognise achievement across educational contexts from primary to further and adult education
- measures achievement in terms of listening, speaking, reading and writing as small steps from beginner through to mastery in functional Can Do statements
- recognises a range of languages including community languages important to the UK, including an individual's proficiency in their mother tongue (not including English)
- complements existing qualifications frameworks including the CEFR
- improves the accessibility of language learning and 'brings achievement within the reach of more people', and
- raises the profile and status of language learning and achievement.

But there were also differences. The Nuffield Inquiry had recommended a system based on informal assessments such as self and peer-assessment while also allowing learners the 'option of taking a test for those who wish to receive certification' (Nuffield Foundation 2002:3). However, in defining the recognition system the Government opted for a formal assessment system, albeit one with two strands: a more formative strand based on teacher assessment, and a more summative strand based on standardised tests and having formal value (Jones 2007). In this way the Government aimed to achieve a more positive relationship between formative and summative dimensions of assessment, enabling summative tests to be a 'positive part of the learning process' (Black 2004:17). Undoubtedly, giving the assessment system formal value within the National Qualifications Framework (NQF) was critical to its adoption, given the basis on which schools performance tables are constructed.

The National Languages Strategy was well funded by the Government. The development, management and roll-out of the voluntary recognition system was the object of an invitation to tender made to national awarding bodies in 2003.

5.2 Designing the Asset Languages framework

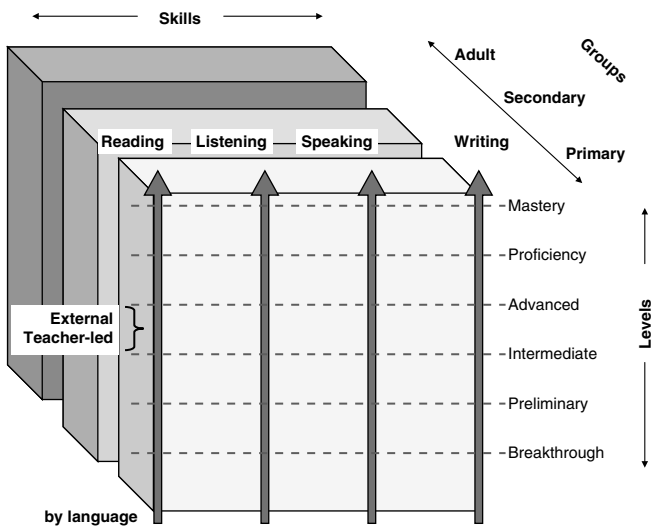
5.2.1 Key features, principles and values

As already described (in Section 5.1.2), the new recognition scheme for which Asset Languages was to provide formal assessment took some key features directly from the Nuffield Inquiry and the subsequent feasibility study. Underlying these features and requirements is the notion of a complex, multidimensional framework, capable of encompassing different levels of proficiency, different skills, languages and groups of learners, as well as different types of assessment (external and teacher-led).

Figure 5.2 gives a schematic view of the major dimensions of the Asset Languages Framework. For each fully implemented language, test materials would be constructed at each of the levels of the Languages Ladder, in each of the skills, and for each of the three target groups (adult, secondary, primary). Each set of test materials comprises external and teacher-led parts.

Practical implementation adds further complexity: there might be both CB and PB versions; the number of different test forms required for each

Figure 5.2 The major dimensions of the Asset Languages framework



language, skill and level must accommodate the requirements of multiple test sessions; there must be some linking between groups to enable comparison, and so on. Constructing this framework as a set of measurement dimensions also implies much work to be done training item writers and markers, in order to standardise understanding of levels across languages.

Later sections offer a more detailed account of the practical process of constructing this complex framework. In this section I shall consider its conceptual basis, particularly as this indicates some fundamental differences between Asset Languages and the existing assessments to which the scheme was to provide a voluntary alternative.

The notion of small steps implies that assessment should provide accessible targets throughout a learner's career rather than one big hurdle at the end of it. It offers the possibility of testing students at each level when they are ready, and not when they reach a certain age. It implies a continuous scale of language proficiency extending from nearly nothing to an advanced level.

The appeal to the CEFR makes clear that this scale is defined in terms of functional language proficiency, with criterion levels that have well-understood meanings. This in turn implies that a level achieved in French has the same meaning (in functional terms) as a level achieved in German or Urdu.

The prominent place of community languages in the scheme, taken together with the by-skill approach, implies that students can receive accreditation for language skills which they possess by virtue of birth, rather than any formal study of the language (in fact, the name Asset Languages was meant to underline that languages should be considered an asset, however acquired).

Underlying all the above features of the Asset Languages framework is the fundamental concept of a proficiency scale. At the outset (Section 2.1.1 above) we reviewed the assumptions underlying the notion of proficiency. Learners are predicted to follow a roughly similar path through a number of levels. The task of assessment is to locate the learner's current level, so that teachers and students are given a clear orientation as to the ground covered and the distance to the next goal. As progression is roughly the same for everybody, proficiency oriented assessments, in theory at least, need not be closely associated with any particular programme of study.

Conceptually this distinguishes proficiency tests from achievement tests, which measure the extent to which a student has mastered a particular set of objectives, for example as specified in a syllabus. In practice, proficiency and achievement tests may seek to measure similar language skills. Proficiency tests do not necessarily provide more meaningful or useful information than achievement tests. Both approaches must find ways of accommodating the fact that language learning and language testing are to an extent context-dependent, so that for example tests developed for adult learners will necessarily differ in content from tests developed for younger children – an explicit issue for the Asset Languages framework to accommodate.

Using the CEFR or the Languages Ladder to compare proficiency levels across languages provides a simple statement of value: in proficiency terms an Asset Advanced grade in French is of equal value to the same grade in German. Speaking of value, we begin to see that the values underlying the vision of the Nuffield Inquiry and embodied in the Asset Languages framework differed in some fundamental aspects from those of existing assessment regimes. The differences are summarised in Table 5.2 and will be further discussed below.

Table 5.2 A comparison of general qualifications and Asset Languages

General qualifications (existing qualifications)	Asset Languages (an alternative framework)
Achievement is tested in relation to a syllabus.	Proficiency is tested in relation to a construct, that is a theoretical model.
Standards are defined in terms of expected levels of achievement – designed to be broadly comparable across subjects.	Standards and comparability are related to criterion levels of language proficiency, in Can Do terms.
Summative – exams come at the end of an extended period of learning.	Summative, with formative aspects, in shorter more regular cycles. Smaller learning steps, each positively accredited.
Test all students in a given age cohort.	Possible to test when ready, at an appropriate level for the individual.
All skills are tested at the same level.	Modular testing of skills at different levels is possible.
Patchy provision for community languages.	A wide range of languages including community languages – accreditation of mother tongue language skills.
Not aimed at lifelong learning.	Serves lifelong learning by design.

5.2.1.1 Standards defined in terms of functional proficiency

Given that the new framework and voluntary recognition system was required to complement both the CEFR and existing qualifications (Section 5.1.3 above), it is important to consider the different value that these two types of frameworks place on achievement and proficiency. A virtue of proficiency frameworks such as the CEFR is that they provide a common basis and discourse for understanding and talking about levels. They achieve this by placing value on what a learner can do with their language skills regardless of their age, or expenditure of learning effort. As Wiliam (2007:241) discusses, in relation to schools subjects generally, '[w]hile the idea of age-independent levels of achievement may be unfamiliar, there is substantial research evidence that in many subjects, achievement does appear to be relatively independent of age'. Within this type of functional proficiency framework, levels are informative to the users, such as teachers and learners, and scores 'from tests based on this scale would thus be comparable across different languages and contexts' (Bachman and Clark in Bachman 1990:5–6).

This contrasts with existing 'high stakes' qualifications such as GCSEs (exams at the end of lower secondary education in the United Kingdom) which test learners' ability after taking a set course for a specified number of guided learning hours, and where comparability across subjects, including languages, tends to be conceived in terms of an expected level of performance, defined in terms of set objectives or attainment targets (see for example Qualifications and Curriculum Authority 2008:4), assuming equal ability to learn, equal motivation, quality of teaching and input, and so on.

For different languages it is of course not the case that a given amount of learning will result in the same level of performance, in Can Do terms. The progress that the average learner makes in a language in a given time period depends predominantly on two key aspects: the nature of the language being learned, and the learner's starting point. For example, in a given period of time, a learner taking a language that is perceived as particularly difficult and dissimilar to the learner's first language is likely to reach a lower level of proficiency than someone taking a language that is perceived as easier and more similar to the learner's L1.

Similarly, a community language learner who starts formal school lessons with some prior knowledge of the language will demonstrate a higher level of proficiency than learners of the more widely accredited languages with no background in the language. High-stakes qualifications such as GCSEs take such expectations into account in the way they are graded, and are therefore not necessarily comparable in terms of what learners are capable of actually doing with their language skills. Some exam boards offer two versions of certain languages: for native and non-native speakers, naturally requiring different standards in functional terms. However, in both cases the qualifications learners receive are intended to have the same currency in terms of what their qualification enables them to do, for example, gain entry into university (Jones 2007). Clearly, for the new recognition system to function as intended by the Nuffield Inquiry and the National Languages Strategy, learners, teachers and decision makers would have to learn to think about learning in a new way, valuing the outcomes-related virtues of a functional proficiency framework.

Given these differences between Asset Languages and existing school language qualifications, it is not surprising that in the English education system comparability across assessments and levels is not always addressed in a rigorous way. Coleman (1996:7) finds, for example, that in a higher education context labels such as *first year level* or *foreign language to degree level* 'are meaningless' due to significant discrepancies in foreign language proficiency across British universities.

5.2.1.2 By-skill testing accommodates different profiles of language skill

The Asset Languages system permits learners to be tested in each of the four skills at a different level, accommodating the fact that language learners

may progress at different rates in each skill. This is particularly important in the case of non-Latin script or community languages where learners may progress more rapidly in listening and speaking than in reading and writing (Little 2006:169). This approach differs from GCSE and A Level, the two key exams at the end of lower and higher secondary education respectively, where although each skill is tested, it is not separately reported and the results are aggregated into a single grade for the language. As a consequence, a strength in one skill could compensate for a weakness in another.

Separate skills testing is not without its critics, as in real life the skills tend to be used together rather than in isolation, particularly in the case of listening and speaking (Mitchell 2003:3). It is also argued that literacy is more complex than can be captured by reducing performance 'to one of the traditional four language skills' (Hudson 2005:206). However, the testing of discrete skills is still widely used as it allows language testers to 'design exercises and test tasks with a clear objective in mind' (Weir et al 2000:19) and facilitates the reporting of ability and comparisons across learners more narrowly than is possible with integrated skills testing. In fact, the Asset approach can readily be justified as favouring authentic and natural language learning: for any native speaker, acquiring the skills of literacy is naturally scaffolded by an existing well-developed competence in the spoken language. Wherever such profiles of skill exist, the stronger skill will naturally scaffold the weaker ones.

5.2.1.3 A graded set of test levels provides individualised, appropriate learning objectives

A key aim of the multi-level Asset system was that students could be tested when ready at an appropriate level, thus favouring a positive outcome and good motivation to master the next level. Where an exam's candidature is determined purely by age, as with GCSEs, the range of observed ability covers every possible level of achievement. Such tests certainly cannot provide meaningful interpretations of performance at every level encountered. In consequence results in such tests are, necessarily, almost entirely norm-referenced, interpretable only in terms of a student's relative ranking in the cohort.

The Asset Languages approach could be expected to have a positive, formative effect by offering a closer match to learners' needs. This should be a positive feature for all learners, but is particularly so for community language learners, for whom under-assessment has been associated with the lowering of expectations and motivation (Datta 2000). The assessments provide more 'formal recognition' of their achievements whilst also helping to promote these languages more widely by addressing the fact that for many people 'community languages are not "foreign" and are part of their everyday experience' (Dearing and King 2006:30). Geach (1996:150)

identifies the monitoring of progress for community language learners as an 'enduring need' while Sneddon (2000:274) states that pre-GCSE there is 'no reliable measure' available to 'assess children's literacy skills in Urdu: the regular tests the children took were primarily designed to assess religious knowledge'.

5.2.1.4 Explicit linking of external and teacher-led assessment

The scheme set out to implement a model of Assessment for Learning, which was defined by the Assessment Reform Group (2002) as 'the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.'

As noted above, the system of graded objectives, with summative assessments available over a range of levels, was itself intended to have positive formative impact. The Teacher Assessment strand of Asset Languages set out to provide an additional formative structure of small learning objectives that provided scaffolding for the external assessment. These short-term goals were designed to be motivational and act as anchor points within a centre's curriculum framework. The development of the Asset Languages Teacher Assessment strand will be presented in Section 5.5.

5.2.1.5 Key features and validation

Asset Languages provides an unusual instance of an assessment system developed wholly to meet the objectives of a reforming educational mission, defined within a national strategy for languages. Its design and development reflect these objectives, which should also lie at the centre of any final evaluation of its success.

Let us refer back to Weir's (2005a) model of validity (presented in Section 2.1) to identify those aspects of design and use which are relevant to evaluating the validity, or fitness for purpose, of the Asset Languages framework:

- Tests should be designed to take into account relevant *learner characteristics*. The potential test population for Asset Languages is very heterogeneous: learners may be of any age, and may have learned the tested language in very different ways, either as a foreign language or one used at home; that is, more or less formally. We have noted above some ways in which the framework accommodated such differences.
- The *content* of tests should be appropriate to the different learner groups, in terms of relevance, accessibility and level of demand. Again, features of the framework design – chiefly, catering for three age groups – addressed this requirement.
- The tests should elicit performance interpretable according to a model of language proficiency – that is, provide evidence of listening, speaking,

reading or writing ability as intended by the test designers. This is *construct validity*. For Asset languages this should relate to the salient criteria for each skill and level as identified by the Languages Ladder.

- The *scores* assigned to learners' responses should be such as to locate learners at a level and grade. This implies not only a degree of reliability but a way of comparing scores on many different test versions targeted at different levels. Within the complex Asset Languages framework this requirement was addressed using an item-banking methodology based on IRT (see Section 2.2.2).
- Scores should predict future performance, in some educational or real-world context of use (*criterion-related validity*). This should be achievable to the extent that the tests demonstrate construct validity, and that results are not skewed by inappropriate preparation, such as rote learning, or teaching to the test. Asset Languages sought to achieve what Cambridge English Language Assessment was doing successfully for foreign learners of English: providing sound tests of communicative language ability which provided a beneficial learning focus for the efforts of students and their teachers.
- The interpretation and use of scores should be such as to have a positive effect on future learning (*consequential validity*). These desirable effects should be observed to the extent that the formative features of the Asset languages system (its graded levels, separate testing by skill, 'on-demand' availability and parallel summative and formative assessment strands) are successfully exploited by learners and teachers. Given the reforming programme of the National Languages Strategy consequential validity would be the single most critical test of success.

The long-term impact of Asset Languages will be taken up in Section 5.6. We have already considered the initial design of the Asset system at some length, but before moving on to review the actual development and rollout process it is important to identify some of the constraints and challenges which the project faced from the outset.

5.2.2 Context, constraints and challenges

5.2.2.1 Specifying the content of the tests

The proficiency-based approach of Asset Languages was a major distinguishing feature, but an important question was: how could teachers work with proficiency tests which set out to be 'content-free' – especially secondary school teachers used to the highly topic-driven and clearly defined vocabulary of GCSE?

With its proficiency framework, its focus on lifelong learning and openness

to learners in any educational sector, Asset Languages could not specify test content as explicitly as is the case of UK GCSE qualifications. The aspiration to be content free responded to what was widely recognised as the negative impact of teaching to tightly specified targets. As the Dearing Languages Review for the DfES (Department for Education and Skills 2007) found:

> The GCSE is the examination which drives the curriculum at Key Stage 4 and casts its mantle over the final year of Key Stage 3. It is particularly in these years that the context of the learning needs to be stimulating to pupils and to engage them in discussion, debates and writing about subjects that are of concern and interest to teenagers. Although outstanding teachers can overcome most barriers to learning, as commonly interpreted the present GCSE does not facilitate this.

While it made sense to counteract the focus on content, it was clear that if the tests were to fit in with what teachers were currently doing then existing curricula needed somehow to be taken into consideration. However, to create something that met teachers' and learners' needs and expectations without being too prescriptive and providing a course of study was a challenge. Finally it was decided to create language specifications. These specifications were tailored for each stage and language. A vocabulary list, as provided for GCSE, was felt to be too restrictive. Instead, the specifications were framed in terms of the language purposes and functional areas that learners might meet in the test.

Appendix C illustrates the approach through sample generic specifications for Breakthrough, Preliminary and Intermediate stages. The following attempt to explain proficiency testing was also included in the initial specification:

> The Asset Languages project requires a framework which links coherently across a wide range of absolute language proficiency levels, different languages, contexts of learning and modes of assessment delivery. A strong theoretical and measurement model is necessary in order to construct this framework. The following are important features of the approach.
>
> Asset Languages grades constitute a proficiency scale. That is, they reflect a view that learners acquiring language skills follow a roughly similar path through a number of levels, which can be described. The task of assessment is to locate the level that a learner is currently at. The predictable nature of progression means that proficiency-oriented assessments need not be closely associated with any particular programme of study. This distinguishes them from achievement tests. Notwithstanding this, language use and language testing are necessarily situated in specific contexts, and different language systems pose specific problems for learners. For this reason general guidance has been provided as to the

> range of topics, and hence vocabulary, in the form of functional areas and aspects of the language system that will be included in assessments (Oxford Cambridge and RSA Examinations/Cambridge ESOL:2005a).

5.2.2.2 Interpreting the Languages Ladder

A further constraint emerged in the requirement to use the Languages Ladder document as the basis for test development and interpretation. How would this impact on the validity of comparisons made within the framework – between adults and children, between different languages, and between different kinds of language learner – the traditional Modern Foreign Languages learner, and the learner for whom the language is perhaps a mother tongue? This section looks more closely at the descriptions of progression offered by the Languages Ladder, which at the time of the development was a work-in-progress document.

The Languages Ladder was the key interpretative framework for Asset Languages. It is a document developed by the DfES (Department for Education 2007), now only available as an archived document on the Department for Education website. It describes a series of levels in Can Do terms. An early draft was given to Cambridge Assessment in 2003, but this was subject to much consultation and revision before its final publication by the DfES in 2007. Throughout the development and finalisation of the Languages Ladder text the Asset Languages teams were offered the opportunity to review and offer feedback on revisions to the Can Do statements.

The text of the Languages Ladder reflected the influence of other existing documents. Within the UK, there were two sets of descriptors of language attainment already available for use at the time of publication of the Languages Ladder statements. The first, used within the National Curriculum and which secondary school teachers were required to use to report on student progress, defined eight levels of achievement at Key Stage 3 (ages 11–14) across the four skills of listening, speaking, reading and writing. Given the familiarity and importance of these levels it was natural that the Languages Ladder should be articulated in a manner closely coherent with the National Curriculum statements. Nonetheless, the National Curriculum statements had been subjected to severe criticism by applied linguists for taking insufficient account of developments and research in second language acquisition. Mitchell (2003:5) commenting on the restrictive ladder-like progression of the Can Do descriptors, states that 'there seems no reason in principle why learners even at the lowest levels should not be engaging in much longer conversational exchanges' and indeed referring from the very beginning to 'past, present and future actions and events' which comes in at Level 6 on the National Curriculum for Modern Foreign Languages.

The second available set of descriptors of language attainment were the National Language Standards, relating to languages in a vocational context.

It was natural that these too should be reflected in the text of the Languages Ladder. It was however unlikely that a satisfactory match could be found between these descriptors of language attainment and those of the CEFR, which, as noted above, was a key point of reference identified by the original advocates of the new framework.

A more insidious and serious problem was the practical necessity of comparing the Languages Ladder levels with those of existing qualifications – in particular with GCSE and A Level. This was clearly going to be important in the eyes of teachers and school principals wishing to get to grips with the Languages Ladder, and yet it was precisely the perceived shortcomings of existing language assessments that had been a major reason for creating the Languages Ladder in the first place. Existing qualifications had been found 'confusing and uninformative about the levels of competence they represented' (Nuffield Foundation 2002:8), while 'beyond 14, student attainment in languages is mainly related to examination targets, and not to performance criteria in 'Can Do' terms, except in vocational courses' (2002:9). This was why they had recommended that the new qualification framework should stress meaningful proficiency levels linked to the CEFR.

The Languages Ladder was a hybrid document in which the CEFR was accommodated alongside National Curriculum attainment descriptors. Among other things, this led to optimistic estimates of how CEFR levels should equate to Languages Ladder stages, particularly at lower levels (Table 5.3 below). To the extent that linking both to National Curriculum levels and to the CEFR presented irreconcilable conflicts, the Asset Languages developers set out to model the Asset Languages stages on the CEFR levels. This made much practical sense. For the developers a compelling advantage of the CEFR levels lay in their links to the Cambridge English exam suite, which result from the intertwined historical development of these two systems (see Section 4.2). For the Levels A2 upwards, Cambridge English exams already existed which provided useful English language models to illustrate the levels, and examples of test tasks appropriate to those levels. These provided useful templates and a basis for developing training materials that could be used to share some common notion of levels among the language specialists recruited for the many community languages. Additionally, the common measurement scale underpinning the Cambridge English levels, a product of many years' research, provided a kind of template for the Asset Languages measurement framework which was to be constructed (see Section 5.3.3.1).

The goal for Asset Languages was to ensure that its assessments could report learners' achievement on the basis of the Languages Ladder whilst also ensuring a close relationship and comparability in relation to the CEFR. As explained previously (5.2.2.1), the Languages Ladder also reflected the National Curriculum. The reference to the National Curriculum was not

ideal for a framework designed to support lifelong learning. While Scotland observed the development from a distance, the Scottish Qualifications Authority made it clear that they would not adopt any system which failed to link to the CEFR. From their perspective the Languages Ladder could be seen as somewhat parochial.

In any case, the Languages Ladder was a slim document not intended as a sufficient basis for test development. As the Asset Languages development team put it, 'the CEFR is a comprehensive resource which we are finding very useful in implementing the Ladder as an assessment framework for lifelong learning. We see many advantages in treating these two six-level systems as essentially equivalent' (Jones 2006). The same article pointed out an important practical difference in how the two frameworks were to be interpreted. 'The Languages Ladder is articulated as learning stages, so that learners are said to be "working at Breakthrough" from zero until they fully achieve the stage at Grade 3. The CEFR is articulated as proficiency levels, so that learners are said to be "working towards Breakthrough" until they fully achieve the level. Thus it is the top band of each Languages Ladder stage which corresponds to achieving the CEFR level.' The relationship is illustrated in Table 5.3.

Table 5.3 Languages Ladder stages and CEFR levels

Languages Ladder learning stages		CEFR proficiency levels
Working at Intermediate	Intermediate 9	B1 Threshold
	Intermediate 8	Working towards B1
	Intermediate 7	
Working at Preliminary	Preliminary 6	A2 Waystage
	Preliminary 5	Working towards A2
	Preliminary 4	
Working at Breakthrough	Breakthrough 3	A1 Breakthrough
	Breakthrough 2	Working towards A1
	Breakthrough 1	

Yet another point of reference for the Languages Ladder was the UK NQF, which had been established by the Quantification and Curriculum Authority (QCA) in 1998 in order to rationalise the plethora of qualifications that had appeared over the years. It identified nine levels of qualification (Table 5.3). The relation of the Languages Ladder levels within the NQF was an important point of reference for Asset Languages, given that it set the new system in an explicit relationship with other existing UK qualifications. This relationship expressed the formal value of Asset Languages stages relative to

the General Qualifications such as GCSEs and A Levels, and was thus critical to whether and how schools would adopt the new system.

There was an expectation on the part of DfES that the Asset Languages stages should correspond neatly to other qualifications within the NQF, via expected levels of attainment in National Curriculum terms. This explains a table published by the DfES in the early stages of the project (simplified in Table 5.4 below), showing Asset Languages stages aligned to NQF levels and also approximate CEFR equivalences, in which Breakthrough straddles CEFR A1 and A2 and Preliminary covers A2 to B1 (Intermediate also being shown as B1).

Some explanation of the table may be helpful. Under General Qualifications, A Level denotes the exam marking the end of secondary education; GCSE denotes the exam marking the end of lower secondary education. Higher grades (A* to C) qualify as a Level 2 qualification, lower grades (D to G) as a Level 1 qualification.

The rightmost column of Table 5.4 shows the approximate CEFR level associated with the Asset Languages levels. This implicitly constitutes an additional statement about the CEFR level of GCSEs and A Levels, so that it would appear that Asset Intermediate Stage qualifications, for example, are equivalent in standard to the A* to C Grade GCSE, both being at CEFR Level B1.

Table 5.4 Asset Languages and General Qualifications in the National Qualifications Framework

NQF	General qualifications	Asset stages	CEFR approx
Level 3	A Level	Advanced: Grades 10 to 12	B2
Level 2	GCSE (A* to C)	Intermediate: Grades 7 to 9	B1
Level 1	GCSE (D to G)	Preliminary: Grades 4 to 6	A2 (B1)
Entry Level	Entry 1–3	Breakthrough: Grades 1 to 3	A1 (A2)

As discussed in Section 5.2.1.1, Asset Languages' adoption of functional proficiency levels as the basis of interpretation set it apart from the General Qualifications, where standards derive from a notion of expected achievement, and are compared across subjects on that basis. In language proficiency terms the interpretation of any general qualification is speculative and would be expected to vary across languages. The qualification 'approximate' in Table 5.4 perhaps recognises this, and subsequent attempts to equate GCSE and A Level grades to the Languages Ladder were resisted.

In matching the Asset Languages Ladder stages to the NQF the first four stages – Breakthrough, Preliminary, Intermediate and Advanced – were felt to fit comfortably at Entry Level and Levels 1–3 of the NQF. The placing of the two remaining stages of Proficiency and Mastery within the NQF was to prove less obvious.

Locating Asset Languages within the NQF illustrates well the possible tensions between the formal value of certificates and the practical value of the skills certificated. The NQF is not a proficiency framework, as is clear if we compare the Asset Language stages to Cambridge English exams which have also been accredited in the NQF (Table 5.5).

Table 5.5 shows the full range of the Asset Languages system, the Mastery and Proficiency levels corresponding to tertiary NQF levels 4 to 8. It also shows five Cambridge English exams at the NQF levels assigned to them by the owners of the NQF, the Qualifications and Curriculum Authority (QCA), together with the CEFR levels that Cambridge English assigns to these exams. The table shows that in formal terms these exams for non-native speakers of English enjoy a considerably lower level of recognition than exams for English learners of foreign languages. This is not especially surprising: the two exam suites were placed in the framework with respect to quite different criteria, not primarily related to proficiency. This reflects a societal value judgement which might even be defended; but it makes the point that Asset Languages' approach to setting standards in terms of the criterion of proficiency was novel, and alien to the established system.

Table 5.5 Asset Languages and Cambridge English levels within the National Qualifications Framework

NQF level	Asset Languages	Asset CEFR levels	Cambridge CEFR levels	Cambridge English exams
Levels 7–8	Mastery	C2		
Levels 4–6	Proficiency	C1		
Level 3	Advanced	B2	C2	CPE
Level 2	Intermediate	B1	C1	CAE
Level 1	Preliminary	A2	B2	FCE
Entry 3 Level	Breakthrough	A1	B1	PET
Entry 2 Level			A2	KET
Entry 1 Level			A1	

As another illustration of how hard it may be for professionals to think outside the value frameworks they are accustomed to: in an internal document a schools language exams expert, familiar with the standard of Entry Level tests, critiqued the specification for the Asset Breakthrough level (A1) as follows:

> The Breakthrough Examinations need to be made more accessible to such pupils as those catered for by the OCR "Certificate of Achievement". Such pupils have very limited concentration spans, and poor memories. They tend to cope badly with structures, and their knowledge tends to be limited to individual nouns, verbs and adjectives. Even these they will struggle to remember. Much of the candidature will have had learning support in the past.

This expert was certainly going to find it hard to appreciate the values which Asset Languages was setting out to promote.

5.2.2.3 Challenges with script acquisition for certain languages

The CEFR framework of levels has two aspects: it specifies and describes criterion levels of functional proficiency, while at the same time it provides a series of accessible learning targets – a learning ladder. For European languages it seems to be assumed that these two aspects relate to similar timeframes: that is, that the criterion levels are staged so that they represent a reasonable set of learning targets, in terms of the time and effort required to move from one level to the next. Thus A1 is generally seen as both practical and useful as a first learning target, as reflected in the curricular objectives of many European education systems.

However, some non-European languages raise difficulties, evident in the development of Asset Languages. There is a specific problem with learning to write or read in languages with complex, non-alphabetic scripts. Learning the writing script represents a significant burden, for native speakers as much as non-natives. For example, to achieve CEFR Level A1 in reading Japanese will predictably take much longer than English or French. For the beginning reader of Japanese, the functional proficiency level A1 requires much more time and learning effort, and is thus too difficult as a first learning target.

As the CEFR text explains and illustrates, it is perfectly possible to subdivide levels into smaller stages, creating as many objectives as necessary. For the multilingual Asset Languages framework, however, it was logistically impossible to develop and administer tests except by imposing a standard model across languages. The standard model did not fit the non-Latin-script languages well.

Given this serious constraint it was necessary to seek an accessible progression by revising the functional descriptions of the lowest levels and providing for a staged acquisition of the script, for example by using defined character sets. Thus the wide range of languages in the Asset Languages system points up a tension between the learning and proficiency orientations of the CEFR which passes unnoticed in the European context, where it is easy to think of learning and proficiency advancing together in the same manner, irrespective of the language.

5.2.2.4 Challenges of applying a common framework to different learner groups

While the Nuffield Inquiry proposed in their feasibility study that there should be 'a single generic' Language Learning Ladder, 'expressed in functional "can do" terms', they also acknowledged that it should have 'different presentational style and content for a finite range of target audiences, such as children to age 14, students 14–19, workplace learners, lifelong learners and speakers of other home languages' (Nuffield Foundation 2002:9).

This relates to North's (1995:446) requirement of the CEFR descriptors, that they should be 'context-free in order to accommodate generalisable results from different specific contexts, yet at the same time . . . context-relevant, relatable or translatable into each and every relevant context'.

Neither the Nuffield Inquiry nor the National Languages Strategy that followed discussed the ways in which learner contexts may differ and how these differences could be taken into account in the assessment. The challenge for Asset Languages was to implement a common framework of levels for learners of different languages and age groups, addressing issues of comparability. Three age groups were identified: primary, secondary and adult. Valid comparison of these groups would require assessments to be tailored to their particular needs and characteristics.

Comparability had both qualitative and quantitative aspects. Qualitatively, test content was differentiated chiefly in the choice of topic and lexis and also by targeting levels of cognitive development appropriate to the age group. The linguistic content was largely common across groups.

Quantitatively, it was felt important to attempt to construct an empirical link across the groups, if the system as a whole were to be defensible as coherent and interpretable. This was addressed for the objectively marked skills of reading and listening by including some common test tasks to function as an anchor across adjacent learner groups. The groups which differed most clearly were primary and secondary, so that selecting appropriate task types for anchoring across these groups presented the greatest challenge.

It would have been logistically impossible to create assessments in all contexts for every language, and in fact, it was not necessary, to the extent that little demand existed for certain combinations. As illustrated in Table 5.6, the development of tests for the three learner groups was only partially implemented: only for those languages commonly taught at primary level, and only

Table 5.6 Provision for three contexts of learning across levels

	Primary	Secondary	Adult
Breakthrough, Preliminary	✓	✓	✓
Intermediate	✓	✓	
Advanced and above	✓		

for the first two Languages Ladder stages: Breakthrough and Preliminary. In all other cases a test suitable for secondary/adult learners was provided.

5.2.2.5 Challenges concerning community languages

Community language learners, or 'speakers of other home languages' as termed by the Nuffield Languages Inquiry, were highlighted as needing a 'learning ladder' with a different presentational style. Again, however, the Inquiry did not analyse closely the ways in which community language learners were different and how these differences might be taken into account. Rather, they simply recognised that resourcing and expertise for such languages was likely to be an issue and stated that if the ladder were specified 'in generic terms applicable to all languages, it will be possible to appoint and endorse assessors for less commonly taught languages, who in turn could 'translate' the generic specification into assessment tests in the language concerned' (Nuffield Foundation 2002:10). In contrast to the modifications made for learners of Chinese and Japanese, the National Curriculum for Modern Foreign Languages made no provision for community language learners. The Inquiry's approach reflected a lack of research in this area. The literacy practices and needs of community language learners, and their teachers who have frequently been educated in their home country, may differ from those of modern foreign language learners (Keenan 2000).

5.3 Development of the external assessment: Pilot phase

5.3.1 Defining the skills to be tested at each level

Internal documents from October to November 2004 (Jones 2004) represent first steps in defining the testing framework, identifying a range of considerations and questions, and requirements to be addressed. Thought was given to capturing the nature of progression across levels, for example:

1. Levels define an *implicational scale*: that is, a given level contains and builds on all earlier levels.
2. Levels become increasingly *complex*, that is, new aspects (dimensions) of ability are added.
3. A level has *salient features* which are critical for distinguishing it from lower and higher levels, and which assessments should focus on.
4. Progression is defined in ways which *emphasise commonality* across different individuals and groups of learners. This favours accurate measurement and comparability across groups. Where there are differences across groups, e.g. between primary and adult learners at Breakthrough level, these are explained.

There were many questions to be addressed in articulating progression. Clearly, there would be no simple answers to many of these:

Progression across levels should reflect:

- the functional scales of the CEFR (which scales are likely to be most useful in capturing a non-language-specific progression of levels)
- pedagogic motives emerging e.g. from KS2 languages strategy (e.g. learning to learn)
- task type features
- task content features.

In what ways is this progression different for certain contexts? This is influenced by:

- stage of cognitive development
- target language uses – purpose for learning
- language background: acquisition or learning, L1 bilingual or L2 situation.

In what ways is this progression different for each language? This is influenced by:

- difficulty of writing system (how long it takes to master it in L1 education system)
- grammatical features: are there notions/functions or skills which require greater or less grammaticisation? Does it follow that this moves them to a higher or lower language level?

In what ways does L1 influence L2 learning?

- notion of language distance
- availability of cognates to speed vocabulary learning, receptive skills.

From these considerations the following steps were proposed for the Asset Languages development:

1. Define a 4-skill descriptive framework of progression across levels which is maximally applicable across languages.
2. Within this, identify characteristic features of particular learning contexts.
3. Revise the Languages Ladder statements as necessary to reflect this framework.
4. For each language, identify any specific issues with particular skill/level descriptions.
5. For each language, construct a curriculum framework in as much detail as is necessary and possible, mapping language-specific features to levels. Make this available to item writers, teachers and other

stakeholders in a form which will enable/promote good classroom practice and a good match between this and the assessment.

6. Devise a standard approach to documentation such that all of the above can be summarised and collated for a range of administrative, professional or research purposes.

The CEFR was studied to identify potentially useful parameters:

- Domains of language use (personal, public, occupational, educational).
- Conditions and constraints, i.e. the external conditions under which communication occurs: Would these serve to provide qualifiers of performance at a level?
- The external context of use (sub-domains defined by locations, institutions or persons): Would these be useful for identifying commonality across the primary, secondary and adult contexts? Would they help distinguish community languages from foreign languages?

Preparatory material for a workshop on reading offered the following analysis of cognitive and metacognitive processes from Macaro (2001) and O'Malley and Chamot (1990):

Cognitive

- *Linking* words or ideas to visual images.
- *Inferring* what a phrase means from the immediate surrounding text.
- *Deduction* – applying rules to understand language.
- *Organisation, or grouping* and classifying words, terminology, or concepts according to their semantic or syntactic attributes.
- *Summarising*, or intentionally synthesising what one has read to ensure the information has been retained.
- *Transfer*, or using known linguistic information to facilitate a new learning task.
- *Elaboration* – linking ideas contained in new information or integrating new ideas with known information (elaboration may be a general category for other strategies, such as imagery, summarisation, transfer, and deduction).

Between cognitive and metacognitive

- *Deciphering* if L2 words look like L1 words.
- *Memorising* a list of vocabulary items.

Metacognitive strategies

- *Planning the organisation*
- *Monitoring or reviewing attention* to a task, monitoring comprehension for information that should be remembered, or monitoring production while it is occurring.
- *Evaluating or checking comprehension* after completion of a receptive language activity, or evaluating language production after it has taken place.

The core Cambridge English examinations provided useful exemplars of tests at each CEFR level, albeit for English, which was not a language tested within Asset Languages. The CEFR descriptors, taken together with the Languages Ladder scales, suggested general linguistic/functional level and appropriate tasks. They identified salient features critical for distinguishing a level from lower and higher levels, which assessments should focus on.

The stage of theorising and planning reflected above finally feeds into the practical process of specifying the content and form of tests. Appendix C provides a shortened view of the 2007 specification for the tests in four skills at each Asset Languages stage to Advanced.

5.3.2 Test administration

Asset Languages' proficiency-based approach and formative purpose raised a number of significant administration issues. Asset Languages had to fit in easily with classroom routine if it were to fulfil its formative function. This would simplify administration, avoiding the need to book examination halls and disrupt teaching, as happens with GCSE examinations. It enabled testing when ready, rather than in a cohort. All assessments would be short enough to fit into a normal class period.

The item-banking approach required the capture of item-level response data for the objectively marked tests. This was vital both for pretesting and to incorporate live data in scale construction and validation. Optical mark recognition sheets were adopted for this purpose. Fears that candidates would find the sheets difficult to use proved to be largely unfounded.

Administration systems enabled centres to take tests on the day of their choosing: 'effectively on demand'. Five 4-week testing windows per year allowed centres to choose sessions and dates. In practice, the busiest sessions were May and June and to a lesser extent, January.

Sample materials for the main languages were created so that students could familiarise themselves with the format of Asset Languages assessments. Test-focused preparation practice, such as rote learning of prepared texts for the speaking test, was explicitly discouraged.

Following trialling, which suggested that the use of dictionaries did not improve performance, dictionaries were banned in the external tests. In place of dictionaries examiners looked at ways of providing additional support within the task, with key words provided both in the target language and in English.

5.3.3 Scale construction, grading and standard setting

Given the multilingual and inclusive nature of the Asset Languages framework, the process of scale construction and the setting of standards was the most challenging technical and procedural aspect of the development. In the model of test validity proposed by Weir (see Section 2.1) it is the area treated under the heading of scoring validity.

Challenges lay in the complexity of the framework and the high degree of comparability between groups (across languages, levels and so on) which it implied. Substantial time and resources needed to be given to all aspects of scoring validity, and particularly to the pursuit of comparability across languages.

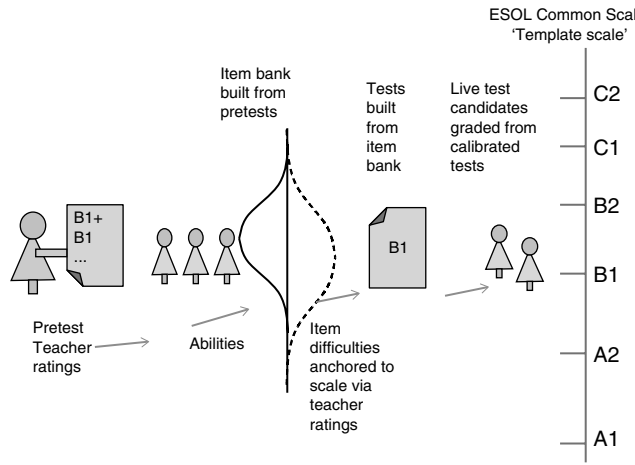
The objective skills of reading and listening and the subjective skills of speaking and writing are presented separately below.

5.3.3.1 Reading and listening

The objectively marked skills of reading and listening were tested using an IRT based item-banking model (see Section 2.2.2). This section provides more detail on how the scale and the standards were established in the difficult, accelerated context of the Asset Languages development. The process necessarily involved iterative approximation, given that the tests needed to go live and start reporting results on the basis of a minimum of empirical data. Scale construction needed to exploit data as it became available within the development schedule, and this meant basing it on pretesting.

Pretesting was conducted in schools interested in adopting Asset Languages, who could volunteer students to complete tests. Teachers in the participating schools were asked to rate each student's level. Initially they did this in terms of National Curriculum levels, as it was felt that these would be most familiar to them (see above for how National Curriculum levels related to the Languages Ladder). Subsequently these were supplemented by a rating scale based on Languages Ladder and CEFR descriptors. The collected ratings were then used in an IRT analysis to estimate the abilities of students, which could be summarised so as to link the group to a proficiency scale. The IRT analysis simultaneously placed the pretested items on the same scale, anchored via the estimated abilities of the students. Therefore live tests subsequently constructed from these items could be used to locate the tested students on the proficiency scale. Figure 5.3 illustrates this process.

Figure 5.3 Provisional grades derived from teacher ratings



The attractive feature of this approach, which I have not seen referred to elsewhere in discussions of standard setting against the CEFR, was that teachers determined the standards, though not in a context where this impacted directly on their own students. A disadvantage was that teachers' ratings could not be well standardised in the conditions under which they were elicited, but one could expect that the error would cancel out to an extent. Above all, it offered a practical way of setting an initial standard.

Figure 5.3 also shows the Cambridge ESOL Common Scale used as a template for the developing Asset Languages item bank. External tests for each Languages Ladder stage were developed and rolled out in sequence, starting with the lowest level. This represented a further constraint on the scale development process: the final objective was a single scale, with well-articulated continuity across the levels, but the scale needed to be constructed level by level, starting at the bottom. In this situation it was useful to be able to make reference to the common scale which existed already for the Cambridge ESOL exams, and use this as a template to which, it was reasonable to believe, the completed Asset Languages scale would approximate. Using the Cambridge ESOL Common Scale as the basis of a template for Asset Languages was defensible to the extent that:

- the Cambridge exam levels were quite well linked to the CEFR
- the response data used to construct the scale reflected the same mix of task types for both exam systems
- the method of scale construction (essentially, linking level by level using anchor tests or other evidence) was the same for both exam systems.

These conditions are important, as they may be expected to influence strongly the form of the scale which emerges. As presented in Section 4.3, the Cambridge ESOL Common Scale is not directly comparable with the Rasch scale underlying the CEFR's Can Do descriptors and level thresholds, as developed by North (1995, 2000). However, the progression demonstrated by the Cambridge exam levels can be seen to have a certain coherence, reflecting a reconciliation of twin goals: to accredit substantive learning gains, while providing a series of accessible learning targets.

This approach allowed a way forward. The Common Scale could be taken as a template. Each level would be anchored to the scale separately, using the level of learners as estimated by teachers at pretesting. Over time the vertical articulation of the scale could be verified and adjusted by specific anchoring across levels. Occasionally such anchors occurred serendipitously in live data, where a group of learners took tests at two levels simultaneously. Other linking data was also collected under experimental conditions.

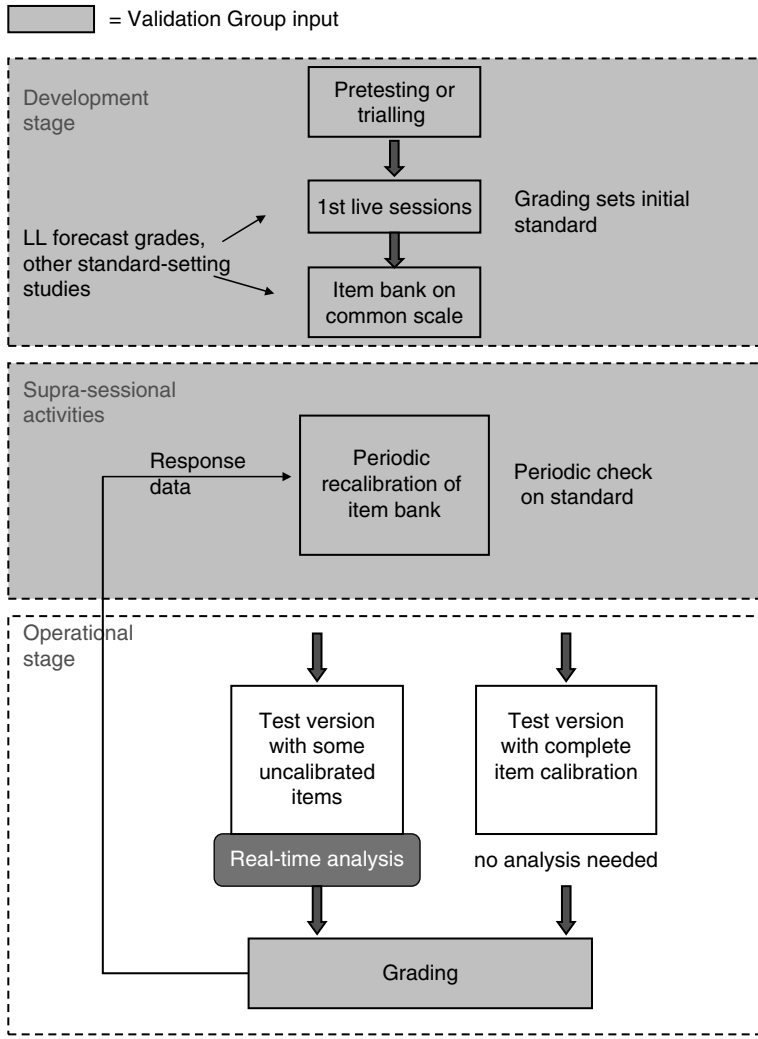
As such evidence became available it could be weighted against the evidence from pretesting. To provide a decision process an algorithm and spreadsheet was developed to derive grade thresholds from data, using explicit weightings of these two sources of evidence – the pretest calibrations and the vertical linking evidence, reconciling them subjectively, based on the confidence placed in them. This enabled progressive approximation to a stable and coherent scale for each language and skill.

Actual setting of grades for particular test administrations was deemed to require a grading meeting with the participation of the subject experts, in line with practice regarding schools exams, and certainly desirable in terms of the professional development of the new teams of examiners needed by Asset Languages. The quantitative evidence described here was presented as one part of the evidence considered at the grading meeting. Grading meetings are treated more fully in 5.3.3.3 below.

The empirical work to develop and apply standards for reading and listening extended over a longer period, and needed to be carried forward in a cycle which was different from the familiar session-based approach used for standard Cambridge ESOL exams. That approach was based on just two sessions per exam per year, and all development and administration phases were focused on a single session. Pretesting, for example, was undertaken for each session, to a quality which provided adequate information for reliable grading. Evaluation of the quality of marking, and statistical compensation for marker effects, also operated on a by-session basis. With Asset Languages the development phase produced material that would be re-used and retired in the operational phase over a number of sessions; necessarily, given the tight development schedule, data to improve the calibration of materials and the articulation of the measurement scale came in a continuous stream.

Thus a higher-level, supra-sessional cycle was defined, enabling periodic review and recalibration of material. Figure 5.4 below sketches this stage for reading and listening.

Figure 5.4 The grading model for Asset Reading and Listening tests



The above description gives an indication of the practical constraints within which the development of the Asset Languages levels had to operate. Full pretesting across the volume of languages and levels proved to be impractical, given development timescales, and as importantly, the

availability of students to do pretesting. Thus it was necessary to adopt approaches for decision making on the basis of sparse data. Particular issues arose with pretesting for languages where there were small cohorts in early live tests, e.g. Panjabi. It was challenging to find sufficient numbers to pretest given that the same candidates might be taking the live tests. It was often necessary for scale construction and standard setting to be done simultaneously within a single session rather than in a linear process extending across several sessions. The process was designed to be iterative and cumulative and therefore as more candidates took tests and more data was available, more information accrued about item difficulty, which led to better informed decisions about standard setting and grading. This means that during the development stage, the evidence for standard setting, i.e. grading, came primarily from pretesting and live pilot administrations, as well as from cross-level pretests (candidates completing a pretest consisting of tasks at two different adjacent stages of Asset Languages) and estimates of National Curriculum levels provided by teachers. During the development stage therefore, grading required close attention from the Research and Validation team, and not all grading decisions could be strongly supported by hard data. This describes the situation when French, German and Spanish at Breakthrough, Preliminary and Intermediate stages were first assessed in February 2005. As qualifications developed, the awarding process changed. By the time that the early qualifications were assessed in November 2007 there was sufficient confidence in the quality of the data to reduce the awarding committee to the Chair and a member of Research and Validation, whose main task was to identify any potential statistical anomalies for further investigation. At this time, Asset Languages could be described as being in the operational stage.

The operational stage begins when test versions can be constructed with at least some tasks reliably calibrated to a common scale for which the grade thresholds have been established. Each series of test administrations requires real-time analysis of response data to calibrate the remaining items and enable grading. Where a version is used unchanged over several series, or is constructed entirely from recycled calibrated tasks, then no real-time analysis will be required, though grades and scaling parameters may still need estimating.

When one block of test administration sessions is complete (the supra-sessional cycle shown in Figure 5.4), the entire item bank of interlinked tasks can be re-calibrated and adjustments made to the standards, using live data and additional experimental data providing the vertical link across levels. This is done before the next block of tests is constructed, ensuring that tests are constructed on progressively more accurate calibrations.

5.3.3.2 Speaking and writing

The subjectively rated performance skills of speaking and writing seem relatively straightforward in comparison with the objectively marked skills of

reading and listening. Learners' proficiency level can be described in terms of observable performance: what tasks they can perform and how well they can perform them. The first step in constructing a scale for speaking and writing is thus to select tasks which are relevant to the construct and which offer an appropriate degree of challenge for each target level, to elicit performance which can be rated with respect to that level. In practice the factors of task difficulty (the what) and performance quality (the how well) are very difficult to disentangle in performance assessment. This is why Can Do descriptions of level such as the Languages Ladder or the CEFR are not on their own sufficient as rating instruments. What is needed are exemplar performances to provide concrete illustration of the target level. For Asset Languages much effort was put into providing exemplar scripts, training, and standardisation. In this way it was attempted to maximise the consistency of standards across exams and across sessions.

An important feature of Asset Languages (and a colourful one, given that representatives of each language tended to come along in national dress) were the cross-language standardisation days. Examiners and moderators for writing and speaking for all of the languages assessed by Asset Languages attended these days. More detail on these events is given below.

Rating scales for evaluating speaking and writing were defined as tables of criterion-related descriptors. These were made available to students and teachers on the Asset Languages website. Mark schemes used the convention that they should be interpreted in relation to the agreed standard for that stage, as enshrined in training and standardisation materials. The Can Do scales of the CEFR were found particularly useful in standardising expectations of performance at each stage. Some participants in training sessions suggested that the CEFR should be incorporated more formally into the articulation of the rating scales.

Specific adaptation of the marking criteria for writing was necessary to accommodate non-Latin script languages. Two amendments were made: the reference to character formation rather than spelling; and reference to use of kanji. The first applied to all non-Latin script languages while the second applied solely to Japanese, clarifying to examiners expectations of candidates' use of kanji. This amendment was necessary as Japanese has three character systems, and it would be possible for candidates to write all the text in one of the simpler systems. Expectations regarding the use of kanji therefore required specification.

Feedback from the pilot phase motivated several changes. The criteria for writing and speaking were more closely aligned. Extra emphasis was placed on interpreting the criterion descriptions in relation to features of expected performance at the level tested. At Preliminary and Intermediate stages the Band 0 descriptor was modified, to counter the reluctance of examiners and moderators to use it. Level 0 is an important concept for a learning ladder, as

it indicates failure to achieve the level of the test. It is quite possible for a candidate to achieve the lower level but to be awarded 0 on the next level up. For examiners used to working with a single exam covering a continuous range of levels this was very hard to accept. This illustrates the conceptual difficulties that the Asset Languages system of discrete proficiency levels presented to markers familiar with other exam systems. Speaking criteria were simplified to two: Language and Communication, with Pronunciation subsumed into the Communication criteria.

5.3.3.3 Setting standards: Grading

In the complex Asset Languages framework (see Section 5.2.1) standards refer to criterion proficiency levels, and should be comparable across languages and contexts. The pursuit of comparability drove much of the materials construction and the marker training described in this chapter. Grading had formal procedures, initially based on the OCR model of an awarding committee for each qualification, with key external personnel. Later, as the number of languages and levels grew and standards became established, awarding committees were substantially reduced in size and outcomes were determined more by statistical evidence.

With reading and listening, the item-banking methodology adopted for Asset Languages was intended to enable a constant standard to be applied using statistical methods. For examiners only familiar with GCSE exams this was a novel concept. However, the stability of the calibrated scale was initially quite uncertain (Section 5.3.3.1), so that for the first year of a qualification human judgement was certainly not to be disregarded. Where recommendations based on statistical findings were questioned the chair of item writers would lead a discussion of individual tasks and the balance of the paper. Thresholds were determined through examiner judgement informed by the statistics, rather than vice versa.

The statistically derived thresholds for the three grades characteristically covered a narrower mark range than OCR examiners were used to, or found plausible. To them, custom and common sense suggested that the grades should be spread out over the whole of the available mark range. We could explain that each grade within a level in fact covered a relatively small part of the ability range defined by the whole system of levels (a contributory factor in the narrow separation of grades was relatively poor item discrimination on some early test versions). It was educational impact which was the focus of discussion. The estimated grade 1 in particular was generally judged as high, compared with the number of items in the paper which examiners considered to be at grade 1. Committees were concerned at the de-motivating effect on candidates who failed to achieve even the lowest grade, and such sentiments led to a lowering of that threshold. Given Asset Languages' formative purpose it would be hard to quarrel with this view, but the discussions at

grading meetings illustrate well the challenges of introducing a proficiency-based, measurement approach into a context where the practice and purpose of assessment is understood in rather different ways.

For speaking and writing, thresholds were pre-determined, given that the mark scheme was criterion referenced to the functional proficiency levels described by the Languages Ladder, and translating scores into grades was straightforward. For these skills the task of the awarding committee was to check on the leniency or severity of markers, evaluate the standards applied and agree any proposed statistical corrections to these. Was the profile of grades consistent with understanding of the cohort? Since significant work had already gone into creating and aligning standards, scaling was not common, and it was more likely to be necessary in the early days of a qualification.

The marks given by teachers in the Speaking assessments were subject to a process of moderation. All teachers conducting Speaking tests first underwent a process of standardisation within their centre. Recordings of the Speaking tests were sent to the moderator, who marked a structured sample of them and compared his marks with those of the centre. A significant difference would lead to the centre being asked to conduct internal standardisation and re-submit. Finally the teacher marks were adjusted by an automated linear scaling process, calculated from a comparison of the moderator and teacher marks.

As explained in Section 5.2.2.2 and illustrated in Table 5.3, the relation of Languages Ladder stages to CEFR levels demonstrates different conceptions of 'level'. In the case of the CEFR a learner is 'at the level' when they demonstrate minimal mastery, and remain at the level until they minimally achieve the next level up. This is the *proficiency* interpretation. The Languages Ladder on the other hand has a *learning stage* interpretation: a learner is considered to be working 'at the level' at the point when they begin to work towards the goal of achieving mastery of the level – that is, when they have just mastered the level below. This difference in articulation of the levels presented a problem for what should qualify as a passing grade. After considerable discussion, a pass was defined as achievement of the *lowest* grade within each stage. Achieving the *highest* band of each Languages Ladder stage would demonstrate full mastery of the level, that is, achievement of the level in CEFR terms. This decision had implications for test construction, making it necessary to produce sufficient questions to test the level of proficiency indicated by the Can Do statements describing the lowest grade.

The publication of grade thresholds was another issue where the IRT-based Asset Languages model came into conflict with traditional practice. OCR's general policy of publishing grade thresholds could not accommodate an IRT model, where the relation of scores to abilities varies from session to session. The 2006/07 report to centres (Oxford, Cambridge and

RSA Examinations 2007) attempted to explain this, while undertaking to work towards more standardised thresholds:

> At present, thresholds vary for different versions, but we plan over time to work towards target thresholds. For the time being, a rule of thumb would be, for example, 13, 16 and 19 out of 25, with the exception of Breakthrough, where the lowest threshold would be 10 . . . .

This ambitious proposal perhaps reflected an intention to refine item writing and test construction for each language over time, so that indeed, raw scores would come to stand in a more regular relationship to ability thresholds; that is, test facility would be more tightly controlled. In the early stages however it was agreed to withhold such information, until the standards of papers became more uniform.

5.3.3.4 Standardisation

It was a challenge to ensure a common understanding of levels given the wide range of languages offered within Asset Languages and the different contexts in which the languages were learned and taught (community languages and modern foreign languages). The main approach was through the use of exemplars in English. English is not a language in the Asset Languages framework, but this was seen as the simplest and most practical way of exemplifying levels in the 20-plus languages to examiners and moderators.

For speaking, English video exemplars of Asset Languages Speaking tests were filmed in both foreign language and second language learning contexts. Writing samples were collected in a similar way. To ensure comparability of judgements across languages it was important that the exemplars represented the Asset Languages test format. Available English exemplars from other Cambridge English exams were considered but not used as they would have introduced extraneous issues. Using the Asset Languages format focused attention on specific issues of how the same construct and the same notion of level could be implemented in perhaps very different languages.

The exemplars also demonstrated how important the test task is to perceptions of level of performance. They included examples of the same learner taking the test at two levels, typically appearing more fluent and confident at the lower level, but struggling at the higher one – that is, demonstrating mastery of one level but failing to achieve the other. This was important in training markers how to use the 0 band, which, as already noted (see Section 5.3.3.2) had proved quite problematic.

In addition to the English exemplars, standardisation exemplars were also to be produced for each language, although in the early stages these were only available for French, German, Spanish and Italian. These were used by principal examiners and principal moderators in standardising raters, and

also in explicit cross-language comparisons of standards. These team leaders were responsible for holding training and standardisation days for their raters and for ongoing monitoring of the quality of their team's work. In addition to the above training, which was given to all raters, cross-language standardisation days were held for principal examiners and principal moderators. These brought together the team leaders to provide training which they could cascade further, and to discuss particular issues. Standardisation through English exemplars achieved something which would otherwise have been impossible: teams engaged on different languages could come together to understand standards in an interactive and collegial context. This was undoubtedly beneficial. More difficult to verify was of course the extent to which a shared understanding of standards in English would enable each examiner to set a similar standard for their own particular language.

There were a few other ad-hoc standardisation events. Once or twice standards issues raised at grading meetings required specific attention. This was to be expected where teams of markers and moderators were still relatively inexperienced. Thus, for example, in one case a multilingual expert was asked to adjudicate on the standard applied across French and German in a particular session.

5.4 Developing the scheme 2005–08

5.4.1 Rolling out all the languages

Development of new qualifications usually follows a schedule agreed with the Qualifications and Curriculum Authority, specifying periods for development, training and evaluation. With Asset Languages political considerations dictated a far shorter timescale, effectively delivering the development phase in nine months rather than two years and evaluating the pilot before the end of the full period. The compressed timescale made it difficult for interested schools to fit the pilot into their timetables. However, national interest in the project was such that there was little difficulty in recruiting pilot and pretesting centres.

From September 2005, the scheme became a 'national pilot', open to any centre who wished to make an entry. Effectively this meant that the pilot stage was carried over into the national rollout.

The scale of the project was ambitious: six stages, covering four skills and 17 grades, external and teacher assessment, primary, secondary and post-16 learners, computer and paper-based assessment and 25 languages. Such a framework evidently could not be delivered in its entirety at the same time: prioritisation was needed. In the first three years it was the Department for Education and Skills who set the priority for which languages and stages to focus on. Breakthrough to Intermediate stages were prioritised, followed by

Advanced. Proficiency and Mastery presented a range of issues which set them apart from the lower stages, and their development was put back to the final phase of the project.

Entries for each language varied widely, from thousands per session for the major languages taught in UK schools (French, German and Spanish) down to very low entries – perhaps a single class within a single centre, for many minority languages with no history of assessment. For these very small qualifications it proved impractical to apply the item-banking development model described above, and a pragmatic approach was needed to setting standards, making use of whatever objective or subjective data might be available.

For the large, popular qualifications, on the other hand, it was important to provide three or four simultaneous versions of the external assessment, enabling schools to organise several administration slots. Schools sought as far as possible to avoid repeated use of the same version. At the same time an important design constraint was that versions should be linked by a proportion of common items, to ensure comparability. The number of versions developed for each language depended upon its popularity and upon the stage. Higher stages were considered higher stakes, requiring more versions for security purposes.

While 25 languages were offered at Breakthrough, Preliminary and Intermediate stages, of these 14 were offered at Advanced, six at Proficiency and only one (French) at Mastery. The final package of qualifications to be offered from September 2008 is shown in Appendix C.

5.4.2 Enabling different assessment models

The modular structure of the Asset Languages framework, with its range of levels and on-demand availability, offered schools a much more flexible form of assessment than hitherto. Whilst virtually all secondary schools offered GCSE qualifications, these were of limited value to significant cohorts of students, and Asset Languages enabled alternative approaches. Initially, Key Stage 3 students (age 11–14) provided by far the largest number of entries. September 2004 saw the end of languages as a mandatory part of the Key Stage 4 curriculum (age 14–16), and language departments were faced with problems of motivation among students at Key Stage 3 who realised they could drop the subject at the end of the year. Asset Languages enabled teachers to emphasise positive achievement using the Teacher Assessments, and to provide formal recognition of the language studied at the end of the year, even if the student did not continue. For many, seeing that their Asset Grade 4 or 5 already represented the standard of a GCSE Grade E or F was sufficient to persuade them to continue their studies into Key Stage 4 with a realistic expectation of success. Having experienced success with these students, schools began to appreciate other groups who would benefit from the scheme:

- students at the transition from primary to secondary schools, either through informal teacher assessment or more formal external assessment
- gifted and talented learners who might have achieved GCSE in the language in year 9 or 10 and wished to continue their studies into year 11, though not necessarily on to A Level
- weaker GCSE students, who could be motivated through the use of teacher assessment and possibly external assessment in place of GCSE
- community language learners supported by the school, often in after-school and lunchtime clubs, who would have means of recognising achievement
- participants in the increasing number of short, intensive language courses in years 10 and 11, which provided a valuable set of language skills, but not necessarily at GCSE level
- participants in language courses offered as part of General Studies at Key Stage 5 (age 14–16) either as continuation of a language studied at GCSE or as a new language to be studied.

However, while the innovators and early adopters were beginning to appreciate the possibilities for the scheme, the process from first expressions of interest up to integrating Asset Languages into the curriculum could be a protracted one.

5.4.3 Specialised Diplomas and the World of Work

As the contract rolled out from its development phase and OCR took over live assessment, a number of variations on the Asset Languages theme were considered, most of which did not finally come to fruition. One which did go forward concerned the new system of Specialised Diplomas proposed by the 2005 White Paper *14–19 Education and Skills*. Concerns were expressed that these made little reference to languages as a core functional skill. In response the DfES proposed renegotiating the Asset Languages contract to include some applied qualifications designed to support the 14–19 vocational agenda, 'applied' being the term used as part of a general attempt to create parity of esteem between vocational and academic qualifications. It was agreed to offer French, German and Spanish at the Preliminary and Intermediate stages, branding the qualification 'Asset Languages World of Work'.

The differences between a general and 'applied' language qualification engendered considerable discussion:

- Do the differences chiefly concern vocabulary?
- Are there significant differences in the nature of functional interaction within applied contexts?

- If an applied vocabulary is important, should it be more clearly specified than in the general Asset qualifications, or should the functional specification be widened?
- Does the notion 'applied' make sense at Breakthrough level?
- Should qualifications be targeted to specific vocational areas?

It was decided not to offer the qualification at the Breakthrough stage. It was also agreed that, in order to ensure comparability at the Preliminary and Intermediate stages, the functional framework should be the same for World of Work and General Qualifications. The World of Work specifications would indicate the kinds of area and vocabulary that would be assessed, including general workplace language but not the technical language of specific areas.

5.5 Research around Asset Languages

From the beginning of the Asset Languages development it was clear that it invited, indeed required, a serious research agenda. Research was needed to address the substantive conceptual challenges implied by the multilingual framework, but these of course all had practical implications for the design. Two researchers recruited to the project at the outset were invited to complete PhDs in the Cambridge University Faculty of Education on relevant topics, while at the same time carrying forward the actual implementation work. An early paper in *Research Notes* (Jones, Ashton and Chen 2005) identified some of the challenges: establishing comparability within a multidimensional framework, and implementing an approach to providing formative feedback. The same issue included discussion of conceptual issues in linking the Languages Ladder to the CEFR and the Cambridge ESOL Common Scale (Jones 2005a).

Regular updates in *Research Notes* give a picture of the project unfolding. By February 2006 progress in constructing the measurement scale could be reported, as pretests linking adjacent stages were introduced in order to improve the quality of the vertical linking. Such tests had been taken by candidates in Chinese and Japanese, and more were in development for other languages. A cross-language analysis of the statistical performance of task types was reported. It found similarities in performance across languages, supporting the view that the skills constructs were comparable, and also identified those task types that generally performed better or worse. A content analysis of Breakthrough stage tasks compared them with the Asset Languages objectives for this level, and raised issues with the interactional authenticity (see Section 2.5.4.2) of some tasks.

While such studies focused on improving the reliability and validity of the assessments, others could be used to look for evidence of positive impact

that the new scheme might have in schools. Coleman, Galaczi and Astruc (2007) studied the nature of learner motivation on over 10,000 school pupils at Key Stage 3 (ages 11–14) and its relationship with gender, age, and type of school. Asset Languages pilot centres were included in the sample, and motivation was found to be higher in these schools (which had demonstrated a commitment to languages by volunteering to pilot-test the new Languages Ladder) than in other similar schools. *Research Notes* reported that pupils were motivated by receiving more specific feedback on their ability than with other assessments, and being assessed separately in listening, reading, writing and speaking.

By August 2006 *Research Notes* could report on work to set the standard of the new Advanced level (CEFR B2), and also on several projects undertaken to investigate the relationship between Asset Languages levels and the levels of current qualifications within the UK education system. In one such project, over 200 candidates sat both Asset Languages and GCSEs in all four skills for French. Self-assessments and teacher ratings of candidates were also obtained using the Asset Languages Can Do statements. Another project was investigating the relationship of Asset Languages and National Curriculum grades. Such studies recognised the importance for the new system of relating it clearly to existing familiar reference points.

Jones, Ashton and Walker (2010) viewed the Asset Languages development from a different perspective: as a case study of using the pilot version of the Council of Europe's Manual (Council of Europe 2009). Given the complexity of Asset Languages as an example of linking assessment to the CEFR, this study stressed the importance for framework construction of minimising human judgement and maximising the use of existing statistical information about the likely form of the framework – the use of the Cambridge ESOL Common Scale as a template for the objectively marked skills.

Ashton (2008) is the PhD completed by Karen Ashton while working on the Asset Languages project. It explored an important issue for the validity of comparing learners of different languages within the same multilingual frame of reference. It compared the reading proficiency of secondary school learners of German, Japanese and Urdu in England to investigate and shed light upon the feasibility of relating learners of different languages and contexts to the same framework. This question also had important implications within education for other frameworks such as the National Curriculum for Modern Foreign Languages (Department for Education and Skills and the Qualifications and Curriculum Authority 1999) and for the application of the CEFR (Council of Europe 2001) to a wider range of languages.

The study employed a mixed-methods approach, using self-assessment Can Do surveys and think-aloud protocols, to compare the reading proficiency of secondary school learners of German, Japanese and Urdu in England. While three common factors were found to best represent learners'

understanding of reading proficiency, there were also strong differences. The difficulty of script acquisition in Japanese impacts on learners' understanding of the construct, while learners of both Japanese and Urdu were unable to scan texts in the way learners of German were able to. Urdu learners underrated their ability, not taking into account the wide range of natural contexts in which they use Urdu outside the classroom; this finding also illustrates how Urdu learners use their knowledge of the spoken language as a resource when reading. Finally, the study demonstrated that the construct of reading in the National Curriculum for Modern Foreign Languages was not reflected by any of the learner groups, which Ashton concluded was worrying for language education and assessment within England, and raised the need for further research.

5.6 Teacher Assessment: Less formal accreditation of learning

The presentation so far of the complex Asset Languages framework has focused on the external, formal assessments. Now we turn to the internal, less formal teacher assessment dimension which was potentially one of the major innovations offered by the scheme. 'Teacher Assessment' was how Asset Languages referred to assessment of students conducted by the teacher, as opposed to the external exams.

The notion of providing teachers with a system for informally accrediting their students' progress was at the heart of the original recommendations of the Nuffield Inquiry. Indeed, the feasibility study which followed the enquiry was clear that the 'LLL [Language Learning Ladder] initiative should not enter the field of formal qualifications, but operate alongside it by defining levels and equivalences in accessible terms, and by offering informal certification' (Nuffield Foundation 2000:12). None the less, the formal certification which was proposed by the DfES and implemented by Asset Languages was undoubtedly essential if such an alternative framework was to compete successfully for the attention of schools. The Nuffield Inquiry did in fact recognise that the 'learning ladder' aspect was most important at lower levels.

Formal assessment was what Cambridge English Language Assessment and OCR understood best. Developing the teacher assessment strand presented new and unfamiliar challenges. The external and internal assessments were doubtless thought of and referred to by many as the summative and formative strands respectively, but in the context of Asset Languages these familiar terms are misleading, because the whole purpose of the scheme, emerging from the National Languages Strategy, was formative – to bring about better learning of languages. The ladder of graded levels was to provide motivating and accessible targets, and the focus on functional proficiency was to make language study meaningful and motivating. By breaking down each major

stage into three teacher-assessed grades the intention was to provide even more accessible goals and opportunities for rewarding achievement, and to ensure a positive synergy between the formal assessments and classroom work.

Jones (2007:15) discusses the various conceptions of 'formative assessment' evidenced in the practice of classroom-based testing. In the UK it was the Assessment Reform Group which had advocated a shift in emphasis away from assessment *of* learning towards assessment *for* learning, as a reaction against what was widely seen as an exaggerated emphasis on national testing for accountability (Assessment Reform Group 1999). In the US, however, the *No child left behind* program had appropriated the term in the context of a proliferation of testing aimed at driving up educational standards. Pellegrino (2003) rechristened the *No child left behind* program *No child left untested.* The developers of Asset Languages were at pains to stress that the scheme was not intended to lead to a test-driven approach to teaching.

But the development coincided with a period in which the role of teachers in assessment was seen in conflicting ways. Proposals for reforms of assessment in the 14–19 age group had recommended professionalising teachers' role in assessment for summative purposes (Department for Education and Skills 2004:57); and the prospect of restoring professional status eroded by standardised national testing led some to advocate a vision of teacher assessment stressing the traditional summative assessment virtues of reliability, objectivity and consistency (Tattersall 2004). In contrast, Leung (2004:21) advocated a view of formative assessment as flexible adaptation to local, immediate learning contexts; it should not consist in measuring achievement against an inventory of externally defined attainment targets.

This view was in line with the intended function of Asset Languages teacher assessment: it should fulfil a formative role by fitting flexibly into existing schemes of work. Teachers should be able to adapt tests to suit their local, immediate contexts. As its publicity material stated, 'there is no specific syllabus. Asset Languages is designed to fit in with any course'. The inherent weakness of this position has already been referred to in Section 5.6 above. In the event adaptation to local contexts was not to be straightforwardly achieved. As described in more detail below, the adopted approach to teacher assessment was simple, and constrained by practical concerns. The form of materials developed comprised a pack for each language and level. For each Asset grade within a level a number of test tasks were to be selected from the pack and administered. Teachers could adapt some of the tasks where necessary, for example, to use already-taught vocabulary. As Jones (2007:30) explained:

> This is a model which is practical to implement initially, but the intention is to develop it further as a community of practice develops among teachers and schools using the system. Ideally the role of Asset Languages would be less to produce materials for teacher assessment, but rather to facilitate the development and sharing of materials and

> ideas by users of the Languages Ladder – effectively, contribute to the "revival and reinvigoration of the principles and practice associated with . . . graded objectives" originally called for by the proponents of the Languages Ladder. The exam board's role would be to do empirical work to ensure the alignment of materials to the framework and thus to the standards of the external assessments. Such a conception of formative assessment would escape the reliance on externally-defined attainment targets while preserving the link to the wider interpretative framework.

The hope that teacher assessment might develop in this way proved over-optimistic, and the grass-roots graded objective movement was, it proved, beyond revival. Practice revealed the issues and tensions embedded in the teacher assessment model.

5.6.1 The approach to Teacher Assessment

Reviewing the approach to the formative Teacher Assessment and the design of the materials, it is striking how little they departed from the traditional summative assessment model. There was, of course, a need to ensure as far as possible that the two assessment strands referred to the same system of proficiency levels. The developers were concerned to avoid the situation that learners might successfully complete a set of Teacher Assessment grades and then fail the external test at the corresponding level. How could comparability be ensured across Teacher and External assessment modes so that grades were aligned? How could Teacher Assessment be quality assured?

The answer to these questions was to opt for making the External and Teacher Assessments identical in style and format across languages, and ensuring that Teacher Assessment should have the same 'look and feel' as the corresponding external assessment. It was clear that the Teacher Assessments in their basic form were to be proficiency based, like the external tests, rather than tests of achievement.

The Teacher Assessments were to be provided in a form which allowed them to be embedded into teaching time and co-ordinated with normal teaching. This would support teachers in their work and provide them with high-quality and standardised assessment materials. Teachers would be given materials which would have a formative function but also provide low-stakes summative assessment of a skill. This philosophy underpinned the approach to development of the Teacher Assessment strand of Asset Languages.

That teachers could issue grade awards (as the teacher assessment certificates were called) directly to their students was expected to be motivational and, through setting a series of short-term goals, to provide a good preparation for external assessment.

The view of Teacher Assessment as a preparation for external tests was

evident in such statements as the following (Oxford, Cambridge and RSA Examinations/Cambridge ESOL 2005b):

> Although learners don't have to do Teacher Assessment to be entered for external assessment, it provides both a good familiarisation with language testing and a clear indication for the teacher of whether a learner is ready to be entered for external assessment.
>
> The tasks are all related to the 'Can Do' statements of the Languages Ladder. As there are tasks at each grade and learners are assessed on the skills needed to fulfil the 'Can Do' statements, Teacher Assessment provides good preparation for Asset external assessment.

At the same time the informal nature of the Teacher Assessments was pointed out:

> Centres should note that, since it does not involve external assessment, a Grade Award cannot be regarded as a formal qualification within the National Qualifications Framework. It does nevertheless provide
> - a short term motivational goal for students;
> - a good measure of students' attainment for use in a range of informal contexts.

Adopting this approach also made the development of the Teacher Assessments a significant task, given that 17 grades of Teacher Assessment for each of four skills implies 68 separate assessments per language (for those languages which covered all levels).

Alternative ideas were explored but not pursued. Early consideration was given to using a portfolio approach, allowing the teacher and/or student to accumulate material and record progress. The model had been extensively used in vocational qualifications, and could have incorporated test templates for teachers to use. Portfolios were seen as motivational and as encouraging learners to take responsibility for their learning. However, feedback from teachers suggested that their use was administratively burdensome. Additionally, the portfolio approach would mean that the Teacher and external assessment strands would differ substantially, which was not considered desirable as it militated against the coherence of the scheme.

5.6.2 The test material – assessment packs

The approach which was developed was based on the use of Teacher Assessment packs. Packs would be developed at each stage (e.g. Breakthrough French) and contain a series of tasks for each Teacher Assessment skill and grade (e.g. Breakthrough French Writing level 2). These tasks related closely to external assessment tasks ensuring coherence between the two forms of

assessment. This meant that the Teacher and External strands were closely integrated and that the Teacher Assessment tasks could be seen as having a formative role while also preparing learners for the external assessments. Tasks also closely matched the corresponding Can Do statement of the Languages Ladder.

Each pack provided a series of tasks that the teacher could administer to learners at an appropriate point in the curriculum, and when learners stood a good chance of success. The majority of tasks were to be administered exactly as they were, while there were also some tasks which could be adapted by the teacher using a specific template. Adapting such tasks was an option for teachers to use if they wished to, and detailed instructions were provided for how adaptation should be done. It was expected to be particularly useful, for example, where teachers might adapt a task to use lexis to that they were currently teaching in class.

An initial proposal was to have five tasks for each skill and grade of Teacher Assessment: three specified tasks plus two others of the teacher's own making, which could then be validated. This would generate a wider pool of shared material that would involve the teachers in its creation. This imaginative idea was quickly judged impractical, and when the first assessment packs were created in September 2005, the format had been settled, at three tasks for listening, four for reading and two for speaking and writing.

The Teacher Assessment pack provided for each language and stage was a substantial document delivered in a weighty file. It contained:

- a set of photocopiable masters of the tests
- two versions of the tests, to allow for resits
- a set of instructions on how to conduct teacher assessment including the pass mark
- a recording sheet to record learners' progress
- templates enabling teachers to create their own tasks (in later versions of the packs).

A later option for Teacher Assessment involved working with course book publishers to endorse certain tasks for use in Asset Languages Teacher Assessment. This was in addition to the option for teachers to create tasks themselves.

5.6.3 Administration

The administration instructions attempted to minimise the impact on the classroom, while preserving a degree of assessment rigour. Accredited teachers undertook to provide security and consistency of administration by ensuring that work submitted was the learners' own, that the environment

was free from distractions, that the content of a question was not known in advance, that test papers were kept secure, and so on.

Tasks were designed to be administered in normal class time, and many of them could be used as whole-class or group activities. This was permitted as long as each candidate could be assessed on their language ability, in line with the above requirements. There were no time limits for completing a task, though this should be within a single sitting. The teacher might assist by explaining instructions, and use the example provided in each task, but not offer further help. Rules also allowed for resitting, reflecting the idea that all students should be enabled to complete the test successfully, sooner or later.

5.6.4 Accreditation and training for Teacher Assessment

At the start of the project, two key roles of 'centre co-ordinator' and 'accredited teacher' were identified, with requirements for training. It was the accredited teacher whose responsibilities included ensuring that standards of teacher assessment were maintained. Initially, only accredited teachers were able to purchase Teacher Assessment Packs, deliver Teacher Assessment tasks and award certificates to students. Accreditation also enabled them to carry out the Speaking component of external assessments. The early training packs for teacher accreditation included sections on adapting and creating tasks, marking writing to the agreed standard, and external assessment of speaking. According to the regulations: 'in order to gain accreditation, a teacher must have satisfied OCR of their ability to administer assessment and apply mark schemes according to the standards laid down by OCR.'

The accreditation process was made as light and accessible as possible, but would still prove a serious impediment to the adoption of the scheme by schools. A survey in 2006 found a significant proportion (25%) of centres concerned that accredited teacher training took too long to complete. Significant dissatisfaction with the approach was revealed. It was clear that many teachers did not complete the accreditation process. New proposals agreed in 2007 attempted to remove this barrier to entry by greatly reducing the complexity of the process. There were also undertakings to increase support for teachers through a web-based forum, low-cost twilight events (after-school sessions) and full-cost additional training.

5.6.5 More support needed for teachers

In practice, while teachers generally welcomed the empowerment given by increased teacher assessment, they did not welcome the associated increase in workload, and were unhappy with the freedom offered, which was interpreted rather as lack of guidance. When faced with the freer, functional approach of Asset Languages many teachers became anxious and asked

for schemes of work and defined vocabulary to be provided. This created a dilemma for the project team. It was clear that the Asset Languages scheme offered clear pedagogical opportunities for UK schools which were welcomed by many. The ability to explore language through a greater focus on purposeful communication, rather than on rote learning, was a fundamental benefit of the Asset Languages approach. But many teachers, after years of following a content-heavy curriculum, were ill-prepared for such a change.

It was perfectly possible to take Asset Languages assessments after following a GCSE course, assuming that the course content would provide adequate preparation for the functional requirements of the stage. Indeed, this was the approach taken by most of the early adopters of Asset Languages, who were positively disposed to making the scheme work. For external assessment this worked well enough, but a problem with the Teacher Assessment packs was that the tasks did not match their schemes of work. Teachers were encouraged to adapt the tasks to an extent, but most were unwilling or unable to do so. Moreover, the second wave of adopters (the so-called early majority) were more cautious and wanted to rewrite their schemes of work before embarking on the course. Many expressed concern that there was insufficient support from OCR to enable them to make the leap to the new curriculum. How could they prepare their candidates for examinations if they did not know the precise topics and vocabulary? They were sceptical of the claim that the tests took their concerns into account – that the scheme would 'fit in with any course'.

The Nuffield Inquiry had referred back to the graded objectives movement, as a successful grass-roots initiative which could be revived with a new focus on the Languages Ladder. In the intervening years the weight of the National Curriculum had perhaps killed off the grass roots; certainly it was not easy to mobilise the degree of creative engagement which the scheme had counted on.

Throughout the project, attempts were made to create local support networks for centres. At an early stage, agreement was reached with the National Centre for Languages (CiLT) to create a regional support network through their Comenius centres (Comenius is an international education programme sponsored by the European Union, which reaches out to schools, colleges and local authorities). The success of these was variable, partly because not all Comenius centre managers had bought into the Asset Languages concept. Nevertheless some centres did offer useful training, but as the National Languages Strategy unfolded, it was local authorities who increasingly became involved with local training of language teachers. Many schools were members of local organisational structures – either through the Strategic Learning Networks or through other informal groupings.

In 2007, the OCR Marketing department began to set up a framework of 'cluster groups' to encourage these groupings and attempted to provide

resource to help with local issues. The resource was normally to be provided free of charge, provided at least five centres could guarantee attendance. A fee was payable for smaller groups. Work with OCR Training also took place to create a series of twilight meetings to support centres.

One solution that was investigated but not pursued was online training. Whilst email support was provided, it proved difficult within a tight timescale to identify a delivery platform for online training that could be supported through the corporate information management programme. Finally centres were kept up to date with developments of the scheme through a regular newsletter each school term.

In retrospect, the Asset Languages approach might have been much stronger had it been developed in partnership with a major curriculum player – either one of the professional associations or CiLT, the National Centre for Languages. This was proposed to the Department for Education and Science, who argued that schools should be allowed the autonomy to develop their own approaches. This view was understandable, but it was perhaps naïve to assume that the majority of schools had the will or the resource to create schemes of work from scratch. The support networks mentioned above were a final attempt to plug this gap in provision.

Further ways of supporting centres were explored. Collaboration with publishers moved towards encouraging the development of schemes of work which supported Asset Languages assessments. Further research was undertaken to provide case studies of centres of excellence and samples of lessons. Work with local authorities increased with a view to including curricular issues as part of training for centres. A contract was agreed with the Department for Education and CILT to develop materials for primary schools.

Meanwhile, Asset Languages attempted to state its position clearly to prospective centres:

- Asset Languages does not endorse a single scheme of work and will therefore not create one that might be considered to be 'recommended'.
- Asset Languages will identify and publish examples of good practice.
- If Asset Languages materials do not match schemes then centres should adapt the Asset tasks and not their schemes (in the first instance).
- Asset Languages may be used as a vehicle for change over time, but it is not necessary to make major changes in order to use Asset Languages. However, in the light of experience in using the tasks centres may wish to adapt their schemes.
- Preparation for the exam should be about encouraging a problem-solving approach to language teaching. Often teachers report that 'a task is too hard' when students use other strengths to tackle it. Students actually like the assessments.

Thus the form of Teacher Assessment which it was practical to implement did not adequately address the needs of the teachers upon whom ultimate success depended. This was partly because the model expected more creativity in constructing a bridge between the assessment and the curriculum than the average teacher was willing or able to supply. In a context where teachers were accustomed to tightly prescribed curricular objectives, the loose match between classroom work and Asset's Can Do objectives presented real problems.

In consequence, the Teacher Assessment tasks as they were generally used were not particularly inspiring and in practice it might be concluded that they did not really achieve the aims of formative assessment as this concept was understood by the development team.

At the same time we should not forget that many successful adopters of the scheme were enthusiastic about the motivational value of the teacher-awarded certificates. These grade awards contributed to articulating the progression of small positive steps which was the central metaphor of Asset Languages and the Languages Ladder.

5.7 Conclusions: The lessons of Asset Languages

Writing in June 2013, the Asset Languages home page hosts the following statement:

> Following a review and consultation in the autumn of 2012, we have decided to redevelop qualifications in French, German, Italian, Mandarin and Spanish. These revised qualifications will be available at Breakthrough, Preliminary and Intermediate stages for first teaching from September 2013. The first assessment series will be in June 2014.
> This means that the four series in 2012–2013 are the last assessment opportunities for Asset Languages qualifications in their current format. These assessment series are also the final opportunity for Advanced stage assessments in any languages and for all assessments in the following languages:

Arabic	Japanese	Turkish
Bengali	Panjabi	Urdu
Cantonese	Polish	Welsh
Cornish	Portuguese	Yoruba
Greek	Russian	
Gujarati	Somali	
Hindi	Swedish	
Irish	Tamil	

This development is not unexpected. A message to the OCR partnerships group in July 2012 explained:

> While OCR recognises the hard work and dedication of students and teachers who deliver Asset Languages, we are forced to re-assess our commitments, particularly since changes to school performance measures and funding policies are likely to impact on the demand for these qualifications.

In September an OCR email update provided more detail on the seriousness of the problem, illustrating the low entries for some Asset Languages compared with their GCSE counterparts, and stating that 'all the evidence suggests that changes in policy relating to accountability measures and funding signal that uptake will decline considerably from this already low base'. The reference to changes in policy refers to the withdrawal of performance points for schools submitting Asset Languages candidates, meaning that schools would get no credit for preparing students for these exams. This effectively signals the end of the road, at least for the major part of the system. Asset Languages won a keen following, but among a fairly narrow group of users. What it was unable to do was compete for mass market share with the established GCSE and A Level qualifications (it was, after all, only a voluntary alternative).

There are important lessons to be drawn from the demise of Asset Languages, although first it is important to place it in context. Over the period covered by the development and deployment of the Asset Languages scheme language learning in England has come under increasing pressure: Asset Languages is just one of the casualties.

Asset Languages was born out of the reforming initiative of the National Languages Strategy. That initiative has also run its course, having been discontinued by the government which took office in 2010. Recent figures show that in England the proportion of students sitting GCSEs in a foreign language fell from 78% in 2001 to just 43% in 2011 in the wake of the decision, in 2004, to make languages optional (see Section 5.1.3). In Wales, foreign languages have never been compulsory in secondary schools and uptake of language GCSEs is the lowest in the UK, representing just 3% of all GCSE subject entries (The British Academy 2013:7).

The same report shows that languages have again become elitist subjects: 'Nearly a third of linguists in Higher Education come from independent schools (while only 18% of the post 16 school population attend these schools), and in state schools just 14% of children eligible for free school meals obtained a good GCSE in a foreign language compared to 31% of other state school pupils' (The British Academy 2013:7). The report concludes that a weak supply of language skills is pushing down demand and creating 'a vicious circle of monolingualism'.

The results of the ESLC (see Section 6.8.1) underline the state of language learning in England. England performed very poorly in the ESLC, with 30%

of students not even achieving A1, and little more than 20 % achieving higher than A1. Those are the figures for the first foreign language (French); results for the second foreign language (German) are, as in most countries, somewhat lower than that. Considering that the students surveyed were sampled from that group of secondary school students in England who actually study a language (less than half), these are alarming figures.

Asset Languages was unable to avert the crisis which has overtaken language learning in England. And yet the survey outcomes make all too clear the shortcomings of the current exam regime to which Asset Languages offered an alternative. An unpublished study (Jones and Benton 2012) compared the CEFR levels with the GCSE exam grades which the England cohort went on to achieve the following summer. These look good and, indeed, were welcomed as 'impressive, and above the national average', by a teacher's association (Association for Language Learning 2012). This begs the question of whether the communicative competence which the ESLC set out to measure is what English students are learning, or what the GCSE is testing. To the extent that this comparison of GCSE grades and CEFR levels is valid it makes the case strongly for what Asset Languages attempted but failed to do: to focus attention on languages as communication tools, and on setting objectives and reporting results in meaningful terms.

Things can only get better, and there are indeed signs that languages are again being taken seriously. While it does not currently appear likely that compulsory language education at Key Stage 4 (age 14–16) will be reintroduced, the government has deplored the current situation, stressed the importance of languages, and introduced an additional reporting measure to encourage schools and students to include a modern or ancient language among their GCSEs. A useful source of information on current initiatives and policy discussions is the website of The Languages Company (The Languages Company 2013), launched in 2008 by Lid King, who as National Director for Languages in England was in charge of the National Languages Strategy from 2003 to its demise in 2011, and a key figure in promoting the Asset Languages approach to assessment. So Asset Languages was launched in unpropitious circumstances, but if we are to learn lessons from the project it is important to look in detail at the contributory factors.

There is the design of the product. The formative aspect was central to the rationale, but its specific implementation, given the numerous constraints and challenges discussed in detail above, was found difficult to work with by many teachers. It was clear during the development that using the existing technical and administrative infrastructure available to the Cambridge English and OCR exam boards to implement a different kind of assessment would set practical limits on what could be offered. Let us concede that there were certain substantive weaknesses in the design.

At the same time it is possible to see Asset Languages as a basically good

product that too few people actually wanted; and we should ask why that is. As a framework for promoting good language learning it had many of the promising features called for by the Nuffield Inquiry and endorsed by the National Languages Strategy: focus on language for communication, meaningful criterion-referenced levels referenced to the CEFR, small steps of achievement, moving through graded objectives, individualised testing when ready, provision for different profiles of skill, an explicitly formative dimension, and so on. Why were these unique selling points not more attractive?

As the England results in the ESLC so clearly demonstrate, the stress on exam grades as indicators of success in language learning has led to the neglect of communicative language proficiency as a goal. But this stress results directly from the government's use of exam grades for accountability purposes. Where teachers and school heads are judged on exam grades, there is no incentive to focus on communicative proficiency as an outcome of language learning. What school heads need are clear specifications which teachers can teach to and students can achieve, by rote learning if necessary. In this climate teachers will not adopt an optional alternative which involves extra work and actually makes test preparation more difficult.

Furthermore, why would schools invest the extra effort in individualising instruction and testing-when-ready? Following a languages ladder of progressive steps is logistically more complex than focusing all attention on a single big-bang test at the end of secondary. Members of the Assessment Reform Group, which promoted formative assessment over a 21-year period until it disbanded in 2010, conceded at that farewell event (Cambridge Assessment Network 2010) that formative assessment is in fact very difficult to do well – too difficult for most teachers. That is a sobering conclusion, indicating that effective teacher training would be a critical element in any successful approach.

Asset Languages did not succeed in its attempt to focus attention on the purposeful use of language for communication. It proposed a different set of educational priorities, but was ultimately unsuccessful in communicating these new values to potential users – teachers and school heads. It demonstrates that on its own a reforming assessment scheme cannot make a difference. Successful educational innovation must integrate coherently curriculum design, teaching practice and assessment of outcomes. Reform which does not encompass the whole system, or which does not communicate its values effectively, is unlikely to succeed. Explaining the results of the ESLC to UK audiences has provided several opportunities to present the above message. But these events have also brought me into contact with many champions for languages in the UK, who remind us of the importance of positive thinking. As Lid King, former head of the discontinued National Languages Strategy, puts it:

> In the absence of a Strategy we therefore have to support positive policy initiatives and to find possibilities for future engagement, learning lessons from both the successes and the failures of the past . . . Firstly we need a clear articulation of our vision for languages, a vision which can be understood and supported by many people, in policy, in schools and in society. This means both understanding and demonstrating how our languages agenda relates to broader educational, social and economic goals. Secondly we need a real desire and commitment to work together in order to realise that vision . . . In the absence of a centrally funded strategy, the way forward is through joint work, collaboration both locally and nationally, making the best of available resources (King 2011).

And there is indeed much work and collaboration, involving a multitude of groups, campaigns and projects. The websites of The Languages Company, the Speak to the Future campaign, or The Association for Language Learning are a good starting point for exploring the range of current activities. There are also many international projects to which UK partners are contributing, such as Language Rich Europe, or LUCIDE – Languages in Urban Communities – Integration and Diversity for Europe.

And while the objectives of the new National Curriculum do seem ambitious, given the current levels of achievement revealed by the ESLC, ambition is surely preferable to acquiescing in failure. The Curriculum states:

> GCSE specifications in a modern language should enable students to:
>
> - develop their ability to communicate coherently with native speakers in speech and writing, conveying what they want to say with increasing accuracy
> - express and develop thoughts and ideas spontaneously and fluently
> - deepen their knowledge about how language works and enrich their vocabulary in order for them to increase their independent use and understanding of extended language in a wide range of contexts
> - acquire new knowledge, skills and ways of thinking through their ability to understand and respond to a rich range of authentic spoken and written material, including literary texts
> - develop awareness and understanding of the culture and identity of the countries and communities where the language is spoken
> - make appropriate links to other areas of the curriculum to enable bilingual and deeper learning, where the language may become a medium for constructing and applying knowledge
> - develop language learning skills to prepare them for further language study and use in school, higher education or employment (Department for Education 2013).

Who could object to any of that?

I remain convinced that the role of assessment is a critical one, and that

the commendable objectives listed above will not be achieved until the central role of communication in language learning is reflected in exam practice and exam preparation. Asset Languages illustrates the problems, but also provides a clear model for how this can be done, employing the multilingual scaling approaches described in this volume. These can provide practical, reliable, and above all demonstrably concrete and convincing measures of the useful language skills specified in the above curriculum objectives.

6 The European Survey on Language Competences: Informing language policy

6.1 The significance of the Survey for Cambridge English

In June 2012 the first European Survey on Language Competences (ESLC) published its findings, bringing to an end a complex 4-year project delivered by a multinational consortium of partners and administered with the assistance of national research co-ordinators in 16 countries or regions. The project's sponsor was the European Commission, and Cambridge English Language Assessment was the contracting partner with the Commission.

The consortium, named SurveyLang, brought together a number of ALTE partners: Cambridge English Language Assessment, the Centre International d'Etudes Pédagogiques (CIEP), the Goethe-Institut, the Università per Stranieri di Perugia, the Universidad de Salamanca, and the Instituto Cervantes. The National Institute for Educational Measurement (Cito) together with Gallup Europe approached Cambridge English Language Assessment with a proposal for participating in a joint bid, and finally these two partners joined the SurveyLang consortium.

The survey was the first such major European project that Cambridge English Language Assessment has participated in. It provided a new context for applying the well-developed theoretical models and the operational experience of working with multilingual frameworks accumulated in the previous two decades. As a major collaboration between partners having highly specialised skills, overall co-ordination of whom lay with Cambridge English Language Assessment, it was a challenging but also a valuable learning experience. As a research study the survey has produced interesting data on factors which impact on language learning. The findings are of considerable relevance to important areas of development for Cambridge English, specifically Learning Oriented Assessment (LOA – see Section 7.4), and studies of the impact of Cambridge English exams in particular educational contexts.

This chapter offers a broad narrative account of the survey, focusing on those aspects most relevant to the theme of multilingual frameworks which is the subject of this volume. Outcomes of the survey – that is, the language test

results and the contextual questionnaire findings – are presented, and interpretations offered of potential value for contributing to Cambridge English Language Assessment's approach to the development and implementation of language education policy.

6.2 Background to the survey

SurveyLang came together under the central co-ordination of Cambridge English Language Assessment, who was the signatory to the contract with the European Commission. A call to tender was issued by the Commission in March 2007 and by June a multinational team was working intensively to assemble a bid.

The invitation to tender had long been expected. In 2002 The European Council in Barcelona had called for further action to 'improve the mastery of basic skills, in particular by teaching at least two foreign languages from a very early age', and for the 'establishment of the linguistic competence indicator'. This decision was motivated by a lack of data on the actual language skills of pupils, and the need for reliable ways of measuring progress towards this new objective.

In 2005 the Commission outlined detailed requirements for a European survey to collect the data necessary to construct a European language indicator, and stressed that it should be carried out as soon as possible (European Commission 2005). Requirements included:

- the survey should cover tests of first and second foreign language competence in the five most taught official European languages in the European Union
- from a representative sample of pupils in education and training at the end of International Standard Classification of Education (ISCED) level 2 (or from ISCED3 if a second foreign language is not taught before)
- test scores should be based on the scales of the CEFR for languages
- the indicator should measure competence in the three language skills most readily testable (i.e. listening comprehension, reading comprehension and writing).

In preparation for the invitation to tender the SurveyLang brand was created and copyrighted, bringing together the group of ALTE members responsible for the language tests, as well as Cito, followed by Gallup Europe, who joined the consortium to take responsibility for the sampling, questionnaire construction and analysis. The contracting partner remained Cambridge English Language Assessment, who was responsible for overall management of the project. A programme board was to be chaired by the Chief Executive Mike Milanovic. The Project Director was to be Norman Verhelst, of Cito.

Verhelst's retirement was expected two years later, when Neil Jones would take over as Director and Jan Wiegers of Cito would take over as chair of the programme board.

A project office was set up in Cambridge, with Karen Ashton as Project Manager. The office was to manage the work of the consortium, and also a large part of the interaction between the consortium and the participating countries, each of which provided a team for administering the survey within the country. A Basecamp website was set up to provide an efficient means of communicating with countries and monitoring progress. Cambridge had also assumed responsibility for co-ordinating the language test construction across the five language partners, a task which was managed by Martin Robinson. The specific work of constructing the English language tests engaged further managers within Cambridge English Language Assessment, co-ordinating the work of a team of external item writers.

6.3 The tender: Language tests and the CEFR

The tender document constructed by SurveyLang provided detailed proposals concerning the approach to the language tests. This was based on the definition of language use offered by the CEFR (Council of Europe 2001:9, see also Section 2.1), and it explained that:

> This model needs relating to the learning context of the 14–16 year-olds who are the objects of the survey, whose communicative language competence may often be the outcome of formal classroom study rather than exposure to the language in a real-life setting. In this case the above-mentioned *conditions*, *constraints*, *texts*, *themes*, *domains* and *tasks* must be understood not as emerging from the daily exigencies of life, but rather as parameters which are carefully selected and manipulated in order to provide a supportive context for learning, the major determinant of progression being, naturally, linguistic.

SurveyLang would focus on those parameters of the above action-oriented approach which relate most strongly to level:

- the communicative tasks to be accomplished
- the range of topics, in the sense of a progression from the immediate and personal through routine and familiar to increasingly unfamiliar and abstract at the highest levels
- the language activities, language processes and strategies which these elicit.

In order to measure language skills validly for 15–16 year-old learners test tasks would:

- engage cognitive skills expected for the age group
- not depend on knowledge of the world which cannot be assumed
- use topics which are relevant and engaging for that age group
- test language functions in contexts relevant to the age group.

Weir's (2005a) socio-cognitive model was presented and its utility explained.

The omission of speaking from the tested skills was, of course, regrettable but reflected the judgement of the Advisory Board that the logistic difficulties and expense of administering a Speaking test made it impractical at least for the first round of the survey. It was retained as an objective for a future round. There was general agreement that testing speaking would require the support of CB systems. The European Commission held a symposium on this issue, the proceedings of which were published (see Galaczi, 2010, Kenyon and Malone, 2010, Van Moere 2010).

The tender paid particular attention to the concept of bias, given the differences across the countries and educational systems encompassed by the survey. The distinction between construct-relevant and construct-irrelevant sources of variance was explained:

> Italian candidates will have an advantage reading a French text over, say, Polish candidates, because of the greater areas of similarity of French to Italian. This is just a fact of life, not a validity problem. However, a French test item that can be answered by recognising a single word which is a cognate of an Italian word should be considered biased, because it can be answered using a construct-irrelevant strategy.

Potential sources of bias included: knowledge of the world, reference to themes which might be considered sensitive in certain participating countries, differences in familiarity with/availability of computers, differences in familiarity with task types, and curricular differences.

Approaches to identifying or avoiding bias in such cases were proposed. In the case of curricular differences the tender actually undertook to complete a study of curricula and course materials in all participating countries prior to finalising the selection of test tasks; something which subsequently slipped the consortium's memory. However, the tender was explicit in stating that:

> There remains the possibility that the survey may identify countries where current language pedagogy is less attuned to the communicative, action-oriented approach assumed by the CEFR. This may lead to lower levels of performance in such countries. Such outcomes would not be considered biased, but entirely valid in the terms of the survey.

Concerning the requirement to link to the CEFR, the tender explained that conceptually, the survey would be composed of five quite independent

surveys, one for each of the target languages. However, it was critical for the purposes of the survey that results for each language could be interpreted within a single CEFR frame of reference. Thus standard setting must include the verification of standards across languages. The tender went into some detail on the range of activities that would be involved in the setting of comparable standards across languages. It noted that to date applications of the pilot manual (Council of Europe 2003) had largely involved single languages. The survey would provide an important opportunity to make progress in establishing comparability of standards within Europe. This required the introduction of an explicitly cross-language dimension into standard-setting activities, for example by:

- using trained plurilingual panellists to make judgements on items belonging to two different languages
- using trained plurilingual raters to rate performances of students in two languages
- identifying students with plurilingual competence to take tests in two languages and to self-rate themselves via Can Do questionnaires with respect to both languages (see Section 3.2.4.2 for such a study done earlier in the context of a multilingual computer-adaptive test).

These proposals were innovative in terms of standard-setting practice to date, particularly the procedures recommended by the *Manual for relating language examinations to the CEFR* (Council of Europe 2009). Not all of them were in practice implemented, as will be discussed under standard setting (Section 6.7)

Each of the work areas was the subject of a chapter in the tender, and an appendix contained sections presenting among other things the SurveyLang partners' long-standing connection with the development of the CEFR, and an extended presentation of the constructs of reading, listening and writing, incorporating work on the series of construct volumes in the SiLT series (Khalifa and Weir 2009, Shaw and Weir 2007, for reading and writing, respectively).

For the most part, the tender's proposals resemble quite closely what happened in practice. Also interesting in hindsight are some of the elements that were volunteered but never required. The services of a European-wide network of support centres were promised (having in mind the offices of ALTE members). Cambridge ESOL's logistics centre at Duxford was depicted dispatching test papers to the four corners of Europe. The language partners were to undertake the marking of writing where requested by countries (for an extra fee). This last proposal was incidentally one which several countries, when providing feedback on their experience of the survey, suggested should be the default in a future round. These visions reflected the experience and expectations of the language partners, who in the course of

the project would come to understand better the specific nature of language testing within an international survey.

One such difference is that the administration of the tests, along with all other aspects of the survey, is devolved to the individual participating countries, each of whom must establish a national research centre to co-ordinate operations. Management of field operations, as this area of the survey is known, is critical to success. The tender suggested that field operations would be co-ordinated by a partner with extensive experience in this area, which would have indicated Gallup Europe. In the event however field operations would be picked up by Cambridge ESOL, as another work area to be co-ordinated from the SurveyLang office in Cambridge.

The contract was duly awarded to SurveyLang, but negotiations put back a possible start on the project by several months. A kick-off meeting with the Commission and their experts finally took place in February 2008. Following the successful submission of the tender, three more reporting stages would be required by the Commission: an inception stage, an interim stage, and a final stage, where a final report and technical report were to be submitted. Payment was to be made by instalments at each stage, subject to the reports being accepted. The following sections refer to these reports and other documents to give an account of each work area, focusing on those areas most relevant to the theme of this volume.

6.4 Language test development

Robinson (2013) describes the processes adopted to develop the language tests, emphasising aspects considered particularly innovative. The European Commission specified The CEFR (Council of Europe 2001) as the framework against which to measure language learning outcomes for the survey, reflecting the widespread impact which this document has had since its initial publication. The language tests developed for the survey set out to reflect the CEFR's action-oriented, functional model of language use, while ensuring relevance for 14–17 year olds in a school setting. The socio-cognitive model adopted was based on the CEFR's model of language use and learning (see Section 2.1). To enable the resulting test construct to be implemented comparably across languages, these abilities were mapped to specific task types, drawing chiefly on task types which had been used successfully by SurveyLang's language partners in their operational exams.

6.4.1 Test content and abilities to be tested

Specification of test content referred to the domains of language use proposed by the CEFR (Council of Europe 2001:43–100). As the CEFR stresses, these categories are illustrative and suggestive, rather than exhaustive. However,

the listed elements provided a useful starting point for selecting appropriate content.

The CEFR identifies four basic domains of language use (personal, public, educational and professional) illustrated in terms of situations (e.g. the locations in which they occur), communication themes (e.g. daily life) and topic-specific notions (e.g. family celebrations and events, relationships, etc.). Considering the relevance of each of these domains to language learners at particular proficiency levels informed a decision on what proportion of tasks relating to each domain mentioned above should be used at each tested level of the survey (Table 6.1 below). Thus for example personal themes dominate at the lower levels, while public themes are used more at higher levels.

Table 6.1 Domain distribution across Levels A1–B2

	A1	A2	B1	B2
Personal	60%	50%	40%	25%
Public	30%	40%	40%	50%
Educational	10%	10%	20%	20%
Professional	0%	0%	0%	5%

Language functions (e.g. imparting and seeking information) are discussed in the CEFR as an aspect of pragmatic competence, providing a general rather than setting-specific taxonomy of language in social use. Together these communication themes, notions and functions provided the basis for categorising and selecting texts for use in the survey. The choice of test content also took into account the characteristics of the target language users, i.e. the 15–17 year old students participating in this survey. To ensure adequate coverage domains and topics were assigned to tasks at the commissioning stage.

6.4.2 Task types

The socio-cognitive validation framework proposed by Weir (2005a) (Section 2.1 above) complements the CEFR's treatment of the cognitive dimension and provides useful practical models for characterising progression across levels. To ensure comparable implementation of the resulting test construct across languages, each skill and level was systematically mapped to specific task types, drawing chiefly on types used successfully by the consortium's language partners in their exams.

For reading and listening it was decided to use only selected response types, for ease and consistency of marking. A repertoire of multiple-choice and matching tasks was specified (see Appendix D). For writing a range of open, extended response task types was used in keeping with the CEFR's

action-oriented, communicative, functional model of language use, e.g., writing an email, postcard or letter, or writing a referential or conative text (intended to inform, persuade or convince).

A particular concern of the Advisory Board was that students should be sufficiently familiar with the task types. In fact, evidence from trialling and pretesting suggested that students had no real problems in understanding how to respond to the test tasks in their PB form. The instructions included in the PB and CB tests were also rendered into the students' first language. The provision of additional on-screen help in the CB mode was thus felt to be unnecessary. Instead, familiarisation material was made available to students or teachers who wished to make use of it, but not as a compulsory part of the test administration. These materials were provided by the National Research Co-ordinators to all participating teachers and were also available on the SurveyLang website.

For the Main Study both PB and CB familiarisation materials were available. The sample CB tests on the SurveyLang website enabled the student to choose a language to be tested in, as well as the language for the on-screen instructions.

6.4.3 Development phases

To develop language tests whose results would be comparable across the five languages close collaboration was required between the partners in the language testing group, and the adoption of shared processes.

Having completed draft specifications and a set of draft task types, the first development stage was to pilot these, gathering feedback from all stakeholders: the European Commission, the participating countries, teachers and students. In 2008 over 100 tasks were piloted in schools made available by the participating countries. Agreement was reached as to the most appropriate task types, in terms of familiarity and transparency, and the test specifications were finalised.

The language team then produced over 500 tasks, comprising more than 2,200 items, which were trialled through both the pretesting stage in 2009, and the Field Trial in 2010, before the best-performing items were selected. For the Main Study in 2011 143 tasks comprising 635 items were used across the five languages.

6.4.4 Test construct, specifications and item writer guidelines

Common test specifications ensured that tasks across languages were almost identical in terms of number of items, number of options, text length, etc.

Detailed item writer guidelines were developed for each of the three skills. These guidelines specified the requirements of each task type at each level in

terms of overall testing aim, testing focus, level of distraction in the options, input text length, etc. They also provided explicit guidance on the selection and manipulation of text types and topics, and the production of artwork and recordings. Quality criteria relevant to each task type were listed and these criteria provided the basis for the acceptance, rejection and editing of tasks as they proceeded through the item production process.

The test items were developed by an expert team of over 40 item writers distributed across Europe, who worked according to specifications and guidance provided by the central project team. Items were moved through various stages of a predefined life-cycle including authoring, editing, vetting, adding of graphics and audio, pilot-testing, Field Trial and so on. Each stage involved different tasks, roles and responsibilities.

6.4.5 Test production process and comparability across languages

The successful delivery of the language test instruments required a shared, collaborative test production process, and achieving comparability across the five languages additionally required some innovative procedures. The usual stages of item writing, editing, pretesting and review were followed, but what made this process unique were additional stages of targeted commissioning, cross-language vetting and cross-language adaptation.

As noted above, a much greater number of items than required for the Main Study were commissioned, in order to allow for selection at the pretest and Field Trial stages. Given the large number of item writers commissioned, it was important to plan for adequate coverage of construct, domains and topics for all tasks at each level across the five languages. Each item writer therefore received a detailed commissioning brief specifying the task types, levels and topics to ensure adequate and consistent coverage of the CEFR domains. The work of creating these tasks was divided among the language partners according to the strengths of each item writing team. For some languages, item writers specialised in certain skills, levels or task types. Item writers were organised into teams and managed by team leaders and specialist language testing product managers.

Initial item writing was followed by cross-language task adaptation. This was an innovation for the language partners which served several purposes. It provided a valuable context for developing collaborative working methods: studying each other's tasks in detail stimulated much critical reflection and interaction. It was seen as a possible way of demonstrating consistency and comparability across languages. Lastly, it offered a straightforward, if not a quicker, way of generating new tasks.

Task adaptation worked as follows. In the Pilot Study, a proportion of the tasks were adapted across languages. Each language partner was asked

to adapt some tasks from two of the other four languages. The Pilot Study review confirmed the value of adapting tasks across languages. It appeared that most task types used in the Pilot Study could be successfully adapted from one language into another, as long as adaptation were seen as more than simple translation. The process needed skilled item writers who were not only competent in two or more of the languages, but also had a comprehensive understanding of the CEFR and its language descriptors. Item writers needed to be aware of lexico-grammatical differences between the languages and how these differences might affect the difficulty of the items.

The only task type that appeared difficult to adapt was the multiple-choice cloze task where the testing focus was largely lexico-grammatical. It was discovered that adaptation was most practical at the lower levels and although possible with some higher-level task types, the longer texts involved meant the extra effort required tended to outweigh the benefits. For the skill of writing, no significant difficulties were encountered with adapting any of the task types.

Thus it was decided to adapt all the writing tasks at all levels and all reading and listening tasks at A1 and A2. This was taken into account at the commissioning stage where each partner only needed to write a proportion of the required writing tasks and reading and listening tasks at A1 and A2.

Cross-language vetting was another important innovation in the survey test production process. Tasks from each language were vetted by at least two other language partners. A vetting form was created to enable a consistent electronic record of comments. Vetting comments were passed back to the original language partner who could then study the comments of their own and their partners' vetters. This approach was trialled during the Pilot Study and a review conducted at the end of that stage confirmed its value, both as an additional quality control, and also as a way of sharing knowledge and experience among the language partners (Perlmann-Balme 2013).

6.4.6 A targeted approach to test administration

It was evident from the outset that in order to test efficiently over the range Pre-A1 to B2 a targeted testing approach was essential. That is, students should be given a subset of tasks appropriate to their level – not too hard and not too easy. This would provide better measurement, firstly because items having very high or low facility provide little information, and secondly because it should improve the validity of students' responses by avoiding frustration or boredom.

Such a targeted approach could have been quite simply implemented had the tests been entirely administered by computer, using some adaptive or semi-adaptive algorithm. However, the terms of reference already specified that both computer- and PB modes of administration were to be offered.

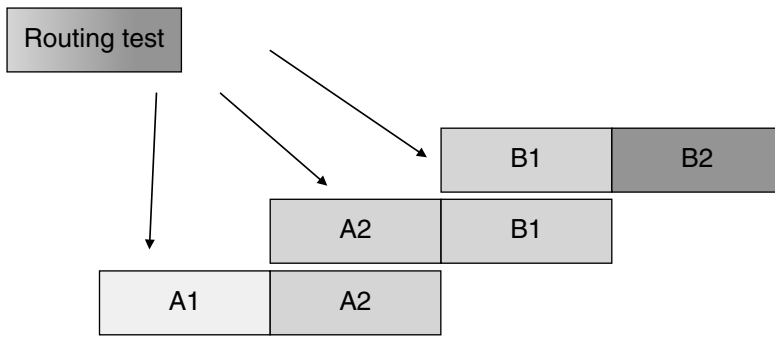
Thus the only alternative was to administer a routing test to every student included in the sample (or who potentially might be included in the sample – both options were offered to countries, to choose whichever they found logistically easier). This had to be done well in advance so that in the actual survey (the Main Study), each individual student could be assigned a test at the correct level, as well as on the basis of randomly assigned parameters (two out of three skills, the particular test booklet within a level). This added a degree of logistical complexity for SurveyLang and for the schools, but in practice worked well.

Unlike the actual survey tasks, which comprised completely new material, the routing tests were taken from language partners' existing item banks and were mostly calibrated on a CEFR-linked scale. This would help the language partners evaluate the levels of the student population to expect in the actual survey. The tests were short and simple, being limited to reading and grammar, but adequate to the purpose of assigning each student to one of three overlapping levels. There was some discussion in the Advisory Board of the possible effect of students being assigned the wrong level test in the Main Study; but SurveyLang explained that the targeted approach would eliminate a proportion of error, and therefore in any case still provide more coherent and interpretable data.

The routing test data was interpreted bearing in mind both the need to give each student an appropriate test, but also to ensure a minimum number of responses for each test booklet which was used. For some languages and levels this involved a compromise, where for example more students might receive a low-level test in order to ensure enough response data to calibrate the items at that low level.

Figure 6.1 illustrates the three-level overlapping targeted design. Each student received a test covering two CEFR levels, and the three tests overlapped each other by one CEFR level. Each sampled student was thus assigned to a test level according to the routing test, but also randomly

Figure 6.1 The targeted test design



assigned to a particular test booklet at that level. Unlike the language testing familiar to the language partners, where the focus is on the individual student, in complex surveys such as the ESLC each sampled student would only see a proportion of the total test material. The total amount of test material is determined by the need to achieve adequate coverage of the construct, i.e. to test all aspects of a skill considered important at a given level. By using an incomplete but linked design this coverage can be achieved, while each student receives only a manageable proportion of the total test material. A complex linked design of test booklets with overlapping content would also allow for the same task to be placed in different positions in different test booklets to negate any potential task order effect. Each individual test or booklet consisted of three or four tasks and lasted 30 minutes. However, the complete design ensured that the whole construct, i.e. all task types, was tested at the cohort level. The design was to be implemented in the same way in each of the five languages, as consistency of approach would maximise the comparability of outcomes. However, a simpler design was adopted for Italian, which would only be administered in Malta, where performance levels were expected to be high.

A further design constraint adopted was that the total language test time for a student should not exceed 60 minutes. A test for one skill would comprise 30 minutes of material. A student would only be tested in two of the three skills. Students would be assigned randomly to one of the three permutations.

6.4.7 Item authoring tool and item-banking system

Close and co-ordinated collaboration between the partners and consistent implementation and presentation of test tasks was made possible by the item authoring, banking and test assembly functionality of the testing tool developed for the survey by the responsible partner, Gallup Europe. The development provided an integrated and rich software platform for the design, management and delivery of the language tests. The platform was built to the specific set of requirements of the survey project, with input and user acceptance testing provided by the language partners. Support for delivering PB and CB tests was built into the design, including the facility to check the rendering of tasks on screen and on paper.

The software platform also linked into the databases which supported the survey process, so that tests individualised to each student could be produced and delivered, either in PB versions, which were printed in each participating country from DVD-ROMs, or CB versions which were distributed to countries on USB sticks, this being judged to be, if not the simplest, the most secure and robust of the possible options for delivery.

The test-item authoring tool supported a distributed and fragmented

development model. It enabled non-technical personnel to create tasks in an intuitive way by means of predefined templates for the various task types used in the survey. At any stage in the development, a task could be previewed and tested to allow the author to see how it would look and behave when rendered in a test. The authoring tool also supported the capture and input of all the metadata elements associated with a task, including descriptions, classifications, versioning metadata, test statistics, and so on.

An item life-cycle was defined on the system, including functionality to create new versions of tasks and adapt them. Adaptation allowed a task developed in one test language to be adapted as a new task in another language.

One of the most innovative features of the item bank was its ability to manage the audio tracks of the listening tasks. Creating high-quality audio is normally a time-consuming and expensive operation. Traditionally the full length track of a task is created in one go and stored as an audio file. If a change is made to this task a completely new recording is thus required. Furthermore, a test-length recording that plays each task twice together with silent pauses creates an unnecessarily large audio file. To avoid this, an audio segmentation model was developed whereby the audio files could be recorded as a series of short fragments. The various fragments were stored along with the other resources of the task and were only compiled into full-length audio tracks at the point of test assembly.

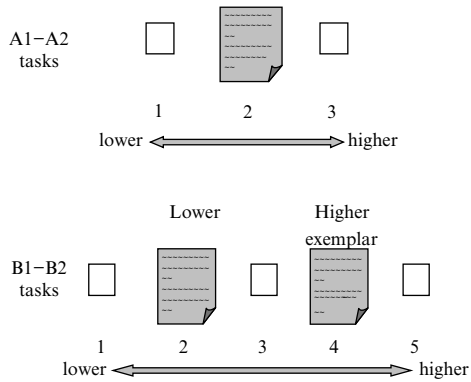
By using a shared online system for the production of the language tests, the language testing group, although dispersed across Europe, could ensure that each language team and all its members were following the same procedures at the same time.

6.4.8 The marking of writing

The approach to marking went through major revisions between the 2008 Pilot Study and the Main Study, aimed at ensuring maximum consistency, given that each country was responsible for their own marking. The basic assurance of consistency was achieved through specifying that a proportion of scripts would be multiple-marked by all markers in a country, and that a proportion of these should be returned to the appropriate language partner for verification. After the Field Trial stage a further innovation was introduced. Rather than make absolute judgements about a student's CEFR level, markers would make a comparative judgement, relative to exemplar texts. For Levels A1–A2 one exemplar defined a 3-point scale, while for the B1–B2 levels two exemplars (a higher and a lower one) defined a 5-point scale. See Figure 6.2 below.

Exemplars were chosen to elicit the widest possible range of marks, informed to an extent by evidence from the Field Trial of the general level

Figure 6.2 Marking of writing against exemplars



of the student population for each language. As explained in training, exemplars were not intended to represent a specific performance level in CEFR terms, but rather a level where a roughly equal number of worse and better performances might be expected to be produced. In other words, the exemplars were norm referenced rather than criterion referenced. This was a necessary feature of the approach, because choice and use of the exemplars could not be allowed to pre-judge the subsequent standard setting.

In preparation for the Main Study further revisions and additions were made to the design of the Writing tests, the marking criteria, training, and quality assurance procedures. The number of tasks a student responded to was reduced to three at Level 1 and two at the higher levels, aiming at quicker marking and fewer missing or partial responses. The same two criteria – Communication and Language – were used for all four test levels, to make marking quicker and easier.

Training procedures were improved, with more stress on practice and standardisation of marking. Detailed, automatically generated feedback on performance was provided in the form of an Excel spreadsheet, in order to improve the accuracy of marking. The marks for a group of trainee raters could be simply input into the spreadsheet, which provided individualised feedback for each rater in the form of a comment on their severity (from very severe to very lenient) and consistency (from very consistent to very inconsistent). All multiple-marked scripts, rather than a proportion, were to be returned to SurveyLang for central marking, aiming at more reliable comparison across countries.

Interesting and somewhat disappointing was that the innovative approach of comparative marking against exemplars did not solve the problem of differing standards across countries: the centrally marked scripts showed clearly that some countries were systematically more severe and others more lenient.

Even more interestingly, when this finding was presented to national research co-ordinators, many of them were able to predict correctly the direction in which they had strayed. In the event it was necessary to use the centralised marking to calculate statistical corrections for the in-country marks. This finding underlines the difficulty of standardising judgements, given doubtless largely unconscious, culturally determined predispositions for markers to punish error or reward achievement.

6.5 Questionnaire development

The contextual questionnaires constitute the second data collection instrument used in the survey. Their development was the responsibility of Cito. Four background questionnaires were administered: to individual students, language teachers, school principals and the National Research Co-ordinators.

The inception report presented to the Commission at the outset of the project explained the purpose of the questionnaires as follows:

> These questionnaire data allow us to detect context factors that are related to foreign language achievement and which, therefore, might be relevant for improving foreign language achievement. The context of foreign language learning differs widely between nations and exactly these differences between nations provide us with a unique opportunity to assess how differences between national system-wide factors are related to foreign language achievement. Furthermore, an in-depth analysis of the context factors related to foreign language achievement within nations allows us to discover where there is room for change and improvement given the social, cultural, educational and economic situation of each nation (SurveyLang 2008).

The first step in the development process was the development of a shared contextual framework. The starting point was an analysis of the conceptual frameworks and variables of similar international surveys. This was important to ensure that as far as possible existing key domains and constructs were identified, so that the results could be linked to other international surveys. In this way previously gathered knowledge could be best exploited, combined from the different scientific fields dealing with educational achievement and specifically foreign language achievement.

The analysis yielded five broad domains to consider for inclusion in the survey context questionnaires. First were concepts reflecting characteristics of students and their teachers, including demographic background (e.g. age, gender, mobility and mother tongue), affective constructs related to foreign language learning (beliefs and attitudes towards foreign language learning/teaching), and experiential variables (e.g. students' out-of-school exposure to foreign languages and the foreign language training of teachers).

The second key domain consisted of constructs reflecting what is actually taught in classrooms, and how: curricular and instructional practices. This domain contains constructs that reflect, among others, opportunity to learn and the student's exposure to foreign language instruction (e.g. length of instruction, whether languages are compulsory or voluntary, and other languages learned).

The three school-level and system-level domains distinguished general conditions (e.g. general affluence of the country or the school's economic resources), more specific linguistic conditions (e.g. the school's foreign language specialisation), constructs related to the foreign language staff (e.g. requirements regarding initial and in-service training), and constructs related to the official curriculum (e.g. whether the foreign language is compulsory or voluntary at the national or school level).

Based on the first exploration a shared contextual framework was developed, involving all stakeholders in the survey. There was collaboration with Eurydice (the Network on education systems and policies in Europe) to exploit available data and avoid duplicating effort.

The International Indicators of Education Systems (INES) model, used in other surveys including the Programme for International Student Assessment (PISA), provided a ready-made system of indicators for cross-national comparisons in education, and a method for ordering the key domains and variables captured by the language tests and context questionnaires. In the INES model concepts are ordered within a grid. Four levels within the educational system are identified: the national level, the school level, the instructional setting (teacher and classroom) and the individual participants (students). At each of these levels three phases of the educational process are identified: antecedent conditions, malleable aspects and outcomes. Antecedent conditions include for example demographic characteristics of individual students and teachers, or the general conditions of schools. These antecedent conditions constrain to an extent the malleable aspects of the educational process, such as the implemented or intended school curriculum. These in turn relate to educational outcomes, such as levels of average achievement in foreign language skills. Table 6.2 illustrates this approach.

Having used the model to agree the major questions to be addressed, the next step was to make a selection of the most relevant variables and for each of these develop one or more questionnaire items to provide the data. The final set of language policies included in the questionnaire design is presented along with the questionnaire outcomes in Section 6.8.2 below.

The questionnaires required customisation for each country. This involved in the first place translation into the national language, or languages. Additionally, some terms and lists of response options needed to be rendered into terms appropriate to the given educational system, a step called localisation. Terms requiring localisation included the names of study

Table 6.2 Example of INES model: Concepts related to informal language learning opportunities

Level	Antecedents	Malleable aspects
Individual participant (Student Questionnaire)	Languages in the home environment	
Instructional setting (Teacher Questionnaire)	Target language exposure and use through home environment, visits abroad, traditional and new media	
Educational institutions (Principal Questionnaire)		
National educational system (National Questionnaire)	National and indigenous language Size and languages of immigrant population	Use of subtitles on television and film

programmes available at the tested levels; the most commonly taught foreign languages and indigenous languages, and the official correspondence of educational levels to the ISCED classification of educational levels.

Localisation also included giving participating countries the opportunity to add their own country-specific questions to the questionnaires. The response data for such questions was provided to the countries by SurveyLang, and used in the national reports which each country or region produced, but they were not referenced in the international report.

6.6 Sampling

Sampling is an interesting area of the survey to discuss here because it is relatively less familiar in the experience of Cambridge English Language Assessment, and it introduces a layer of complexity in how response data is elicited and interpretations developed. Exam boards typically focus on assessing individual students, and all considerations of reliability and validity relate to this focus. With a survey we are not interested in the individual except as an indirect indicator of the distribution of ability in the whole country (or rather, those individuals who make up the eligible group for the survey – in the case of the ESLC, students at the end of lower secondary education). The language ability of individuals does not even need to be precisely measured: as the presentation of the linked booklet design above shows, each student only sees a small fraction of the content of the tests. What is essential however is that the performance of each individual can be linked in a way which enables all performances to be brought into the same frame of reference, and that the selected group of students in the sample are perfectly representative of the population from which they are drawn. To take the range of research which Cambridge English Language Assessment engages with: studying the impact of a Cambridge English exam in a particular institutional setting can probably be done with a fairly simple approach

to sampling. However, a benchmarking study at national level would require the full sampling and analysis approach.

6.6.1 Grades and languages tested

Each participating educational system identified the first and second most taught languages on the basis of the latest available documented data from Eurostat (they could not elect to differ from this). The eligible grades for sampling students were defined in terms of ISCED levels. The international target population corresponded to the total number of students in eligible grades (ISCED2 or ISCED3) that were both attending educational institutions located within the educational system and also had studied the language to be tested for a minimum period of one academic year prior to testing. The survey was targeted primarily at the last grade in ISCED2, i.e. the end of lower secondary education, and participating educational systems were strongly encouraged to aim for this level. The survey standards allowed exceptions only in special situations where the use of ISCED3 level could be justified. In difficult cases SurveyLang worked with the national research co-ordinators to come up with appropriate rules to address issues specific to an educational system. Having two eligible levels (a preferred one and an acceptable one) reflected the situation in a country where a second foreign language might only be taught at the higher level. The practical requirement to define two eligible populations did, of course, complicate interpretation.

6.6.2 Sampling: Schools and students

A two-stage stratified sample design was used. That is, first schools were sampled, and then students within each sampled school. To test students in the first and second foreign language in each country two separate independent samples were chosen. The two samples could overlap, and so procedures had to be defined to ensure that no pupil was sampled (and therefore tested) in both foreign languages.

Stratification involves dividing schools up into homogenous groups according to relevant stratification variables. Use of stratification has several advantages:

- it maximises the efficiency of the sample design, thereby improving the reliability of survey estimates
- it enables using different sample designs, including disproportional sampling across different school strata
- it enables adequate (or minimum) representation of schools from different school groups and guarantees that all population segments are incorporated in the sample
- it thus enables reliable estimates for specific strata where necessary.

Examples of stratification variables used in the survey include:

- regions (for example, states/provinces)
- school size (large, medium or small)
- school types (for example, public/private)
- school programmes (for example, academic/vocational)
- urbanisation (rural areas, urban areas)
- socio-economic status (for example, low/medium/high income).

There are two types of stratification variable: explicit and implicit. The explicit variables illustrated above provide the flexibility to implement disproportional sampling across explicit strata when necessary to ensure adequate representation of certain types of schools (size, public/private etc.) or geographic regions. Implicit stratification involves sorting the schools within each explicit stratum by a set of implicit stratification variables before randomly sampling them with a specified sampling interval. Its advantage is that it ensures randomness in selection with respect to the implicit strata.

National research co-ordinators were requested to suggest stratification variables (explicit and implicit) appropriate to the specific features of their educational systems. SurveyLang then reviewed those suggestions and finalised the stratification variables. One variable always chosen for stratification, for example, was school size (the number of eligible students enrolled), generally using three levels (large, medium and small).

The sample allocation of schools across all explicit strata was done such that the proportion of students sampled in any explicit stratum was roughly the same as the population proportions of eligible students in the corresponding explicit stratum.

Within each stratum, schools were selected using probability proportional to size (PPS), where size refers to the number of eligible students enrolled for the language to be tested (for practical reasons, the figure from the previous academic year was used for this purpose). Using PPS ensures that larger schools are more likely to be sampled, so that each student in the country has a more equal chance of being selected.

The goal of sampling for any language was to select 25 students, representing students learning that language only but also students learning both languages. This involved a stratification of students in a school into three groups (first language, second language or both), with proportional selection from each stratum.

Student lists contained names, assigned IDs, and all relevant information captured in the student sampling form, including the results of the routing test, which enabled each student to be assigned a test at the appropriate level (this was not, of course, a variable used for sampling).

The selection of the students was by simple random sampling from the list of eligible students. Students learning both languages were then assigned

to one of the two languages, as no student was to be tested in more than one language.

Once the student sample was selected for any language, each student was randomly assigned for testing in two of the three skills: reading, listening and writing.

6.6.3 Ensuring quality of data

In order to ensure data quality, highest priority was given to the task of minimising the coverage error, i.e. for minimising the difference between the national desired target population and the international desired target population. SurveyLang made all possible efforts to limit exclusions from the national target population. Countries might wish to exclude schools from the sample in cases where the school is particularly small, or is a special need school, or where physical access to the school is difficult. In fact, according to the data submitted, no national research co-ordinators excluded specific regions on the basis of problematic access in any of the educational systems covered. Hence in terms of geographic coverage, the national target population matched the international target population in every entity surveyed.

Besides school-level exclusions, student-level exclusions constituted another quality indicator of the national survey samples. To avoid countries defining within-school exclusions differently, SurveyLang requested national research co-ordinators to follow specific rules. Within-school exclusion rules might be applied to three groups: functionally disabled students, intellectually disabled students, or students with insufficient command of the questionnaire language of the educational system. Any other reason for within-school exclusion was to be documented in detail on the sampling form. The objective was to limit the overall school-level and within-school exclusions to 5% of the national target population. Response rates also have an important effect on the quality of data collected. It was important to determine minimum participation rates for schools as well as for students. The survey set the bar at a minimum participation rate of 85% of originally sampled schools, although it was accepted that sampled schools choosing to opt out of the test might be substituted with 'replacement schools'. For each sampled school in the Main Study, up to two replacement schools were assigned from the sampling frame at the time of the selection of the main sample. For each sampled school, the schools immediately preceding and following it on the sorted list (also known as the sampling frame) in the same explicit stratum were designated as its replacement schools.

Along the same lines, the bar for students was set at a minimum participation rate of 80% within participating schools (sampled and replacement). It was acknowledged that follow-up sessions might be necessary in some schools where too few students took part in the tests originally conducted.

It was left to the School Co-ordinators and Test Administrators to decide together with the national research co-ordinators whether additional sessions were needed.

Data quality also depends on the accuracy and precision targeted by the sampling design. In the school sample a minimum of 71 schools for each of the two designated languages were selected in most of the participating educational systems.

Roughly 1,775 (71*25) students were sampled, in general, for each language. This was the standard sample size requirement at the national level for any educational system to participate in the survey. Based on an overall response rate of 85%, about 1,500 students per educational system per language were expected to be tested. Given that any student only responded to two of the three tests (Reading, Listening, Writing), an average sample (or cluster) size for any single skill of 14 (=25*(2/3)*0.85) per school was expected to be achieved.

Such careful specification is necessary in order to control the desired level of precision in the final estimates. Predicting the precision or accuracy of an estimate is not straightforward in complex surveys. It depends on the *effective sample size* which in turn depends on the underlying *design effect*. The design effect reflects the fact that the sample is not a simple random sample of every student in the country. Defining sampling in terms of a two-stage process (schools then students) is standard practice in educational surveys because it is far more practical and economical than taking a simple random sample. However, it introduces an effect due to the intra-class correlation of the groups thus created. The size of the effect is difficult to estimate in advance, but in practice values found for similar studies can be used to estimate the likely effect size, and thus the sample size needed to achieve some desired level of precision. As noted above, for any single language and skill, an average cluster size of 14 per school was expected to be achieved. Given this cluster size, and anticipating an intra-class correlation coefficient of 0.1, the design effect could be roughly estimated to be about (1 + 13*0.1) = 2.3. This was an approximation and was expected to vary depending on the exact value of intra-class correlation coefficient in specific educational systems and estimates. However, based on this simplifying assumption, the effective sample size corresponding to 1,000 completed cases was expected to be around 437 (1,000/2.3=437). This was expected to result, at the educational system level, in a minimum precision (or maximum sampling error) of ±4.7% for estimation of an unknown population proportion, which was considered adequate given comparable surveys. The precision associated with any estimator for any other subgroup (region, demographic groups etc.) was of course dependent on the corresponding sample size and also on the nature of the estimator. Stratification was employed in the sample design to further reduce the variance of the survey-based estimators.

6.6.4 Sampling: School personnel

The survey school personnel sample was self-selecting – each of the participating schools' principals and language teachers teaching the test language at the testing level were invited to complete the School or Teacher Questionnaire, respectively. Where a school was selected for both test languages, the school principal was randomly allocated to complete the School Questionnaire for one test language only rather than having to complete the two questionnaires, one for each test language. There was no official participation criterion for the teachers and principals. Educational system samples were eligible to be included in the international sample, even if the response rate for questionnaires among teachers remained low. In the event, there was one country where low participation rates proved somewhat problematic for interpretation of the questionnaire data.

An issue which was discussed in the survey Advisory Board was whether teachers should be linked in the data to particular students, there being some concerns that students' responses might somehow be linked back to particular individual teachers. With the sampling approach described above this was not possible. It is arguable that this design decision rendered the Teacher Questionnaire data somewhat less useful, as at best it could only be linked to student performance at the level of the school.

6.7 Standard setting

As argued in Section 2.4 above, standard setting in relation to a multilingual framework remains a difficult and under-theorised area, and orthodox standard-setting approaches are not wholly appropriate. Standard setting was a challenging aspect of the survey. It is interesting that much of Cambridge English Language Assessment's earlier work on scale construction (Chapter 3 above) actually had very little requirement to engage with standard-setting issues, for the reason that the standards embodied by the core Cambridge English exams were inherited, rather than set. Thus the first applications of the Rasch model could take standards directly from the traditionally applied grading judgements of examiners, applied to quite well-established exams and stable candidate cohorts. The Cambridge Common Scale project simply involved empirically linking these inherited standards into a single scale. The gradual transition from traditional to statistically driven grading obviated the need for standard setting. With an item-banking system in place, new exams could take their standards from statistical anchoring to existing exams. The ALTE Can Do project (Section 3.4 above), which linked Can Do descriptors to examination grades, enhanced the description of the grades, but only in a few cases challenged the standards. Thus unlike many other areas of the survey test

development, which could be based on well-rehearsed Cambridge English practice, standard setting was at a disadvantage.

It may seem curious that the language partners in the survey chose to make so little use of existing standards. Each of them have, of course, exams which are linked to the CEFR, and a reasonable amount of work could be cited (the ALTE Framework and Can Do project, several multilingual CBT projects) to base a claim of comparability on. However, no empirical work was undertaken to link survey material to CEFR levels via the partners' existing calibrated tests, although this was suggested as a possible validation study.

This reticence reflected the consortium's sensitivity to the potential accusation of imposing its own standards on a European benchmark. Perhaps too it was felt that despite the efforts made over the years to construct a common scale for the ALTE examinations (see Sections 3.2, 3.3 and 3.4) the empirical evidence for the comparability of standards across the language partners could not be sufficiently neatly presented. Furthermore, there was a view among some consortium members that standard setting should be made as visible and accountable a process as possible, inclusive of all the participating countries. Countries not included in the standard-setting process would not accept the resulting standards, or so it was argued.

6.7.1 Standard-setting proposals in the tender

SurveyLang's tender to the Commission (Cambridge ESOL 2007) proposed a complex approach to standard setting which in the final event was only partially implemented. The importance of verifying standards across languages was pointed out, as was the fact that applications of the pilot manual (Council of Europe 2003) had largely involved single languages. The survey would provide an important opportunity to make progress in establishing comparability of standards within Europe. Task-centred procedures would be used for the receptive skills (reading and listening), specifically, the Cito variation of the bookmark method (Council of Europe 2009:82); and learner-centred procedures for the productive skill of writing. All procedures were to follow the steps described in the *Manual for relating language examinations to the CEFR* (Council of Europe 2009).

It was explained that the psychometric analyses which would define the scales for the five languages could not be generalised in a simple way to a common scale for all languages together. Therefore comparability required the judgements of experts able to judge the relative difficulty of tasks or the relative merits of performances in two or more languages. A procedure was described for deriving a common scale for the performance skill of writing, whereby judges would rank-order a set of students' works, half in one language and half in another.

Considerable attention was paid to the validation of the outcomes. It was pointed out that the Manual's linear treatment of the different phases in linking examinations to the CEFR (specification – standardisation of judgements – standard setting – empirical validation) should be modified so as to run the processes as far as possible simultaneously, and to allow multiple independent sources of information to influence each other. Five sources of information would thus be exploited:

1. The test construction process would produce tests with highly comparable CEFR-linked content, such that a provisional standard setting could be derived from them.
2. The students' responses would provide the basis for calibrating the tasks, thus giving judges an explicit view of the progression in difficulty defined by the tasks.
3. The judgements of the panel members would be an independent source of information, reflecting the quality of the training and their familiarisation with the CEFR and with the test material. A high degree of inter-judge agreement would promote international acceptance of the standards, hence their validity.
4. An independent source of information as to the ability of students would be teacher judgements, for practical reasons probably based on a subsample. Judgements would be based on a number of Can Do statements taken from the CEFR or DIALANG, plus a number of exemplar tasks which the teacher would judge as within the students' ability to accomplish, or not.
5. Finally, students would be asked to self-evaluate their ability in relation to the same set of Can Do statements.

Validation would be built into the standard -setting processes, for example by having several rounds, where the nature of the judgement would be slightly different. The responses to the Can Do statements would be treated as items and calibrated together with the test responses, so that the difficulty of tasks could be related to the difficulty of Can Do statements. It was also foreseen that standard setting would be a two-stage process, with a first round implemented as soon as possible after the field trial, and a second one after the data collection for the main study.

The inception report explained that regardless of the method chosen, a degree of disagreement among panel members would require means of reconciling differences, in order to produce a standard enjoying the support of as broad a range of participants as possible. A footnote added, cautiously:

> It would be naïve to think that carefully carrying out a set of procedures must lead to a valid result by necessity. Linking test results to the CEFR is a complex endeavour, for which reasonable (looking) procedures are

> being currently proposed. The proof of their success can only be delivered by a series of well conducted studies, of which the ESLC is one, although prominent, example (SurveyLang 2008:105).

In the event some aspects of the above approach were not implemented.

The use of multilingual judges for all three skills, ensuring that cross-language comparability could be addressed at each stage of the process, was not implemented. As proposed, a cross-language alignment study for writing was undertaken (see Section 6.7.4), but specifically multilingual participants were not recruited for the standard-setting conference, and all judgements took place on a by-language basis. Thus ensuring that standard setting constituted an open and inclusive event with all interested parties welcome to participate finally took precedence over assigning the task to a smaller body of multilingual experts.

Of the above-listed approaches to standard setting and validation, the use of teacher ratings was not pursued, and although student self-ratings were collected in an appendix to the student questionnaire, these were not used in the standard-setting process, nor were they included as 'items' within the standard-setting process. Reasons for this are presented below.

Standard setting was done only after the Main Study: the proposed event after the field trial was not pursued.

Several additional studies were proposed in the tender but not followed up. One would replicate the study described in Section 3.2.4.2 for BULATS, where students with plurilingual competence took tests in two languages and provided Can Do self-ratings with respect to both languages. Also not followed up was the proposal for small-scale studies to be scheduled into the pilot and field trial stages of the development, which would involve selected students taking tests in two languages, and the conduct of plurilingual standard setting involving suitably qualified panels. In practice the aspiration to implement a series of such additional research-based studies in the context of delivering an international survey proved impractical.

6.7.2 The standard-setting conference

This event took place in Cambridge from 26 September to 30 September 2011. Standard setting was done by panels of judges in separate, monolingual contexts: there was one panel per language. Panels varied in size from 21 for English to eight for Italian. Standards were set for the three tested skills of listening, reading and writing. As stated in the inception report, a task-centred approach was used for the objectively marked skills of reading and listening. The model used was the Cito variation on the bookmark method described in the Manual (Council of Europe 2009), or rather, a variation on the variation, reflecting the approach taken to calibrating the tasks. The Survey tasks

were calibrated using a partial credit IRT model, so that students' performance was estimated from their score on the whole task rather than the separate items scored right or wrong. This analysis approach had been taken to deal with violations of the local independence assumption which could be predicted from features of some of the task types used (see Section 2.3.2). The standard-setting procedure is explained in more detail in the Technical Report (European Commission (2012b).

Each judge was provided with a sheet displaying each task as a horizontal line with points for each possible score, all shown in relation to an ability scale (Figure 6.3).

Figure 6.3 Two steps in the standard-setting procedure

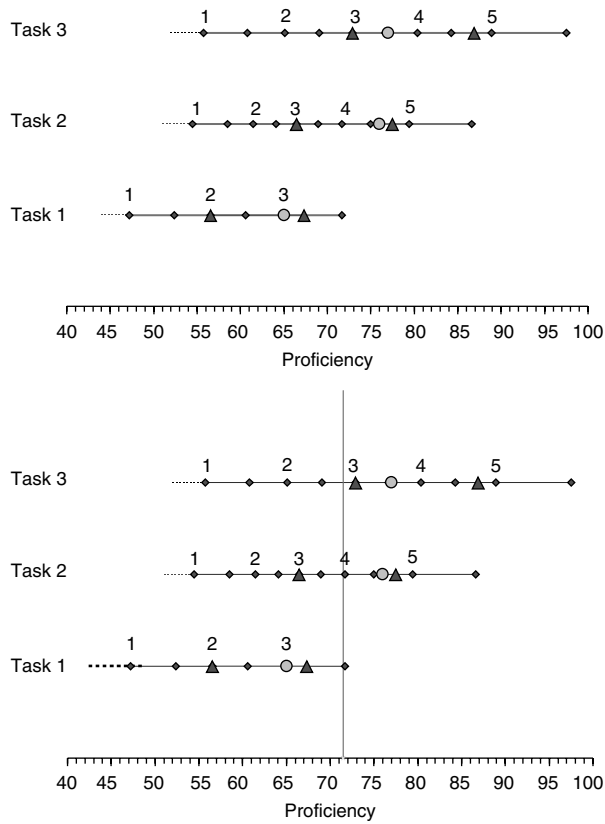


Figure 6.3 illustrates a sheet with three tasks. For each task possible scores are shown, from 1 up to the maximum score minus a half. Additionally, three levels of mastery of a task are defined for each task, separated by triangles:

- Full mastery: a score of 80% or higher
- Moderate mastery: a score between 50% and 80%
- No mastery: a score of less than 50%.

Tasks are ordered vertically on the sheet according to their 50% mastery point (harder tasks higher on the sheet). Central to this procedure is the concept of a borderline person. In the above example panel members are asked to imagine a person who is a borderline B1. The task for the panel members then consists of two steps:

- First they study each task in the task booklet, decide on the expected score for such a borderline person and indicate this on the graph for this task with a cross. This is illustrated at the top of Figure 6.3.
- Second, they draw a single vertical line which they believe best reconciles differences between the placement of the crosses. Thus the outcome for each judge is an estimate of a cut-off point between two levels, on the numerical scale along the bottom of the sheet.

In the first two rounds four standards were set, on three different answer sheets:

- for A1 and A2, the sheet displayed the tasks with intended Levels A1 and A2
- for B1 the sheet displayed all the A2 and B1 tasks
- for B2 the sheet displayed all the B1 and B2 tasks.

As a consequence the A2 and B1 tasks were displayed twice.

In round 3 (the validation round), the same kind of graphs were used, but with all tasks displayed on a single sheet. The task for the panel members was to set standards for all four levels on this sheet. For the third round a table was available for the panel members where the intended CEFR level of each task was indicated.

For standard setting of writing a student-centred method was used. The complete writing performances of a number of students per language (i.e. two or three tasks, depending on the assigned test level) were sampled and transcribed. Each of the eight tasks used in the Main Study were sampled (except for Italian, which was administered only in Malta to a candidature expected to be relatively advanced). In this case only three tasks were used, one at each of the Levels A2, B1 and B2, with no students expected to be assigned to the low level. Twelve performances were selected for each of the eight tasks (or three for Italian), selected from a pool of about 50 different students (20 for Italian).

For rounds 1 and 2, a variation of the Body of Work method was used. For each task the 12 performances were to be sorted into passes or fails, i.e. those which dealt adequately with the demands of the task and those which did not. The approach of judging each task separately was chosen for two reasons:

1. CEFR levels are usefully understood in relation to the tasks a student can perform at a level. The challenge presented by a task influences students' performance as well as raters' perceptions of it. Thus it was thought important for raters to be aware of, and judge in relation to, the demands of the task.
2. The by-task approach was coherent with that used in the Writing Alignment study (see Section 6.7.4) and was thus hoped to elicit comparable behaviour.

Round 1 responses were captured and presented in the form of a table with tasks in the columns and raters in the rows, with a zero or one in each cell (passed or failed) and marginal percentage totals. This was provided as normative information in round 2. It showed each rater which students had been generally rated higher or lower, and which raters had been generally more or less severe. Raters were asked to consider and discuss their ratings, and if they wished, change them.

In round 3, the same set of tasks were presented at student level: for each of 30 students the panel member had to assign a CEFR level on the basis of their complete set of performances (two or three tasks). Raters were offered eight CEFR categories (with higher and lower categories for each of A1, A2, B1), and these were collapsed to five (Pre-A1 to B2) for analysis.

6.7.3 The standard-setting conference – results

The standard setting produced mixed results in terms of coherence and interpretability. For all skills, significant differences were found for the method used in rounds 1 and 2 and that used in round 3. Also within methods, significant differences were found among panel members and/or tasks. This confirmed that standard-setting results are sensitive to the exact procedures used, and shows the value of having incorporated the third, validation round – even if it threw up differences which would need reconciling. Reconciling outcomes used both informed judgement and statistical evidence. The approach attempted to identify a standard reflecting the individual judgements of as many of the panel members as possible.

For reading and listening, as described in greater detail in the Technical Report (European Commission 2012b:283–285), it was found that the standards set on the four different sheets used in the three rounds were in the case of several raters seriously inconsistent, producing outcomes which the raters themselves certainly did not intend (for example, with the sequence of CEFR levels reversed). A close analysis revealed two patterns of rating, one of which evidenced confusion on the part of some raters arising from the form of presentation of the rating task. This motivated discarding some of the data relating to some of these raters.

Other effects could also be hypothesised: the fact that the pre-A1 and A1

cut-offs were set on the same set of tasks (the A1/A2 level test) was likely to lead judges to separate these cut-offs more clearly, perhaps placing the A1/A2 cut-off higher than they might otherwise have done. This would be consistent with round 3 outcomes, where all levels and tasks were shown on a single sheet of paper, and where in consequence a greater discrimination of levels was also observed.

For reading and listening a process of cross-language comparison and reconciliation had to be completed following the standard-setting event (European Commission 2012b:286–288). The problems arising during the event made this requirement more evident, but it was always predictable, given the absence of an explicit cross-language dimension to the standard setting, as had originally been proposed in the inception report.

In finalising standards significant weight was given to the test materials themselves, following the argument that the test construction process was such as to have produced tests in each language broadly comparable in terms of the construct tested and the absolute level of the tasks. Thus the very final step in reconciling standards across languages was, for each skill, to compare standards across languages on a whole-test score metric, that is, the modelled total score were a student to complete all tasks (in practice each student completed only a small fraction of the tasks, as shown in Section 6.4.5 above). This enabled a direct comparison across languages, on the assumption that scores should be broadly comparable. Differences between languages at each level threshold were then reconciled by taking an average value. There were two arguments for this approach:

- It should apportion uncertainty about the 'true' standard more equally across languages.
- It should ensure that the proportional width of each level is similar across languages. This satisfies an important requirement of a language indicator. It is highly desirable that the proportion of students achieving each CEFR level in any future round of the survey should depend solely on levels of achievement, and not on variations in the proportional placement of cut-offs. Imposing consistency in the first round should simplify interpretation in future.

This is in line with the argument put forward by Jones (2005b, 2009a), in the context of imposing a top-down approach to constructing the Asset Languages framework. Wherever in standard setting consistency across languages can be imposed by appeal to a common principle then this should be done, in preference to allowing every level in every language to be fixed according to the judgements of independent groups (see Section 2.4).

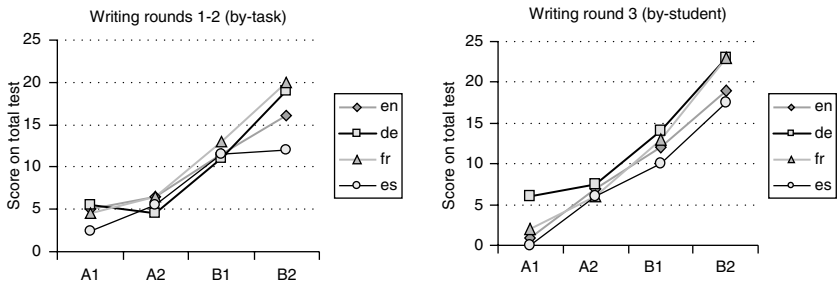
For writing the judgements were modelled using logistic regression with language proficiency as predictor. Just as observed in the case of reading and listening the different methods used in round 1–2 and in round 3 produced

different results: the by-student judgements stretched the distance between A1 and B2 more than did the by-task approach. It was not clear whether this reflected differences in the logistic models used, in the degree of agreement elicited by the two methods, or substantive effects in how raters view a student's complete body of work or a single task. Some approach to reconciling these two sources of data was needed.

As with reading and listening, standards were finalised by comparing across languages and using a whole-test score metric.

Figure 6.4 shows the cut-offs derived from logistic regression (by-task) or polytomous regression (by-student) for four languages. The A2 and B1 cut-offs are similar for both methods. The B2 cut-off varies substantially for the by-task method, whereas it is more consistent for the by-student method. The A1 cut-off is poorly discriminated from the A2 cut-off in the by-task method, and is consistently lower in the by-student method.

Figure 6.4 Writing cut-offs from round 1–2 (by-task) and round 3 (by-student)



The A2 and B1 cut-offs were set by averaging over the two methods and the four languages. The A1 and B2 cut-offs were set to reconcile differences between languages and weighting the outcomes of the by-student method. Italian, which as explained above used a reduced task set, was subsequently fitted to the other four languages.

6.7.4 The Writing Alignment study

For writing the cross-language comparability of the standards set could be validated against the outcomes of the Writing Alignment study, which had been completed prior to the standard-setting conference. For the performance skill of writing samples of performance in different languages can be directly compared by suitably multilingual judges, and in this way inform an alignment of standards. The eight writing tasks used in the Main Study are essentially cloned across the five languages, making comparison by task straightforward. The purpose of the study was to inform alignment of the

standards set. It was not intended to impact on the underlying scale construction for each language and skill from Main Study response data.

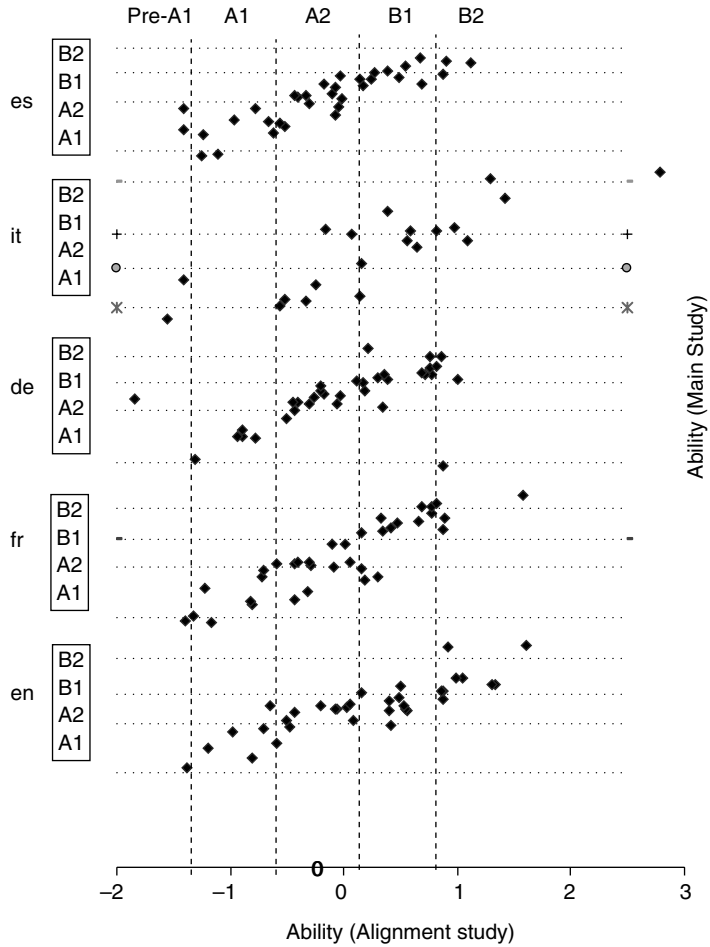
The approach was based on ranking: raters with a competence in two languages were asked to rank-order sets of samples, half in each language. (see Section 2.3.4.3 for more on ranking). Students were selected from the multiple- and centrally marked Main Study scripts, randomly from each level of a distribution stratified by booklet and average mark. This ensured linking across levels, and a gradation of levels of performance. Selection was done for each language by the appropriate language partner, who verified the suitability of the performances of each selected student and could select an alternative if necessary. All the performances of selected students (two or three tasks) were transcribed and entered into a database. The samples used for the standard-setting conference were selected from these. The alignment study samples are a subset of the standard-setting samples. For a more detailed account of the design and analysis see European Commission (2012b:289). For examples of writing performance at each of the tested levels see Appendix D.

Figure 6.5 provides a view of the data comparing student abilities as estimated from the Main Study and the alignment study. The x-axis shows the student abilities on the common metric provided by the alignment study. The y-axis shows, for each language, estimated abilities in relation to the cut-offs set for that language.

This view shows that the number of data points potentially informative about each cut-off and each language is quite small, and although the correlations are high, there is clear variation between the two sets of estimates. There was disagreement within the consortium as to whether the data was adequate to support a statistical approach to alignment. The vertical dotted lines represent a subjective placement of CEFR cut-offs on the alignment study scale. It is possible to place these lines so that the transitions between level indicated by the Main Study cut-offs on the vertical axis correspond quite well to the transitions indicated on the alignment scale. This is reasonable confirmatory evidence that the final standards set per language are comparable. At least there is no evidence of gross variation. Figure 6.5 shows the final standards. These reflect the one change made on the evidence of the alignment study: the English B2 cut-off was slightly lowered.

The alignment study is an innovative aspect of the survey which addressed the empirical validation of standard-setting procedures, as advocated by the Manual for relating examinations to the CEFR (Council of Europe 2009, Chapter 7) under the heading of external validation, and as proposed in the inception report. As a demonstration of the value of the ranking approach to alignment it is somewhat less compelling than the Sèvres study reported in Section 2.3.4.3 above, but still contributes to experience of using ranking as a technique for cross-language alignment.

Figure 6.5 Main Study and Alignment study abilities



6.7.5 Can Do self-evaluation

Can Do statements were included in the Student Questionnaire as a potential means of validating test outcomes. How they might be used for this purpose was explained in the Inception Report (SurveyLang 2008), but the approach proposed there was in the final event not adopted. Students responded to 16 Can Do statements, evaluating their own competence in the tested language. The statements were administered as part of the Student Questionnaire but were analysed separately from the questionnaire responses. The statements were taken directly from CEFR descriptor scales or adapted slightly to ensure their relevance to the target population.

Table 6.3 shows the Can Do statements. Statements for speaking were included, even though speaking was not a skill tested in the survey, because it was considered worthwhile to elicit students' own perceptions of their competence in speaking relative to the tested skills of reading, listening and writing. As shown in Figure 6.6, student perceptions of their relative competence in the different skills were quite similar across the tested languages.

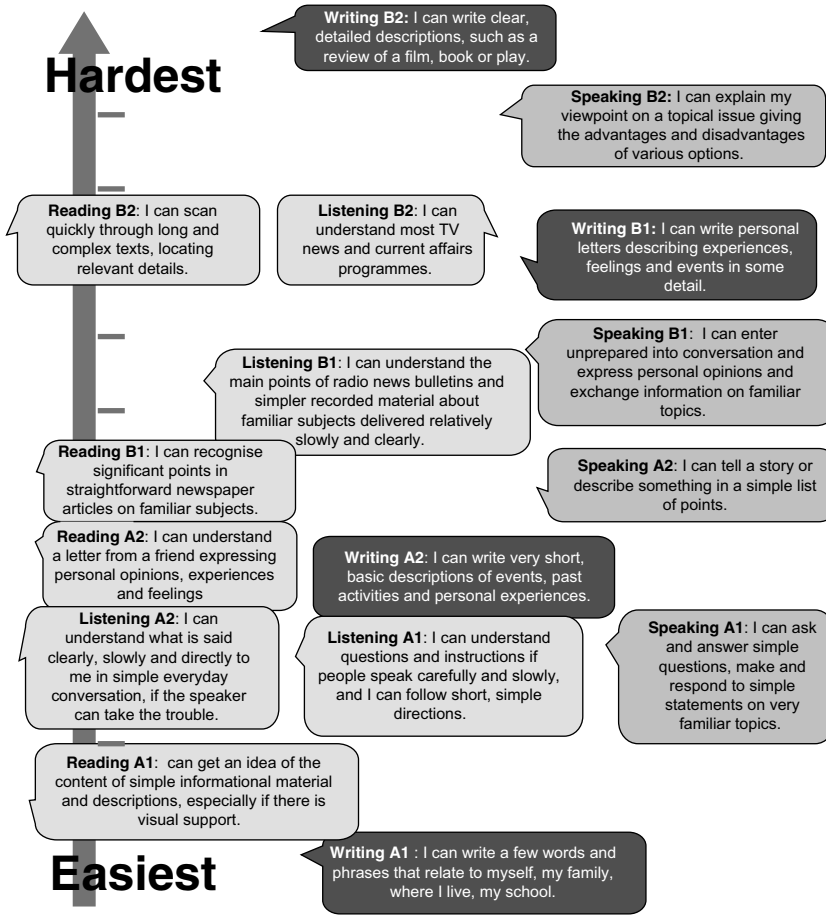
Table 6.3 CEFR Can Do statements included in Student Questionnaire

	Reading	Listening	Writing	Speaking
B2	I can scan quickly through long and complex texts, locating relevant details.	I can understand most TV news and current affairs programmes.	I can write clear, detailed descriptions, such as a review of a film, book or play.	I can explain my viewpoint on a topical issue giving the advantages and disadvantages of various options.
B1	I can recognise significant points in straightforward newspaper articles on familiar subjects.	I can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly.	I can write personal letters describing experiences, feelings and events in some detail.	I can enter unprepared into conversation and express personal opinions and exchange information on familiar topics.
A2	I can understand a letter from a friend expressing personal opinions, experiences and feelings.	I can understand what is said clearly, slowly and directly to me in simple everyday conversation, if the speaker can take the trouble.	I can write very short, basic descriptions of events, past activities and personal experiences.	I can tell a story or describe something in a simple list of points.
A1	I can get an idea of the content of simple informational material and descriptions, especially if there is visual support.	I can understand questions and instructions if people speak carefully and slowly, and I can follow short, simple directions.	I can write a few words and phrases that relate to myself, my family, where I live, my school.	I can ask and answer simple questions, make and respond to simple statements on very familiar topics.

6.7.5.1 Analysis of student responses to the Can Do statements

Analysis is described in more detail in the Final Report (European Commission 2012a:7). Figure 6.6, taken from the Final Report, summarises an analysis estimating the difficulty of each Can Do item, thus giving a simple picture of progression by skill, as self-assessed by students. The figure shows the calibrated statements arranged in descending difficulty.

Figure 6.6 Calibration of 16 Can Do statements

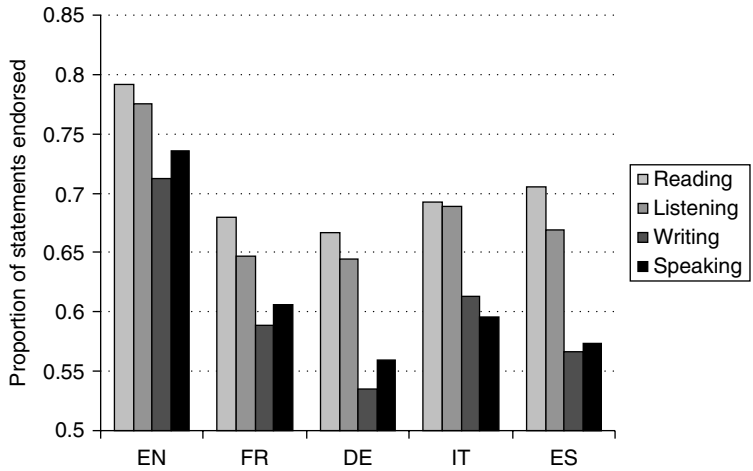


A second analysis allows a summary view of how the difficulty of the four skills is rated by students (Figure 6.7).

In terms of relative proficiency level, students of English rate themselves higher than other languages, which is not unexpected given that English is the first target language in most educational systems. However, the relative levels claimed for the other languages are not confirmed by the language test outcomes.

As Figure 6.6 shows, the perceived relative difficulty of the four skills is similar across all five tested languages: generally, reading is perceived as easiest, followed by listening, then speaking, then writing. Italian shows a different order, with reading and listening nearly equal in difficulty and writing

Figure 6.7 Can Do statements, all educational systems, by skill and language tested



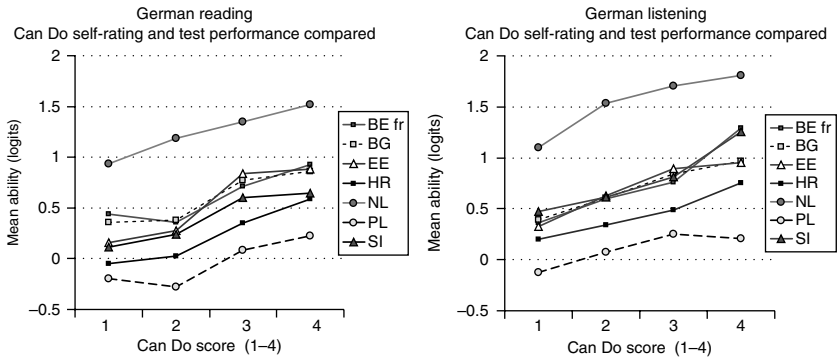
slightly easier than speaking. This may reflect characteristic features of the uniquely Maltese context in which Italian was tested.

As the Inception Report made clear, the Can Do statements were included as a potential way of validating or moderating the standard setting. The similarity of students' perceptions of their relative ability in the different skills might have motivated an adjustment to the standards for listening and reading, to make reading slightly easier. Within the constraints of the project, without the possibility of further validation, it was decided not to use the Can Do evidence in this way. However, further research might explore whether and how such evidence might be validly used in future iterations of the survey.

In fact, comparison of students' self-ratings with their actual level of performance in the language tests reveals an interesting phenomenon: their understanding of CEFR Can Do statements strongly reflects standards in their own educational system. Students rate their language ability relative to their peers, who are a familiar point of comparison, rather than relative to the fixed criterion intended by the Can Do statements. The self-ratings thus function normatively within each country.

Figure 6.8 below illustrates for German reading and listening (all five languages are shown in Appendix 8.1 of the Final Report (European Commission 2012a). The horizontal axis shows Can Do scores from 1 to 4, that is, the number of statements pertaining to each skill which students endorsed. A score of 4 indicates that all statements up to B2 were endorsed. Zero scores pose some problems of interpretation and are not shown in the figures.

Figure 6.8 Can Do scores and test performance by educational system: German Reading and Listening tests



The vertical axis shows the mean ability of the group endorsing a given number of statements, as estimated from the language test responses. The lines ranged on the vertical axis show the results by country (or region).

Within each educational system students are generally reasonably accurate in estimating their relative ability. However, the actual results of educational systems vary considerably. Students in the lowest performing educational system who rate themselves at B2 level are actually achieving lower levels than students in the highest performing educational system who rate themselves at A1. This general pattern is observed for all tested languages. What these graphs also demonstrate is that the Can Do statements discriminate far less than the language tests.

The above figures show for reading and listening that although individual students' self-ratings taken alone may not predict their absolute CEFR level, within one educational system they may predict their ranking quite well. However, it is clear that these self-ratings are context-dependent and relative. This indicates that they can contribute little evidence for where the criterion-referenced CEFR standards should lie, and for this reason it was concluded that they could not contribute to finalising the standard setting.

6.7.6 Evaluation of standard setting

As noted in previous sections, there are respects in which the approach to standard setting attempted less than originally proposed in the Inception Report. There were several reasons for this. The lack of an explicit cross-language dimension, for example, reflected the concern of some consortium members that the standard-setting event should be as inclusive and open as possible, which militated against requiring multilingual credentials of participants. The notion that standard setting should be a public event serving

the purpose of garnering support and acceptance for the standards seems to reflect conceptions of standard setting carried over from other contexts (considered in Section 2.4 above), and not necessarily applicable to the context of the survey. An alternative would have been to treat standard setting as the final stage in a single process of test construction and interpretation: as a professional task undertaken by the survey language partners, in an explicitly multilingual, comparative context.

As described above, the analysis of students' self-ratings on CEFR Can Do statements demonstrated clearly that, while they provide an interesting demonstration of the relative, context-dependent nature of standards, they could not contribute usefully to validating the absolute CEFR levels established for the survey.

The proposal to validate standards against teacher Can Do ratings of students was shelved for practical reasons, including the fact that in the questionnaire data teachers were not linked to students except at school level. It is in any case quite likely that they would have demonstrated the same norm-referenced effects as in the student self-ratings and for that reason their interpretation would have proved difficult.

The Technical Report concludes the chapter on standard setting by offering an evaluation, describing the final set of standards as defensible and practically useful, but acknowledging some uncertainty about their status (European Commission 2012b:294). That uncertainty relates in part to the current status of the CEFR itself, the nature of its levels and the different interpretation placed on them by different users, depending on local standards and custom. The aim of the survey standard setting is to set a common standard which may promote convergence of use in future, but the standard-setting event and the data collected inevitably reflect this uncertainty.

Uncertainty about the standards also relates to the very nature of standard setting: the impossibility of claiming absolute validity for the outcomes of research conducted in a particular context within particular practical constraints. The absence of speaking in the survey is recognised as an issue in using the outcomes to inform the setting of a European indicator or benchmark.

The standard-setting process itself, as presented in the Technical Report, illustrated the sensitivity of the process to apparently minor differences in the structuring of the standard-setting task. Standard setting inevitably involves reconciliation of opinions or of conflicting evidence. The reconciliation done after the standard-setting event to finalise the standards could be seen as just part of this process; but it is possible that a more integrated, in-house process with an explicitly cross-language dimension would have enabled greater confidence in the outcomes. This would be a useful area for experimental research following up the survey, to inform the design of a future round, and articulate methods for ensuring continuity of interpretation, linking back from a second round to the first one.

Finally we should consider whether the goal of a universal, common understanding of criterion-referenced standards such as intended by the CEFR levels is practically achievable. The ESLC, like other projects described in this volume, illustrates many of the problems: judgements of standards tend to be norm-referenced relative to general levels of achievement in the context familiar to the judge; what judges pay attention to, penalise or reward depends on particular conceptions of their role in language learning; attempts to formalise standard-setting judgements prove sensitive to features of the instrument used; and so on. Are these merely technical or communicational problems that can be addressed by better training and 'standardisation', or are there unresolvable ontological issues? Is reality simply not to be captured within a measurement framework of the kind proposed? If we recognise the heuristic, practical purpose of the endeavour we will probably agree that it is not about pursuing and sharing some absolute truth. To that extent we may expect that better training procedures, and the availability of more authoritative points of reference – such as the survey results – will in time lead to greater common understanding and convergence of practice. Certainly within a particular well-controlled domain, such as that of a single exam board, standardisation can be considered a practical matter of training.

The problems begin when different assessment cultures are expected to adopt a common standard. Simple inertia, and the practical difficulty of tinkering with accustomed standards and qualifications, are enough to derail the process. But it is also important not to think of 'standardisation' as a kind of reprogramming of the recipient to see the world the same way as the expert doing the standardisation. Standardisation is not something that should be done to people, but rather something that should grow out of what people do in their professional lives as educators, and how they can relate this to an external frame of reference. As argued earlier, we should see linking to the CEFR as an active process, working from the specific context of learning – and, of course, also paying attention to the major part of the CEFR text which is about the nature of language learning, not about the framework of levels. As more contexts actively participate in developing the use of the CEFR then the dimensions of the complex reality may take on more shape and finally impact on the descriptive framework of the CEFR itself.

The finding reported in Section 6.4.8, showing that markers of writing, even when the marking task was presented as a comparative exercise, still managed to demonstrate differences in severity systematically across countries demonstrates that one should never underestimate the human capacity to resist standardisation. This behaviour was presumably not conscious, but rather reflected habits of severity or generosity rooted in some notion of the markers' role in the educational process. Perhaps such systematicity can be harnessed; perhaps the multifaceted Rasch model presented in Section 2.2.3.1

would be a useful tool; at any rate, even if it is difficult to propose a specific methodology, I do believe that ideally standards, or common ground, should be discovered rather than imposed.

6.8 Outcomes of the European Survey on Language Competences

The following sections summarise the findings of the language tests and the questionnaires. Besides their inherent interest they also point up the relevance of the findings to Cambridge English Language Assessment, and the importance of the survey as an indicator of future directions in which their role in language education and language testing may develop.

6.8.1 The language tests

This section is based on the survey's Final Report (European Commission 2012a) which offers a fuller presentation of the results. Figure 6.9 and Figure 6.10 below show for first and second foreign language respectively, the percentage of students in each country achieving each CEFR level from Pre-A1 (i.e. failing to achieve A1) up to B2 (the highest level tested in the survey). In these figures the results are summarised across the three tested skills by taking an average of the percentage achieving each level in each skill.

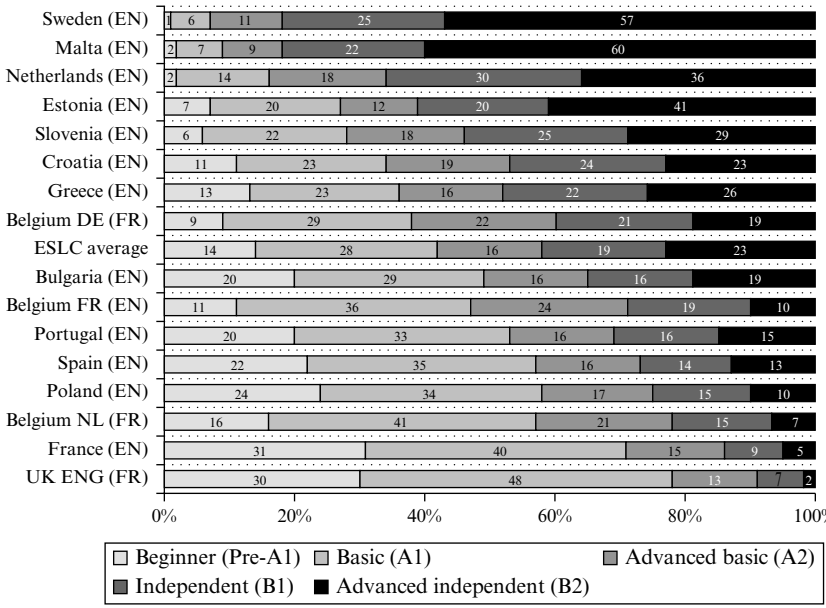
The countries are shown ordered from the highest performing (i.e. having more students at higher CEFR levels and fewer at low levels) to the lowest performing. This has the advantage of clear presentation, but the disadvantage that it suggests a simple 'league table' approach to evaluation. In fact, contexts of language learning differ so greatly across countries and languages that to understand the situation in a given country requires a much more qualitative and differentiated approach to evaluation.

None the less, the bare language test results tell a story: there is clearly a very wide range of achievement across countries and education systems. Figure 6.9 presents results in first foreign language, which is English for all countries except England itself, and the Flemish and German communities within Belgium, for whom it is French. Figure 6.9 shows that Sweden is the highest performing country, with 57% of students achieving CEFR Level B2. England is the lowest performing country, with 30% of students failing to achieve even CEFR A1, and only 2% achieving B2.

Figure 6.10 presents the picture for second foreign language. It can be seen that German (DE) is the second foreign language in eight education systems, French (FR) in three, Spanish (ES) in two and Italian (IT) in just one. Note that 'first' and 'second' here relate to the five tested languages.

The picture for second foreign language shows the same wide range of achievement as first foreign language. Achievements in CEFR terms are

Figure 6.9 First foreign language–Percentage of pupils at each level by educational system using global average of the three skills



somewhat lower, reflecting in part the generally much shorter duration of learning. For England, where 'first' and 'second' foreign language have more equal status, levels of performance are quite similar, with 30% of students failing to achieve A1 in either language, and 80 or 82% achieving no higher than A1 in French and German respectively.

For both first and second languages the number of students achieving no higher than A1, or not even achieving that, is high in many countries.

6.8.2 Questionnaire results

The questionnaires are organised around a number of language learning policy issues identified as being of interest to the European Commission (see Section 6.5). The Final Report presents the questionnaire findings in two ways: as simple tabulations of the data by country, for example, showing parents' mean knowledge of the target language, which was part of the concept 'language spoken in the home environment'. Then the relationship of that concept with results of the language tests was reported using regression analysis. To illustrate using this example:

Figure 6.11 shows the tabulated data on parents' language knowledge, reflecting the students' estimates on a scale from 0 (not at all) to 3 (very well).

Figure 6.10 Second foreign language – Percentage of pupils at each level by educational system using global average of the three skills

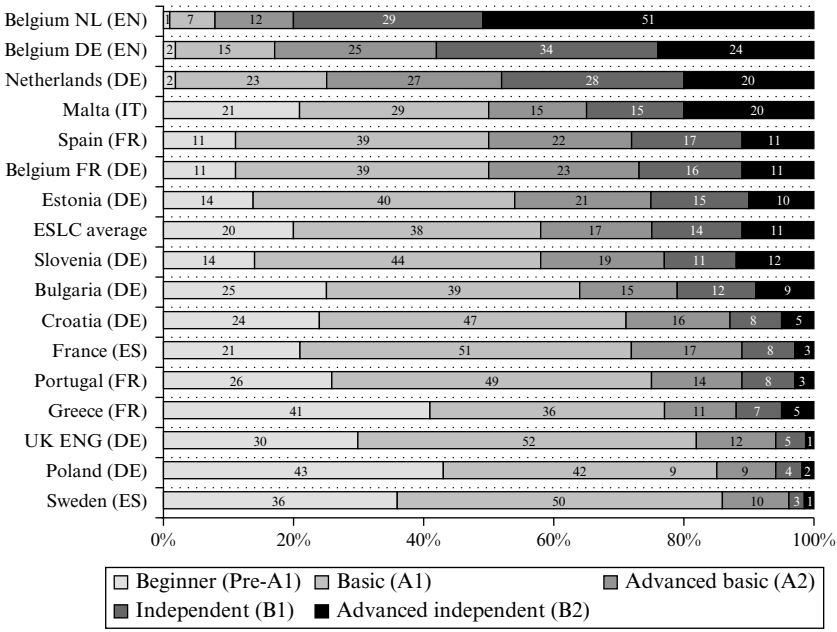


Figure 6.11 Parents' target language knowledge (mean)

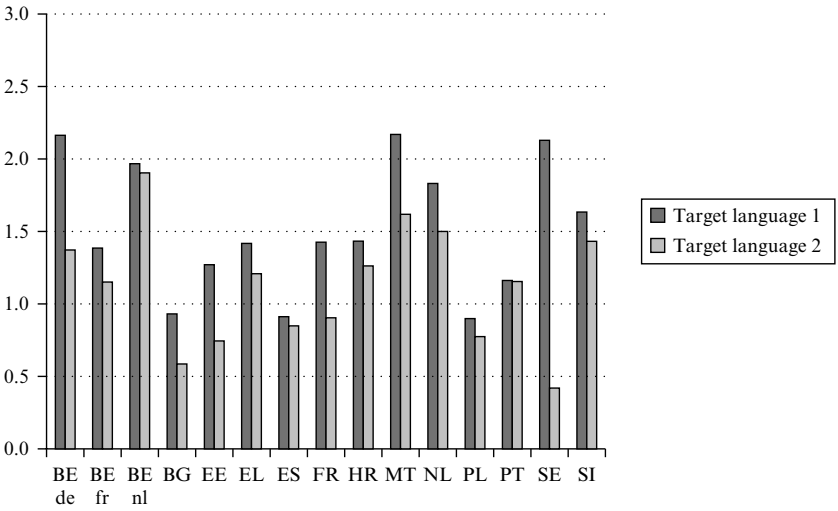


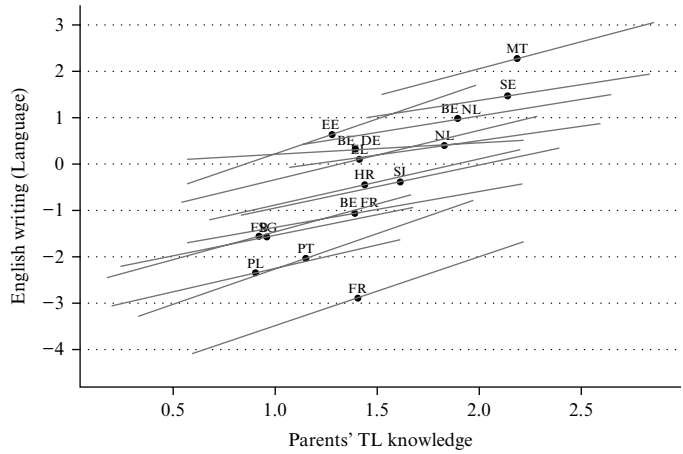
Figure 6.11 shows rather large differences between educational systems and within educational systems between target languages. Notably, in Sweden (SE) respondents' parents have very good knowledge of the first target language (English), and the least knowledge of the second target language (Spanish). This was closely reflected in the outcome of the language tests, where Sweden showed the highest performance in first English and the lowest in Spanish.

Then in the chapter reporting the results of the regression analyses the findings are described: 'In general, the effect of parental target languages knowledge is positive for all educational systems and languages, meaning that more parental target language knowledge goes with higher scores on the language tests. This effect is strongest for writing, followed by listening and to a lesser extent for reading. For writing, the effects are sometimes substantial' (European Commission 2012a:49).

The above description indicates that the nature of the relationship between a questionnaire index and test performance is a complex one. All the regressions were done separately for each educational system (country), target language and skill. How to identify what should count as a significant effect was thus an issue. As the Final Report explained: 'We used a rule-of-thumb for determining whether an overall effect is found or not. This rule-of-thumb is: if two thirds of the effects are in the same direction (either positive or negative) and one third of the effects are significant, we say that there is an overall effect' (European Commission 2012a:68).

Figure 6.12 illustrates the complexity by showing the relationship between parents' knowledge of target language and language scores for just one

Figure 6.12 Relation between 'Parents' knowledge of target language' and language scores for English writing



language and one skill: English writing. Fifteen graphs (five languages by three skills) were needed to illustrate the full picture for this single index. Thus for reasons of space the graphs were removed from the final version of the report.

For many of the policy issues addressed it was possible to show significant relationships with language test outcomes. For others it was not possible, simply because countries did not differ much regarding the specific issue at question. Where there is little difference, perhaps because all countries stand equally high or low on some index, then the regression will not detect an effect, even if the effect exists. Thus, for example, it was not possible to show a positive impact of using Content and Language Integrated Learning (CLIL), because it is generally still rather rarely adopted; but this does not suggest that CLIL is not an effective approach to language learning.

Another necessary limitation on how we may interpret regression findings is that we cannot treat a significant relationship as a proof of causation. For example, the survey found a significant relationship between the number of languages studied and ability in the tested language. But we cannot infer that studying more languages makes you a good language learner; it might be that people choose to study more languages *because* they are good language learners.

The contextual information collected through the questionnaires focuses on those factors which can be modified through targeted educational policies, such as the age at which foreign language education starts, or the training of teachers. The survey maps out differences within and between educational systems regarding three broad policy areas, and evaluates which of these relate to differences in language proficiency. Other factors which are largely beyond the control of policy such as general demographic, social, economic and linguistic contexts are not explicitly discussed in the Final Report, although data on socio-economic status is collected and is available for analysis by educational systems.

The following summary of significant questionnaire findings is condensed from the Final Report (European Commission 2012a:Chapter 6). Under each of the following headings a description of the raw data is followed by a summary *in italics* of the significant relationships found with language test outcomes.

6.8.2.1 An early start to language learning

Generally pupils report a rather early start to foreign language learning (before or during primary education) and most commonly they learn two foreign languages. However, considerable differences are still found across educational systems in the exact onset of foreign language learning, the current teaching time and the number of languages offered and learned.

The results of the survey show that an earlier onset is related to higher

proficiency in the foreign language tested, as is learning a larger number of foreign languages and of ancient languages.

6.8.2.2 A language-friendly living and learning environment

Policy also aspires to create a language-friendly living and learning environment, where different languages are heard and seen, where speakers of all languages feel welcome and language learning is encouraged. Clear differences between educational systems are seen in the informal language learning opportunities available to pupils (such as pupils' perception of their parents' knowledge of the foreign language tested, individual trips abroad, the use of dubbing or subtitles in the media, and the pupils' exposure to the language through traditional and new media).

A positive relation is observed between proficiency in the tested language and the pupils' perception of their parents' knowledge of that language, and their exposure to and use of the tested language through traditional and new media.

6.8.2.3 The language-friendly school environment

Differences are found in schools' degree of language specialisation, the availability of ICT facilities, the number of guest teachers from abroad and provisions for pupils with an immigrant background. However, exchange visits for pupils, and participation in school language projects display a relatively low take-up and most aspects concerning classroom practice display relatively less variation across educational systems (such as the use of ICT for foreign language learning and teaching, the relative emphasis teachers place on particular skills or competences, emphasis on similarities between languages, and pupils' attitudes to their foreign language study, its usefulness and difficulty). Only the amount of foreign language spoken in lessons shows clear differences across educational systems.

Pupils who find learning the language useful tend to achieve higher levels of foreign language proficiency and pupils who find learning the language difficult lower levels of foreign language proficiency. Also a greater use of the foreign language in lessons by both teachers and pupils shows a positive relation with language proficiency. Overall, differences in language specialisation, hosting staff from other language communities, and provisions for immigrant pupils show no clear relationship with foreign language proficiency.

6.8.2.4 Teacher qualifications and training

Improving the quality of initial teacher education and ensuring that all practising teachers take part in continuous professional development is identified in European language policy documents as a key factor in securing the quality of school education in general. Overall, most language teachers are well qualified, are educated to a high level, have full certification and are specialised in teaching languages. Also relatively little variation was found

between educational systems concerning in-school teaching placements and teaching experience even though differences exist in the number of different languages teachers have taught. Generally, across educational systems only a small proportion of teachers have participated in exchange visits, despite the availability of funding for such visits in a number of educational systems. We did find considerable differences between educational systems in teacher shortages and in the use of and received training in the CEFR, and, to a lesser extent, in a language portfolio; the actual use of a portfolio appears rather low. Concerning continuous professional development, despite clear differences found in the organisation of in-service training (such as financial incentives, when teachers can participate in training and the mode of training), reported participation in and focus of in-service training display less variation across educational systems.

The different indices related to initial and continued teacher education show little relation to language proficiency. For many indices this lack of a relation can be attributed to a lack of differences within educational systems. For others however, such as the use of and received training in the CEFR, considerable policy differences have been found, and yet these differences do not account for differences in language proficiency.

6.9 Conclusions: How to interpret the European Survey on Language Competences

In interpreting international survey findings caution is always recommended. The Organisation for Economic Cooperation and Development (OECD), who sponsors the PISA survey, reminds us that: 'on their own, cross-sectional international comparisons such as PISA cannot identify cause-and-effect relationships between certain factors and educational outcomes, especially in relation to the classroom and the processes of teaching and learning that take place there' (Organisation for Economic Cooperation and Development 2012:22). Sturman (2012:16) states: 'probably the greatest risk in the use of large-scale international datasets is the ease with which it is possible to draw overly simplistic – or erroneous – conclusions'. The International Association for the Evaluation of Educational Achievement (IEA), who runs Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS), is more positivistic, claiming that: 'the diversity of educational philosophies, models, and approaches that characterise the world's education systems constitute a natural laboratory in which each country can learn from the experiences of others' (International Association for the Evaluation of Educational Achievement 2012). However, this statement serves rather to underline the major problem, that educational systems operate in complex contexts whose parameters are difficult to identify and impossible to control – unlike laboratories, where the controlled

experiment is a fundamental concept. Schleicher's defence of surveys such as PISA is better: 'by showing what is possible in education, [international surveys] can help policy makers, researchers and practitioners to look beyond the experiences evident in their own systems and thus to reflect on some of the paradigms and beliefs underlying these' (Schleicher 2012). The idea that surveys prompt critical reflection on one's own practice, rather than simple models to be adopted wholesale, seems nearer the truth.

International surveys provide at best a partial picture. Critics claim that the evidence they provide plays too dominant a part in determining how countries compare and evaluate themselves: they get too much attention, relative to other high-quality educational research (Alexander 2012, National Research Council 2003). While the notion of 'evidence-based policy' is widely accepted, such criticism questions how evidence is actually constructed and used (Hammersley 2005). Policy makers pay lip service to the idea of evidence-based policy, but will too often ignore the larger part of the evidence that is available to them, or indeed, persist in dogmatically motivated policies which are directly contradicted by available evidence (Sahlberg 2011).

The survey reports submitted to the European Commission constitute a rich and complex set of findings. How much interpretation and evaluation of these findings was it appropriate or necessary for SurveyLang to provide? This issue was discussed at some length with the European Commission, and between SurveyLang partners. On the one hand, the Commission wished the survey to demonstrate its usefulness by producing clear recommendations – a view to which Cambridge English Language Assessment was sympathetic. Against this was the more psychometrically oriented opinion that the survey should report matters of strictly empirical fact and leave interpretation to others. This divergence of views also extended to the Commission's expert advisors. In the end it was the latter view which generally prevailed in the editing of the Final Report. However, the executive summary, a document published in several languages under the Commission's imprint to coincide with the release of the survey results, included evaluative section headings inserted by the Commission, stating for example that:

- language competences provided by educational systems still need to be significantly improved
- there is a wide range of ability across countries in Europe
- English is the language pupils are most likely to master.

A final section was also inserted entitled *Challenges for language learning in Europe*, identifying five specific areas requiring attention. In shortened form:

- language competences still need to be significantly improved, and educational systems must step up their efforts to prepare all pupils for further education and the labour market

- language policies should address the creation of language-friendly living and learning environments inside as well as outside schools and other educational institutions
- the wide range of ability among Member States in language competences indicates the rich potential for peer learning in language policy and learning, e.g. concerning early onset of foreign language learning, and promoting the teaching of foreign languages for meaningful communication
- the importance of the English language as a basic skill and as a tool for employability and professional development requires concrete actions to further improve competences in this language
- the need to improve language skills for employability in a globalised world must be combined with the promotion of linguistic diversity and intercultural dialogue.

A summary of key findings on the Commission's multilingualism website, updated in February 2013, also conveys the negative accent in its overall evaluation of the situation of languages in Europe:

- the survey reveals that Europeans still need to improve their knowledge of foreign languages and there is a wide range of ability across the participating countries
- only 42% of tested pupils were found to be competent (i.e. achieve B1) in the first foreign language tested and just 25% in the second
- a large number of pupils did not achieve the level of a basic user: 14% for the first foreign language and 20% for the second foreign language
- the level of independent user is reached by only 4% in Sweden (Spanish) and 6% in Poland (German) (European Commission 2013).

The publication of the survey results received quite wide coverage in the press (though not in England, predictably enough). The reaction of countries suggested that by and large their performance in the survey tended to confirm a pre-existing conviction that they were either good or bad at languages. If this were true, it rather undermined any expectation that the survey might have a positive impact on language learning in Europe. It would be a shame if good or bad results in the survey simply confirmed countries in the view that there was nothing that needed to be done, or alternatively, nothing that could be done.

In my view the survey findings could readily be interpreted to provide a clear model for language learning, of relevance to all countries and all languages. Thus at the Commission-sponsored conference 'Multilingualism in Europe', held in Cyprus in November 2012, I presented an account of the survey which provided a simple interpretation. As extended for later publication in *Research Notes* (Jones 2013b:5) it goes like this:

The above findings present quite a complex picture. However, a very brief summary of some of the significant findings does provide a compelling portrait of the successful language learner: a language is learned better where motivation is high, where learners perceive it to be useful, and where it is indeed used outside school, for example in communicating over the internet, for watching TV, or travelling on holiday. Also, the more teachers and students use the language in class, the better it is learned.

These conclusions are not surprising: they probably confirm what we already believed. However, it is an important achievement that the survey has provided empirical evidence in support of them. What the paragraph above describes is language being used for motivated, purposeful communication. It is this which favours learning: we learn in order to communicate, and we learn by communicating. Moreover, the Survey shows that this ideal learning situation is approximated only in some countries, and effectively, only for English.

Both the language test and questionnaire results confirm that English appears as a special case. It is learned to the highest level (note in [Figure 6.10] that it is the Flemish and German communities of Belgium which come at the top. English is their second foreign language, but they still perform more highly in it than in their official first foreign language, French). The questionnaires indicate that English stands distinct from other languages in terms of student perceptions of its usefulness, its visibility in life outside school, and its use as a medium for communication – a *lingua franca*. The successful learner of English appears to perceive and experience it in ways which are characteristically different to less successful learners of any language – including English.

Does English provide a model which other languages can follow? Clearly, it has advantages that other languages do not: above all, its higher visibility in many kinds of media. In Sweden, which as shown above performs very strongly in English and very poorly in Spanish, English is the language of a significant proportion of television programming. This is not the case for France, which performs poorly in English. In England there is an evident credibility problem: motivation to learn a foreign language is low because it is widely perceived as unnecessary, in a world where everyone else is believed to speak English.

However, the fundamental importance of placing communication at the heart of successful language learning applies to any language. The European Commission is inclined to acknowledge the special status of English as the language for business and for mobility (and to justify an emphasis on English in terms of pressing economic need), while stressing the cultural importance of other languages. However, we could argue that this is not a necessary either/or choice. Effective intercultural communication goes beyond the merely utilitarian or transactional, and being able to talk to an interlocutor in their own language, even to a modest level, is an asset, in business as in social life. The survey provides evidence that using language in purposeful communication favours

> learning, and that is perhaps the most important message to take from it: languages *will* only be successfully learned as communication tools. And in the age of social networking the current generation of learners have no shortage of things to communicate about. We may assert that an appropriate language policy for Europe in the 21st century will place communication and intercultural competence at the centre. For most learners this may well start with English, but it need not and should not finish there (Jones 2013b).

Is this just another tendentious, selective and dogmatic interpretation of an international survey? It can be defended on several grounds. Firstly, it is supported by the survey findings. Secondly, it confirms what many qualified experts would endorse on the basis of their experience. Experience is just the name we give to evidence collected over a lifetime. It is what makes experts expert; conversely, a definition of dogma might be that it is a strong belief not supported by the experience of experts in the field.

Further, one could argue that it is already understood in the premises of the survey itself, which set out to measure communicative competence in relation to the CEFR. The near ubiquitous presence of the CEFR in European language education reflects a general acceptance in theory of its communicative principles. What is striking in the survey outcomes is that in many countries what is accepted in theory is not evidenced in practice. Thus it is a particularly important interpretation to emphasise.

Finally, the only valid reason for conducting international educational surveys is in some way to impact positively on educational outcomes. The Commission has confirmed that a second round of the ESLC will be held, although at the time of writing we do not know whether Cambridge Assessment will have the opportunity to contribute to it. If that is the case, one would hope to be able to present and shape the next round as a positive intervention in European language learning, underpinned by a clearly articulated set of educational values. More will be said on this in the final chapter.

Research Notes 52 offers several further commentaries on the ESLC: Szpotowicz (2013) and Culej (2013) discuss the impact of the survey in Poland and Croatia respectively; Ashton (2013) reflects on some of the challenges, limitations and lessons learned over the 4-year period of intensive work; McKenna (2013) discusses the treatment of the survey in the European media.

7 Frameworks for the future

7.1 Frameworks so far

Over the period covered by this text the focus of work on scaling and framework construction has continually shifted, as the account given in this volume shows. This development reflects the evolving priorities of the Research and Validation Group, which now, as then, has the dual function of doing research around the exams as well as providing fundamental operational support for their administration. At the cost of some simplification it is possible to trace a progression through the different research themes that have been focused on over the years.

The initial focus was on the analysis, the administrative systems and the skills needed to develop measurement scales. Ensuring good quality of measurement and a grasp on reliability was an urgent issue, as the Cambridge-TOEFL Comparability Study had shown (Bachman et al 1995). The Rasch model was introduced and in-house software developed to build it into the operational test cycle. Classical item analysis also became routinely available, providing feedback that helped to improve the quality of item writing. Understanding the candidature better was also a priority. Early on a Candidate Information Sheet was developed to be completed by all candidates, providing for the first time some basic demographic information, vital for understanding the test populations and for monitoring trends. All this activity reflected a test development model (Saville 2003) which focused predominantly on the quality of data – on scoring validity in Weir's model (see Section 2.1).

Next there came a greater focus on interpretation – on criterion validity in Weir's model. The ALTE Can Do project and work to link this to the CEFR aimed to provide accessible characterisations of what it means to achieve a particular Cambridge English exam level. This then shifted to a more explicit consideration of construct validity, culminating in four volumes in the SiLT series co-authored with Weir (Shaw and Weir 2007, Khalifa and Weir 2009, Taylor (Ed) 2011, Geranpayeh and Taylor (Eds) 2013).

Progress having been made on these major aspects of validity and reliability, attention has shifted more recently to aspects of test use, and the development of impact studies as a research priority. Learning Oriented Assessment (LOA), introduced below, is a new focus which develops logically out of the study of test impact.

This progression can also be detected in the chronology of topics addressed by volumes in the SiLT series. Table 7.1 shows an illustrative selection of titles (leaving out general collections such as conference proceedings) arranged by topic and by time period.

The account of multilingual frameworks in this volume thus describes one thematic aspect (though a central one) of a broader range of developments. The changing focus of research interest is reflected in the history of scale construction and interpretation, and the priorities which determined the direction of developments. But history does not stand still, and in concluding the account it is necessary to look at current areas of interest and possible future frameworks.

7.2 Beyond the CEFR: A framework for language education

This volume has dealt largely with frameworks in relation to summative assessment. Asset Languages offers one case of a framework which came closer to the classroom, although the implementation of its formative strand remained relatively basic (see Section 5.5). Today however Cambridge English examinations are being widely adopted into the state educational sector, and projects are undertaken which encompass curriculum design, teacher training and system evaluation. The role of English as a medium of instruction in many contexts brings into focus the role of languages across the curriculum. Increasingly, Cambridge English assessments impact on important issues of language policy and education in general.

The CEFR explicitly limits its scope to foreign language learning, so it cannot be criticised for excluding other aspects of language education. However, as argued in Jones (2013a), the contexts in which language assessment expertise is called into play are becoming increasingly complex, and the framing of language policy is something that *should*, we might agree, be carried on within an inclusive, coherent framework. This could be seen as a higher-level, more heterogeneous and inclusive system, of which the CEFR with its focus on foreign language learning would form one part.

Even as a framework for foreign language learning, the CEFR is somewhat narrowly defined. This is evident not so much in the conceptual framework presented in the chapters of the CEFR book, but rather in the descriptive apparatus which underpins the framework of *levels* through which the CEFR has proved so influential. For most users it is the descriptor scales which are most salient in their understanding of the CEFR, and they have acquired more importance than was the original intention.

Two foreign language learning contexts in particular are not best treated by the CEFR:

Table 7.1 Selected volumes in the SiLT series showing changing focus over the period 1995–2013 (titles abbreviated)

	1995–1999	2000–2004	2005–2009	2010–2013
Test methods, validity	Cambridge-TOEFL Comparability Study (Vol 1) Issues in Computer-Adaptive Testing of Reading Proficiency (Vol 10)	The Equivalence of Direct and Semi-Direct Speaking Tests (Vol 13)		
Learners and test performance	Test Taker Characteristics and Test Performance (Vol 2) Learner Strategy Use and Performance on Language Tests (Vol 8)			
Research methods	Verbal Protocol Analysis in Language Testing Research (Vol 5)	A Qualitative Approach to the Validation of Oral Language Tests (Vol 14)		
Subjectively assessed skills	Performance Testing, Cognition and Assessment (Vol 3)	Testing the Spoken English of Young Norwegians (Vol 20)	IELTS speaking and writing (Vol 19)	
Washback and impact			Changing Language Teaching through Language Testing (Vol 21)	

Table 7.1 (continued)

1995–1999	2000–2004	2005–2009	2010–2013
		Impact Theory and Practice: IELTS and Progetto Lingue 2000 (Vol 24) IELTS Washback in Context (Vol 25) The Impact of High-stakes Examinations on Classroom Teaching (Vol 22)	
Construct and levels, the CEFR	The Componentiality of L2 Reading in English for Academic Purposes (Vol 12) CPE (Vol 15)	Examining Writing (Vol 26) Examining Reading (Vol 29) Business English (Vol 17) Academic English (Vol 23)	Examining Speaking (Vol 30) Examining Listening (Vol 35) Aligning Tests with the CEFR (Vol 33) Components of L2 Reading: Linguistic and Processing (Vol 32) Measured Constructs (Vol 37) Exploring Language Frameworks (Vol 36)

- young learners, because there is no explicit treatment of cognitive stage
- CLIL because language for learning is not clearly distinguished from language for social use.

These two contexts are related: CLIL may involve young learners learning school subjects through the medium of a foreign language. This is in fact the case for a wide variety of L2 learning contexts.

Taking these two factors into account requires us to supplement the proficiency dimension by two additional dimensions – age and academic content area, enabling us to describe a learner at a specific proficiency level, at a specific age, studying a specific subject. The WIDA consortium's English Language Proficiency Standards for English Language Learners in Kindergarten through Grade 12 (WIDA 2012) illustrates just such a matrix.

In understanding the role of languages in education an important distinction can be made between language used for social interaction and its use as the medium of learning. Cummins (1979, 1984) identified Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP): respectively, the 'surface' skills of listening and speaking, which are typically acquired quickly by many students, and the basis for coping with the academic demands placed upon a learner in school.

In a CLIL or L2 setting CALP requires specific attention from an early age and possibly from a low proficiency level. This is not reflected in the descriptor scales of the CEFR. Rather, these reflect the customary progression in a language school, where it is only at the C levels that 'academic' use of language is envisaged. Clear distinction of these two aspects of proficiency would favour clearer formulation of language policy and more effective practical action.

An extended framework would also require clearer treatment of the differences between native and non-native speakers. The CEFR's treatment of the native speaker is confusing. It asserts that native speakers have a higher level than C2: 'level C2 . . . is not intended to imply native-speaker or near native-speaker competence. What is intended is to characterise the degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners' (Council of Europe 2001:36). And yet there are C2 descriptors which clearly identify competences *beyond* the level of many native speakers, for example, in Table 2 of the CEFR, entitled Common Reference Levels: self-assessment grid (Council of Europe 2001:26–29):

- **Reading**: I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.
- **Spoken production**: I can present a clear, smoothly flowing description or argument in a style appropriate to the context and with an effective logical structure which helps the recipient to notice and remember significant points.

- **Writing**: I can write clear, smoothly flowing text in an appropriate style. I can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

Very successful foreign language learners may acquire such educated competences, while many native speakers never do. At the same time, there is much that native speakers naturally acquire which remains beyond the reach of all but a few foreign language learners. Naturally, these competences are not described in the CEFR because they are not relevant to the foreign language context. Mother tongue language is characterised by the linguistic reflexes of a developed socio-cultural competence (culture in the 'broad' sense): a shared grasp of idiom, cultural allusion, folk wisdoms, and so on. The native speaker can understand and move freely between linguistic codes appropriate to insider groups or more formal communication (Bernstein 1971).

Shifting the focus to the individual learner we should recognise the fundamental role which language plays in the individual's development: everyone is a native speaker; everyone learns through language; everyone can benefit from learning new languages and be enriched by the experiences and opportunities which languages afford. A general framework could connect all these aspects, enabling a coherent approach to language education, where proficiency in foreign languages is an integrated element. Such an approach would align more closely with the Council of Europe's concept of plurilingualism, which sees an individual's linguistic repertoire as a complex whole, reflecting the entirety of experience within society, rather than as a set of discrete competences with respect to separate languages.

Identifying the limits of the CEFR also identifies certain limitations of the multilingual frameworks which are the theme of this volume. All the projects and developments undertaken so far and described herein are concerned squarely with measuring foreign language proficiency, rather than with providing a comprehensive system within which language policies can be developed and implemented. This final chapter looks forward to a reasonably proximal future in which such a bold re-orientation might be accomplished.

7.3 Engagement in education: Impact studies and the ESLC

Delivering the ESLC has given us a close-up view of language learning in Europe. The wide range of achievement across countries demonstrated by the survey has clear implications for educational policy makers tasked with carrying forward a policy for languages. There is no European norm – every country is different, not only in terms of the parameters studied in the survey,

but in other important respects: its educational traditions, the structure of its industry and business, and above all perhaps the cultural, historical and linguistic factors which contribute to the image a country entertains of itself as being 'good' or 'bad' at languages – possibly powerful stereotypes with self-fulfilling positive or negative impact. Such attitudes are clearly visible in the national press commentaries provided on the European Survey outcomes in different countries.

Given such heterogeneity, successful educational initiatives will need to be designed and carried through on a by-country basis. Assessment regimes are doubtless an important factor in success, although as the Asset Languages story illustrates, assessment must be integrated within a coherent overall programme if it is to stand a chance of success (see Section 5.6). The adoption of Cambridge English exams at national or institutional level in countries in and beyond Europe is an important development for Cambridge English, changing the composition of the traditional candidature and locating the exams in new and more complex contexts. Accordingly, studying the impact of these exams has become a research priority for Cambridge English Language Assessment. *Research Notes* issue 50 (Cambridge ESOL 2012) reports on a number of such case studies. Many of these contexts involve collaboration with education systems at national, regional or institutional level. The work already done in the area of impact shows Cambridge English Language Assessment engaged in working with countries in Europe and elsewhere on the formulation of language education policy, and assisting in its implementation.

The ESLC has given Cambridge English Language Assessment a further opportunity to engage with important issues in European language education, and to contribute information for making language policy. The experience complements and extends work on impact, which continues through bilateral collaborations in a number of countries. Such experience helps us to conceptualise better the nature of the relationship between assessment and language education, and, possibly, a vision for languages in Europe which we can contribute to realising.

As concluded above in the context of the Asset Languages project, assessment systems cannot impact on language learning outcomes on their own, but only as part of wider programmes. Nonetheless, in my interpretation (see Section 6.9) the ESLC confirms the value in foreign language education of setting concrete, criterion-related goals, and using assessments which measure them. If the central importance of teaching language for purposeful communication is accepted, then the Cambridge English suite of exams provides a good model, and underlines the value of the CEFR as a point of reference, not only of relevance for assessment, but as a resource for shaping approaches to learning and teaching.

As the survey outcomes show, there is a great deal of work to be done

to make language learning more effective, and assessment of communicative language proficiency has an important part to play. But we should also agree with the *Guide for the development and implementation of curricula for plurilingual and intercultural education* (Council of Europe 2010:29) when it asserts: 'language teaching in schools must go beyond the communication competences specified on the various levels of the CEFR'.

The title of the CEFR presents it explicitly as a framework for *learning, teaching and assessment*, and it remains an obvious point of reference for linking these three concepts; but there are compelling reasons for extending the CEFR in at least two ways:

- Firstly, it defines itself narrowly as a framework for foreign language learning, and as explained above, a coherent policy for languages should reach across the curriculum and include mother tongue.
- Secondly, we need theories of action to help articulate how learning, teaching and assessment should come together. Politicians (and assessment professionals) may be tempted to see testing as the driver of learning, but we need a more ecological concept, defining complementary roles for teaching and assessment.

Language education implies more than achieving some level of proficiency in a language. It comprises a range of learning skills and learning objectives that are critical to becoming competent learners not just in one language, but more importantly, given that the languages we need in later life are probably not those we learned at school, of languages generally. Though the focus of impact studies may be on English, going forward we may increasingly be treating English as just one element in a coherent comprehensive policy for language education in a given context.

Moreover, language education impacts crucially on educational outcomes generally. Hawkins wrote of an 'apprenticeship in languages':

> We will no longer measure effectiveness of the apprenticeship in languages by mere ability to 'survive' in a series of situations, but by how the foreign language experience contributes to learning how to learn through language, and to confidence as a (mathetic) language user (Hawkins 1999:138).

Mathetic means: serving discovery, understanding and learning. Hawkins emphasises the importance of mother tongue competence to success in school, and of foreign languages in developing awareness of how language works.

Embracing this all-encompassing view of language education does not require us to abandon the constructs and procedures described in this volume, that have served the measurement of language proficiency well; but it will be necessary to extend them. The following section outlines one model for how this may be done.

7.4 Learning Oriented Assessment

The term LOA has been adopted recently by several writers (Boud 2006, Carless 2007, Jones, Saville and Hamilton 2013, Purpura 2004, 2009).

The model of LOA presented here is one which Cambridge English Language Assessment has been working on for some years, and is currently developing as a theory of action for achieving positive impact, particularly where Cambridge English exams are adopted in institutional educational settings, that is, as a significant intervention in language learning. As described above, Cambridge English is devoting considerable research effort to studying the impact of its exams in such settings, where it is possible to work with local partners to create beneficial links between classroom practice and the examination which is a final target, with the aim of achieving 'positive impact by design' (Saville 2009, 2012).

Every context of learning must be treated on its own terms, as the great differences between countries identified by the ESLC clearly confirm. Policy that will impact positively on language learning must be made country by country, on the basis of case studies.

Cambridge Assessment, and Cambridge English Language Assessment as a part of that organisation, have grown up in a British (or European) tradition where educational assessment is linked into educational processes (rather than stands outside them, as in the US psychometric tradition). Most of our candidates are still found within educational settings. The range of 'wraparound' services and products offered as support to teachers and learners has expanded steadily over the years. As recounted in Chapter 5 on Asset Languages, Cambridge English Language Assessment has even grappled with an implementation of formative assessment.

However, LOA is a bigger idea, offering a systemic model of assessment operating on multiple levels in an educational context and taking on many different forms (Jones et al 2013). It encompasses both the *macro level* of framing educational goals and evaluating outcomes, and the *micro level* of individual learning interactions which take place in the classroom or outside it. It defines complementary roles for teaching and assessment expertise.

LOA sets out to define a new, beneficial and principled relationship between assessment and learning at a time when the traditional relationship is coming under pressure to change from powerful external forces. These partly reflect technological developments which subject teaching and assessment to similar pressures, above all to treat learning and testing as commodities to be efficiently 'delivered' through the use of information technology. This in turn creates commercial pressures, as both learning and assessment become a target for media companies.

There are also ideological forces at work, as some governments turn to assessment as a lever for raising educational standards. Sahlberg (2011)

describes an ideology, dominant particularly in the Anglophone world, which he christens the Global Education Reform Movement, or GERM.

Sahlberg's concept of GERM has five features which are worth listing in detail:

1. **Standardisation** of teaching and learning through performance targets and standards for teachers and students, centrally prescribed curricula, external testing, inspection and evaluation.
2. **Focus on 'basic' skills** of literacy, numeracy and to a lesser extent science, as the sole indices of student achievement and national educational and economic success.
3. Use of **low-risk ways of maximising achievement** in this narrow area: standardised textbooks, prescribed pedagogy, conformist professional culture discouraging teacher choice.
4. **Corporate management models** from the business world, applied both systemically – businesses fund and control state-maintained schools – and institutionally, through business-derived management approaches: targets, measurement, conformity and control.
5. **Use of high-stakes testing** for teacher, school and system accountability, published league tables, rewards and sanctions (Alexander 2012:8).

An example of GERM in action is the *No Child Left Behind* programme launched in the United States in 2001, which imposed rigid accountability through extended testing of students – 'No child left untested', as it is described by many. British readers will also doubtless find examples closer to home. As a polemical device Sahlberg's invocation of GERM certainly serves to remind examination providers engaging more closely with education that there are pitfalls to be avoided. Assessment and education are values-laden concepts, and we may need to be explicit about the values we accept or reject.

Given these pressures to change the nature of the relationship between educational assessment and the process of learning and teaching, what options are there for shaping how this new relationship should look? Historically, assessment impacts on learning in a simple way: exams provide curricula which define learning objectives, and final accreditation for successful learners. Although much has happened in the last 20 years, so far the traditional relationship between assessment and learning has not fundamentally changed: exams remain external, summative commentaries on learning, which may at best have washback effects that impact positively on outcomes.

To support learning in radical new ways requires us to question our current trait-based conceptions of assessment – that is, the conception of measurement at the centre of the projects described in this volume. In his introduction to *Test Theory for a New Generation of Tests* (Frederiksen et al 1993), Mislevy states: 'Educational measurement faces today a crisis that would appear to threaten its very foundations. The essential problem is that

the view of human abilities implicit in standard test theory – IRT as well as classical true-score theory – is incompatible with the view rapidly emerging from cognitive and educational psychology.' Trait-based measures, it is explained, fail to capture the complexity of abilities in the way necessary to understand and impact on the process of learning. For this purpose detailed cognitive models are needed, and approaches to measurement that can deal with them (Frederiksen et al 1993, Pellegrino, Chudowsky and Glaser 2001).

The conception of LOA which Cambridge English Language Assessment is developing shares Frederiksen et al's recognition that the trait-based approach on its own is insufficient as a model for supporting learning. However, it questions the cognitive modelling approach, for a number of reasons. Firstly, it is not an appropriate metaphor for language learning, which is a unique subject in the range of learner attributes – cognitive, psycho-motor and affective – which it engages. The cognitive modelling approach seems to treat learning tasks in terms of *content* (that is, a close analysis of the thing to be learned). LOA's socio-cognitive approach (see Section 2.1) understands learning tasks more in terms of the *interactions* which they generate.

Secondly, the cognitive modelling approach appears reductionist: learning may be better seen as an emergent property (Larsen-Freeman and Cameron 2008) which cannot be reduced to its atomic parts as simply as the cognitive modelling approach attempts to do. Thirdly, the model cannot on its own explain classroom learning: diagnosing what a student knows or does not know is only a starting point for formative activity, which entails interaction, and, at least in the LOA view, a pivotal role for teachers in co-ordinating that interaction.

Therefore the Cambridge English model of LOA defines a complementary relationship with teaching. We acknowledge the limitations of the traditional trait-based concept of assessment, but also recognise its value: locating learners on a measurement scale offers them an orientation, perhaps motivation, and enables a degree of profiling or diagnosis. Linking the scale to criterion levels of performance makes clear the value of the learner's achievement. Trait-based assessment can provide the vertical, quantitative dimension of learning. A second – horizontal, qualitative – dimension is needed, to capture how learners at the same global level differ in terms of their cognition, experience, and learning needs. The vertical dimension is the primary domain of assessment experts; the horizontal dimension is the primary domain of the teacher. The LOA model thus foresees a central role for the teacher in creating an environment productive of learning, complementary to the role of assessment.

In the LOA model the fundamental assessment or learning event centres on a task, which engages the learner's cognition, eliciting language activity in

suitable conditions (such as the provision of comprehensible input, or scaffolding). These conditions enable learning mechanisms, related in particular to communicating personally significant meanings. Feedback is generated, performance is judged, through self-evaluation or evaluation by others. Interactional authenticity is the notion that tasks, though not in themselves authentic, can engage the learner's cognition in authentic ways: 'a function of the extent and type of involvement of task takers' language ability in accomplishing a test task' (Bachman 1991:691, citing Widdowson 1978). It applies equally well to test tasks as to activities in the classroom. Thus LOA links both learning tasks and test tasks to real-world tasks in the same way.

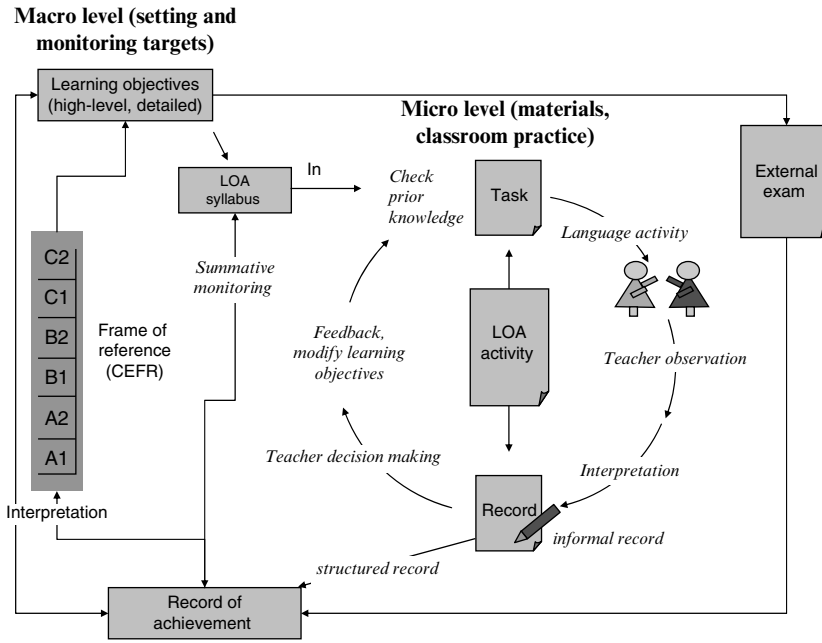
Interaction is thus central to learning. Assessment helps *adapt* teaching to the global level of a learner, but it is *interaction* at that level where learning happens. This view is consistent with that of several other writers. The 'interaction hypothesis' (Gass, Doughty and Long 2007, Long 1996), sees the negotiation of meaning as the means by which learning takes place. The 'output hypothesis' (Swain 1985) argues that production and practice is necessary for the self-monitoring which enables the learner to test and modify hypotheses about the language. These positions are all consistent with the socio-cognitive model presented above, and stress the centrality for learning of purposeful language activity prompted by engagement with a task.

The validity of LOA will be demonstrated by studying the synergies between the qualitative and quantitative (assessment and teaching) dimensions which LOA enables. An implementation of the LOA model is sketched in Figure 7.1 and briefly described below.

At a macro level, learning objectives are defined and progress monitored. The CEFR provides the interpretative frame of reference. The micro level concerns the classroom. A classroom LOA activity engages learners in purposeful language activity, has an explicit learning objective or objectives, and produces a record, that is, evidence. Evidence is the basis of evaluation, feedback and self-monitoring, and is thus a central concept in LOA.

Within a learning-oriented view of assessment trait-based scaling and measurement technologies still have an important part to play, and as noted above, more complex psychometric models are being developed, and new purposes will be found for the item-banking, IRT techniques for scale construction which have been presented in this volume, helping to create learning environments that adapt to the individual learner. Following on from the EPP, Cambridge English Language Assessment has funded a new institute within the University to undertake research on Automated Language Teaching and Assessment (ALTA). Launched in October 2013, it brings together the expertise of several departments of the University – the Computer Laboratory, the Department of Engineering and the Department of Theoretical and Applied Linguistics – and will further the development of assessment applications for automated speech recognition and marking of

Figure 7.1 Elements of LOA



writing, as well as explore the emerging science of computer-driven adaptive language learning.

These applications of artificial intelligence will doubtless impact on the Cambridge approach to language assessment, as well as certain scenarios within LOA. But they will be only one aspect of development, complemented by other kinds of more qualitative information, and based on data from observations elicited in far less structured forms of interaction than the traditional testing formats.

7.5 In conclusion

This book has tried not to claim too much for proficiency frameworks, presenting them as heuristic devices justified by their practical utility, rather than as windows on aspects of an absolute truth. When I first discovered IRT, accessibly presented in *A Guide to Language Testing* by Henning (1987), what caught my attention was the notion of making latent traits visible. I glimpsed it as a metaphor for providing motivating feedback into learning, which was an issue much on my mind as I attempted to introduce concepts of communicative language learning in the unpromising context of a Japanese junior college. I was sufficiently excited about the idea to step away from language

teaching and undertake a PhD in the topic. Having completed the PhD the only place to move on to – within the UK at least – was Cambridge English, where the commitment to do language assessment better was clear, as were the opportunities to do innovative work with like-minded colleagues. Now many years later the language learning question that first interested me in assessment seems to be moving centre stage in the form of LOA.

Of course, assessment must continue to fulfil its traditionally important roles in society, which depend on such traditional values as objectivity, reliability, validity, security, probity and quality. However, the potential value of assessment as a positive force within language education is far too rarely realised, and it is this area which is in greatest need of reconceptualisation and reforming action. Assessment expertise cannot bring about reform or progress on its own, as the Asset Languages experience confirms; it must be enabled to work alongside other forms of expertise in larger, integrated interventions. Nonetheless, assessment remains a powerful force within education, in many contexts significantly determining, for good or ill, how languages are conceived and taught, and how useful and beneficial the outcomes of learning are.

The findings of the ESLC can be readily interpreted to show that successful language learning requires a focus on language for purposeful communication. Language exams which test purposeful communication may move teaching in the right direction, and the complex, coherent and inclusive conception of assessment provided by LOA represents a more explicit model of intervention within which to pursue closer alignment of assessment and teaching.

LOA is a conception which encompasses assessment at all levels, from defining national objectives down to individual teaching moves in the classroom. Such a complex system will generate complex data. That data must be processed to provide information, which becomes evidence, which informs decisions and actions. A greater focus on learning means a greater focus on the individual learner. The quantitative, vertical dimension represented by proficiency scales allows learners to be ranked and located relative to each other, and to criterion levels of achievement. That is where summative assessment stops, but it is the point where the orientation towards learning begins, using available evidence to arrive at optimal decisions and actions to move each individual forward.

In other words, the new area of learning-oriented work outlined in this concluding chapter will continue to build on the measurement frameworks described in the preceding ones. That area of work is far from completed, as the problems and issues raised in the conclusions to several of the chapters make clear. In fact, it follows from the pragmatic, heuristic status of such frameworks that the work never will be finished, because there are no right answers, only ones which are better or more useful than those we currently have.

Appendix A

CEFR Common Reference Levels: Qualitative aspects of spoken language use

Table A1 Common Reference Levels: Qualitative aspects of spoken language use

	Range	Accuracy	Fluency	Interaction	Coherence
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turn-taking, referencing, allusion making, etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his remarks in order to get or to keep the floor and to relate his/her own contributions skilfully to those of other speakers.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.

Table A1 (continued)

	Range	Accuracy	Fluency	Interaction	Coherence
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he/she searches for patterns and expressions. There are few noticeably long pauses.	Can initiate discourse, take his/her turn when appropriate and end conversation when he/she needs to, though he/she may not always do this elegantly. Can help the discussion along on familiar ground confirming comprehension, inviting others in, etc.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some 'jumpiness' in a long contribution.
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.

A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can answer questions and respond to simple statements. Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord.	Can link groups of words with simple connectors like 'and', 'but' and 'because'.
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like 'and' or 'then'.

Appendix B

Sample illustrative descriptors

ALTE Can Do project: Example statements

Listening/speaking

A1	CAN understand simple directions, e.g. 'turn left at the end of the road'.
	CAN ask very simple questions for information, such as 'What is this?'. CAN understand 1 or 2 word answers.
	CAN understand simple replies, for example 'Yes. We will deliver on Friday.'
A2	CAN give simple explanations about familiar places.
	CAN understand a simple phone message and confirm details of the message.
	CAN understand and answer simple predictable questions.
	CAN express simple opinions using expressions such as 'I don't agree'.
B1	CAN take more complex messages, provided that the caller dictates these clearly and sympathetically.
B2	CAN keep up a conversation on a fairly wide range of topics, e.g. personal and professional experiences, events currently in the news.
	CAN present her/his own opinion, and justify opinions.
	CAN give a clear presentation on a familiar topic, and CAN answer predictable or factual questions.
C1	CAN deal with unpredictable questions.
	CAN show visitors round and give a detailed description of a place.
	CAN follow up questions by probing for more detail. CAN reformulate questions if misunderstood.
C2	CAN use the telephone confidently, even if the line is bad or the caller has a non-standard accent.

	CAN argue effectively for or against a case, and has sufficient language to be able to talk about/discuss most aspects of her/his work.
	CAN take an active part in most kinds of seminars or tutorials. IS LIKELY to understand cultural references.

Reading

A1	CAN understand basic hotel rules and signs, for example 'Dining-room'. CAN understand basic hotel information, for example, times when meals are served.
	CAN understand store guides (information on which floors departments are on) and directions (e.g. to where to find lifts).
A2	CAN understand a letter which describes people or events.
	CAN understand price labels and a range of advertisements such as 'Special Offer' in a department store or counter service shop.
	CAN understand the main points of information given on posters.
B1	CAN understand most articles and reports of a 'general' nature.
	CAN understand a letter expressing personal opinions.
	CAN understand most tourist brochures, guidebooks etc.
B2	CAN understand the general meaning of a report even if the topic is not entirely predictable.
	CAN understand most correspondence likely to be received.
	CAN understand most factual product literature within own work area.
C1	CAN understand complex opinions/arguments as expressed in serious newspapers.
	CAN scan texts for relevant information, and grasp main topic of text.
C2	CAN use appropriate strategies for efficient reading (skimming, scanning, etc.)
	CAN understand abstract concepts and argumentation.
	CAN read quickly enough to cope with the demands of an academic course.

Writing

A1	CAN write a simple routine request to a colleague, of the 'Can I have 20 X, please?' type.
	CAN leave a simple message giving information on e.g. where he/she has gone, what time he/she will be back.
A2	CAN convey personal information of a routine nature to, for example, a pen friend, and CAN express opinions of the 'I don't like ...' type.
	CAN complete most forms related to personal information.
	CAN write a short, simple letter introducing her/himself to a host/exchange family containing basic, factual information such as name, age etc.
B1	CAN write a simple narrative or description, for example, 'My last holiday', with some inaccuracies in vocabulary and grammar.
	CAN write letters on a limited range of predictable topics related to personal experience.
	CAN write to a hotel in order to confirm accommodation, etc.
B2	CAN draft a set of straightforward instructions, regulations etc.
	CAN present arguments, using a limited range of expression (vocabulary, grammatical structures).
	CAN write letters of thanks, sympathy and congratulations.
C1	CAN write a report that communicates the desired message. WILL need more time to write the report than a native speaker would.
	CAN write an essay with only occasional difficulties for the reader, whose message can be followed throughout.
C2	CAN write an essay that shows an ability to communicate with few difficulties for the reader.
	The essay shows a good organisational structure, which enables the message to be followed without much effort.
	CAN write with an understanding of the style and content appropriate to the task.
	CAN write any type of letter necessary in the course of his/her work.

Cambridge ESOL Common Scale for writing

LEVEL C2

MASTERY
CERTIFICATE OF PROFICIENCY IN ENGLISH
Fully operational command of the written language

- Can write on a very wide range of topics.
- Is able to engage the reader by effectively exploiting stylistic devices such as sentence length, variety and appropriacy of vocabulary, word order, idiom and humour.
- Can write with only very rare inaccuracies of grammar or vocabulary.
- Is able to write at length organising ideas effectively.

LEVEL C1

EFFECTIVE OPERATIONAL PROFICIENCY
CERTIFICATE IN ADVANCED ENGLISH
Good operational command of the written language

- Can write on most topics.
- Is able to engage the reader by using stylistic devices such as sentence length, variety and appropriacy of vocabulary, word order, idiom and humour though not always appropriately.
- Can communicate effectively rare inaccuracies of grammar and vocabulary.
- Is able to construct extended stretches of discourse using accurate and mainly appropriate complex language which is organisationally sound.

LEVEL B2

VANTAGE
FIRST CERTIFICATE IN ENGLISH
Generally effective command of the written language

- Can write on familiar topics.
- Shows some ability to use stylistic devices such as variety and appropriacy of vocabulary and idiom though not always appropriately.
- Can communicate clearly using extended stretches of discourse and some complex language despite some inaccuracies of grammar and vocabulary.
- Can organise extended writing which is generally coherent.

LEVEL B1

THRESHOLD
PRELIMINARY ENGLISH TEST
Limited but effective command of the written language

- Can write on most familiar and predictable topics.
- Can communicate clearly using extended stretches of discourse and simple language despite relatively frequent inaccuracies of grammar and vocabulary.

LEVEL	WAYSTAGE
A2	**KEY ENGLISH TEST** **Basic command of the written language** • Can write short basic messages on very familiar or highly predictable topics, possibly using rehearsed or fixed expressions. • May find it difficult to communicate the message because of frequent inaccuracies of grammar or vocabulary.

Cambridge ESOL Common Scale for speaking

LEVEL	MASTERY
C2	**CERTIFICATE OF PROFICIENCY IN ENGLISH** **Fully operational command of the spoken language** • Able to handle communication in most situations, including unfamiliar or unexpected ones. • Able to use accurate and appropriate linguistic resources to express complex ideas and concepts and produce extended discourse that is coherent and always easy to follow. • Rarely produces inaccuracies and inappropriacies. • Pronunciation is easily understood and prosodic features are used effectively; many features, including pausing and hesitation, are 'native-like'.

LEVEL	EFFECTIVE OPERATIONAL PROFICIENCY
C1	**CERTIFICATE IN ADVANCED ENGLISH** **Good operational command of the spoken language** • Able to handle communication in most situations. • Able to use accurate and appropriate linguistic resources to express ideas and produce discourse that is generally coherent. • Occasionally produces inaccuracies and inappropriacies. • Maintains a flow of language with only natural hesitation resulting from considerations of appropriacy or expression. • L1 accent may be evident but does not affect the clarity of the message.

LEVEL	VANTAGE
B2	**FIRST CERTIFICATE IN ENGLISH** **Generally effective command of the spoken language** • Able to handle communication in familiar situations. • Able to organise extended discourse but occasionally produces utterances that lack coherence, and some inaccuracies and inappropriate usage occur. • Maintains a flow of language although hesitation may occur whilst searching for language.

	• Although pronunciation is easily understood, L1 features may be intrusive. • Does not require major assistance or prompting by an interlocutor.
LEVEL B1	**THRESHOLD** **PRELIMINARY ENGLISH TEST** **Limited but effective command of the spoken language** • Able to handle communication in most familiar situations. • Able to construct longer utterances but is not able to use complex language except in well-rehearsed utterances. • Has problems searching for language resources to express ideas and concepts resulting in pauses and hesitation. • Pronunciation is generally intelligible, but L1 features my put a strain on the listener. • Has some ability to compensate for communication difficulties using repair strategies but may require prompting and assistance by an interlocutor.
LEVEL A2	**WAYSTAGE** **KEY ENGLISH TEST** **Basic command of the spoken language** • Able to convey basic meaning in very familiar or highly predictable situations. • Produces utterances which tend to be very short – words or phrases – with frequent hesitations and pauses. • Dependent on rehearsed or formulaic phrases with limited generative capacity. • Only able to produce limited extended discourse. • Pronunciation heavily influenced by L1 features and may at times be difficult to understand. • Requires prompting and assistance by an interlocutor to prevent communication from breaking down.

Appendix C

Asset Languages

Example Languages Ladder statements (Listening)

LISTENING

Breakthrough	Grade 1	I can understand a few familiar spoken words and phrases.
	Grade 2	I can understand a range of familiar spoken phrases.
	Grade 3	I can understand the main point(s) from a short spoken passage.
Preliminary	Grade 4	I can understand the main points and some of the detail from a short spoken passage.
	Grade 5	I can understand the main points and simple opinions (e.g. likes and dislikes) of a longer spoken passage.
	Grade 6	I can understand spoken passages referring to present and past or future events.
Intermediate	Grade 7	I can understand longer passages and recognise people's points of view.
	Grade 8	I can understand passages including some unfamiliar material from which I can recognise attitudes and emotions.
	Grade 9	I can understand the gist of a range of authentic passages in familiar contexts.
Advanced	Grade 10	I can understand the main points of an authentic spoken passage/conversation involving one or more speakers.
	Grade 11	I can understand the main points of authentic spoken passages and conversations in a range of different contexts.
	Grade 12	I can identify the majority of points and am able to infer the meaning of a range of authentic passages/conversations spoken at near native speed.

Breakthrough: Grades 1–3

On *completing* this stage, you should be able to understand a basic range of everyday expressions relating to personal details and needs. You may need to listen several times to get the information you need, depending on how fast and clearly the speaker talks. You should have some understanding of a few simple grammatical structures and sentence patterns. You should be familiar with the sound system of the language. You should be aware of how to address people both formally and informally as appropriate.

Preliminary: Grades 4–6

On *completing* this stage, you should be able to understand standard speech relating to a range of predictable everyday matters, providing that it is spoken clearly and directly. You should be able to recognise the difference between past, present and future events and be familiar with simple forms of the verb tenses.

Intermediate: Grades 7–9

You should now be comfortable with a range of tenses, and should be able to understand authentic passages on familiar matters. On *completing* this stage, you should be able to follow much of what is said at near normal speed on familiar matters or in predictable situations. You should be able to give an oral or written summary of what you have heard.

Advanced: Grades 10–12

You should now be comfortable understanding a range of tenses and a variety of registers. On *completing* this stage, you should be able to understand the majority of what you hear in the target language, including references to the culture and society of countries/communities where the language is spoken.

Sample generic specifications for Breakthrough, Preliminary and Intermediate stages

Generic Breakthrough specification content

1.1 Language purpose and functions

At Breakthrough stage there are only two broad categories of language functions:

- imparting and seeking factual information
- expressing and finding out opinions.

At Breakthrough stage, the realisation of these functions will be in a basic way. The following list is not exhaustive but gives some examples of this.

Greeting and responding to greetings
Giving personal details – name, age
Counting and using numbers
Talking about the weather
Talking about food etc.
Following and giving simple instructions
Expressing thanks
Telling what day or month it is
Describing some simple objects – colour, size, place
Describing people (from Grade 2)
Expressing ability (can) (from Grade 2)
Expressing likes/dislikes
Giving information about everyday activity – food
Suggesting activities (from Grade 3)
Accepting/declining (from Grade 3)
Agreeing/disagreeing (from Grade 3)
Expressing opinions (from Grade 3)
Offering (from Grade 3)

1.2 Grammatical areas

This list illustrates, using English examples, the grammatical structures minimally required to express the above language purposes and functions at the Breakthrough stage.

Imperatives, in context of games and classroom instructions
Singular and plural nouns
What's this? It's a . . . . What are these? They're. . . . . .
What is there (in your classroom, living room, etc.) There's a. . . . . There are. . . . . .
Have you got . . . . . .? I've got/I haven't got. . . . . .
What do you like/eat. . . . . .? (with nouns) I like. . . . . .
Where do you live? I live in. . . . . street.
Parts of present tense of To be – I'm, he's/she's/it's. . . . My name is. . . . .
How old are you / is he/she ?
How many. . . . . .are there?
Where's / where are the. . . . .? + location
Articles – masculine/feminine/neuter (where applicable)
Range of common adjectives (and adverbs?)

Range of common present tense verbs
Range of prepositions for location, etc.
Can
Want/like + noun or verb form (e.g. infinitive or gerund)
Must, need
Present tense verb forms
Present used with future reference

1.3 Vocabulary areas

Animals
Basic prepositions of place
Classroom objects
Clothes
Colours
Common adjectives: e.g. big, small
Family
Film, play, concert, music, band, names of instruments

Food and drink
Furniture and other household machines and objects
Leisure and holidays
Methods of communication: email, fax, post, stamps
Names of occupations
Names of sports
Numbers
Parts of body
Places: shops, cinema, park, beach, etc.
The home
Time (including months, days)
Ways of travelling: by bus, by plane, etc.
Weather

Classroom instructions
Come in. Listen! Quiet, please! Look at . . . . . Open your book at page . . . . Close your books. Put your books away. Stand up! Sit down! Touch. . . . . Find. . . . Show me. . . . . Point to . . . . Put your hand up. Hands down. Give me a . . . . . Tell me about. . . . . .

Other classroom phrases
yes, no, please, thank you, excuse me, I don't understand (Mainly for teacher to use): Very good! Excellent! That's right! Again! Everyone! Is this right? Try again. Now, let's begin. . . . Now. . . .

Greetings
Hello. Goodbye. How are you? Fine, thanks. Here! (for register) He/she isn't here.

Phrases used in assessment (receptive)
Are you ready? Let's start now! Tick, colour, box
Letters of the alphabet

Generic Preliminary specification content

1.1 Language purpose and functions

At the Preliminary stage there are three broad categories of language functions:

- imparting and seeking factual information
- expressing and finding out opinion
- socialising.

At the Preliminary stage the realisation of these functions is at a very brief and simple level. The following list is not exhaustive but gives some examples of this.

Greeting and responding to greetings
Introductions – incl. phone
Asking for and giving personal details – incl. forms
Asking and giving info on routines, habits
Expressing and responding to thanks, invitations
Apologies and excuses
Agreement, disagreement, contradiction
Expressing preferences, feelings, opinions, needs, wants
Checking meaning, spellings, asking for repetition, interrupting
Using numbers
Asking and giving date, time, etc.
Buying and selling
Following and giving instructions
Asking way, giving directions, travel info
Describing education, job, people
Describing accommodation
Talking about food, weather, health
Understanding signs and notices
Describing shape, size, use etc. of objects
Expressing purpose, cause, result, reasons
Understanding and producing simple narratives

1.2 Grammatical areas

Candidates are required to show knowledge of a list of grammar and linguistic structures nationally agreed by the regulatory authorities for use by all Awarding Bodies for GCSE.

Language-specific lists of Grammar and Linguistic Structures will be given in language-specific supplements.

Generic Intermediate specification content

1.1 Language purpose and functions

Intermediate stage follows on closely from Preliminary stage and covers the same three broad categories of language functions:

- imparting and seeking factual information
- expressing and finding out opinions
- socialising.

At Intermediate stage, the realisation of these functions:

(a) builds on the specifications at Preliminary stages by adding length and complexity to the language used

(b) uses a greater range of areas.

The following are some examples of functional areas additional at Intermediate stage:

- expressing opinions and making choices
- understanding and producing simple narratives
- justifying opinions and choices and persuading
- describing education, qualifications and skills
- drawing simple conclusions and making recommendations
- criticising and complaining
- talking about physical and emotional feelings.

1.2 Grammatical areas

Candidates are required to show knowledge of a list of grammar and linguistic structures nationally agreed by the regulatory authorities for use by all Awarding Bodies for GCSE.

Language-specific lists of Grammar and Linguistic Structures will be given in language-specific supplements.

Asset Languages: The final list of languages offered from September 2008

Table C1 Final list of languages offered from September 2008

	Breakthrough/ Preliminary/ Intermediate	Advanced	Proficiency	Mastery
Arabic	✓	✓		
Bengali	✓			
Chinese (Cantonese)	✓	✓		
Chinese (Mandarin)	✓	✓		
Cornish	✓			
French	✓	✓	✓	✓
German	✓	✓	✓	
Greek (Modern)	✓	✓		
Gujarati	✓			
Hindi	✓			
Irish	✓			
Italian	✓	✓	✓	
Japanese	✓	✓		
Panjabi	✓			
Polish	✓	✓	✓	
Portuguese	✓	✓		
Russian	✓	✓	✓	
Somali	✓			
Spanish	✓	✓	✓	
Swedish	✓			
Tamil	✓			
Turkish	✓	✓		
Urdu	✓	✓		
Welsh	✓			
Yoruba	✓			

Note: Offered at Breakthrough and Preliminary Stages only

Appendix D

The European Survey on Language Competences

Task types: A complete list

Described below are the full set of Listening and Reading task types in terms of their testing focus, text type, the kind of response elicited, and CEFR levels targeted. Appendix 1 in the Technical Report has examples of all of these task types, for a selection of languages.

Table D1 Main Study Listening task types

Task type ID	Test focus	Text type	Task type	Levels
L1	Identifying key vocabulary/information (e.g. times, prices, days of weeks, numbers, locations, activities).	A simple dialogue.	Candidates match the name of a person to the relevant graphical illustration.	A1 A2
L2	Identifying the situation and/or the main idea (A1/A2) or communicative function (B1/B2).	Series of five short independent monologues or dialogues, e.g. announcements, messages, short conversations, etc.	Candidates choose the correct graphic (A1/ A2) or text (B1/B2) option from a choice of three.	A1 A2 B1 B2
L3	Understanding and interpreting detailed meaning.	A conversation or interview.	True/False	A2
L4	Understanding and interpreting the main points, attitudes and opinions of the principal speaker or speakers.	Dialogue	3-option multiple choice	B1 B2
L5	Understanding and interpreting gist, main points and detail, plus the attitudes and opinions of the speaker.	A longer monologue (presentation, report)	3-option multiple choice	B1 B2

Table D2 Main Study Reading task types

Task type ID	Test focus	Text type	Task type	Level
R1	Identifying factual information relating to personal and familiar themes.	Short personal text (email, postcard, note).	3-option multiple choice with graphic options. Candidates choose the correct option.	A1
R2	Finding predictable factual information in texts such as notices, announcements, timetables, menus, with some visual support.	Notice, announcement etc. on everyday topic, with graphic support.	3-option multiple choice with short text-based options focusing on information. Candidates choose the correct option.	A1 A2
R3	Understanding signs, notices, announcements and/or labels.	A set of notices or signs etc. and a set of statements or graphics paraphrasing the message.	Candidates match the statements or graphics to the correct notices/announcements.	A1 A2
R4	Understanding the main ideas and some details of a text.	A newspaper/magazine article on familiar everyday topic.	Candidates answer 3-option multiple-choice questions.	A2
R5	Understanding information, feelings and wishes in personal texts.	A personal text (email, letter, note).	Candidates answer 3-option multiple-choice questions.	A2 B1
R6	Reading 3 (B1) or 4 (B2) short texts for specific information, detailed comprehension and (at B2) opinion and attitude.	A set of 3 (at B1) or 4 (at B2) short texts (e.g. ads for holidays, films, books), and a list of information/attitudes that can be found in the texts.	Candidates match the information to the text it is in.	B1 B2
R7	Reading for detailed comprehension and global meaning, understanding attitude, opinion and writer purpose. B2: deducing meaning from context, text organisation features.	A text on familiar everyday topic.	Candidates answer 3-option multiple-choice questions.	B1 B2
R8	Understanding text structure, cohesion and coherence.	Text from which sentences are removed and placed in a jumbled order after text.	Candidates match the sentences to the gaps.	B2

Illustration of CEFR levels: Writing

Four of the eight tasks – one at each CEFR level – are presented below to illustrate the progression. The tasks themselves are presented in all five language versions, enabling the reader to judge the comparability of the tasks across languages.

Performances are then presented to exemplify the progression of levels. For each task, a performance which demonstrates ability at the intended level is shown, alongside a performance which fails to achieve the level.

Table D3 An A1 level task: Holiday photo

	EN – Holiday photo
	You are on holiday. Send an email to an English friend with this photo of your holiday. Tell your friend about: • the hotel • the weather • what the people are doing Write 20–30 words.
FR – Photo de vacances	**DE – Urlaubsfoto**
Tu es en vacances. Tu envoies un email à un ami avec cette photo de tes vacances. Tu utilises la photo pour parler de: • l'hôtel • le temps • les activités Tu écris 20–30 mots.	Du hast Ferien. Schreib deiner deutschen Freundin eine E-Mail mit diesem Urlaubsfoto. Schreib deiner Freundin über: • das Hotel • das Wetter • was die Leute machen Schreib 20–30 Wörter.
ES – Foto de vacaciones	**IT – A1 level not tested**
Estás de vacaciones. Envía un e-mail a un amigo español con esta foto de tus vacaciones. Escribe sobre: • el hotel • el tiempo • qué hace la gente Escribe 20–30 palabras.	

Table D4 An A2 level task: New hobby

EN – New hobby	
You have a new hobby. Write an email to an English friend about your hobby. Say: • what your new hobby is • when you started it • why you like it so much Write 25–35 words.	
FR – Nouveau passe-temps préféré	**DE – Neues Hobby**
Tu as commencé une nouvelle activité. Tu écris un email à un ami français et tu lui dis: • quelle est ta nouvelle activité • quand tu as commencé cette activité • pourquoi tu aimes cette activité Tu écris 25–35 mots.	Du hast ein neues Hobby. Schreib einer deutschen Freundin eine E-Mail. Schreib: • Was ist dein neues Hobby? • Wann hast du damit angefangen? • Was gefällt dir an dem Hobby? Schreib 25–35 Wörter.
ES – Nuevo hobby	**IT – Nuovo hobby**
Tienes un nuevo hobby. Escribe un e-mail a un amigo español sobre tu nuevo hobby. En este e-mail debes decir: • cuál es tu nuevo hobby • cuándo empezaste a tenerlo • por qué te gusta tanto Escribe 25–35 palabras.	Tu hai un nuovo hobby. Scrivi un'email a un tuo amico italiano e dici: • qual è il tuo nuovo hobby • quando hai incominciato • perché ti piace tanto Scrivi 25–35 parole.

Table D5 A B1 level task: Favourite family member

EN – Favourite family member	
This is part of an email you receive from an English pen friend:	
In your next email, tell me about someone in your family that you like a lot. What sorts of things do you do together? Why do you get on well with each other?	
Write an email to your friend, answering your friend's questions.	
Write 80–100 words.	
FR – Membre de la famille	**DE – Familienmitglied**
Voici un extrait d'un message que tu as reçu de ta correspondante française.	Von einem deutschen Brieffreund bekommst du eine E-Mail. Darin schreibt er.
Dans ton prochain mail, parle-moi d'un membre de ta famille que tu aimes vraiment beaucoup. Qu'est-ce que vous faites ensemble ? Pourquoi est-ce que vous vous entendez bien tous les deux ?	*Bitte schreibe mir in deiner nächsten E-Mail, wen du in deiner Familie besonders gern magst. Was macht ihr gemeinsam? Warum versteht ihr euch gut? …*
Tu écris un email à ta correspondante française et tu réponds à ses questions.	Schreib eine E-Mail an deinen Freund und antworte auf seine Fragen.
Tu écris 80–100 mots.	Schreib 80–100 Wörter.
ES – Miembro de la familia	**IT – Familiare preferito**
Aquí tienes parte de un e-mail que has recibido de un amigo español.	Questa è una parte di un'email che hai ricevuto da un amico italiano.
En tu próximo e-mail, háblame de alguien de tu familia que te guste mucho. ¿Qué tipo de cosas hacéis juntos? ¿Por qué os lleváis bien?	*Quando mi scriverai la prossima email, parlami di una persona della tua famiglia che ti piace molto. Che tipo di cose fate insieme? Perché andate così d'accordo?*
Escribe un e-mail a tu amigo en el que contestes a las preguntas que te hace.	Scrivi un'email al tuo amico e rispondi alle sue domande.
Escribe 80–100 palabras.	Schreib 80–100 Wörter.

Table D6 A B2 level task: Exchange student

EN – Exchange student

You see this newspaper advertisement:
Experience England!
Exchange trips organised by the StudentWorld agency
Would you like to be an exchange student in an English school and live with an English family?
Apply now for one of only 20 free places!
Tell us:

- *what you would like to learn about life in an English family*
- *what you would like to do with your English classmates*
- *why you think you should be given this opportunity*

Write your letter of application.
Write 120–180 words.

FR – Échanges scolaires

Tu vois cette annonce dans un magazine.
Découvrez la France !
Échanges scolaires organisés par l'agence " Le monde des études"
Aimerais-tu participer à un échange pour découvrir un collège français et vivre dans une famille française ?
Dépose ta candidature maintenant.
Il n'y a que 20 places !
Dis-nous:

- *ce que tu aimerais apprendre en vivant dans une famille française*
- *ce que tu aimerais faire avec tes partenaires du collège français*
- *pourquoi tu penses que cette expérience serait une bonne opportunité pour toi*

Tu écris une lettre de candidature.
Tu écris 120–180 mots.

DE – Austauschschülerin

In einer Zeitschrift findest du diese Anzeige:
Erlebe Deutschland!
Austauschreisen mit der Organisation „Wechselspiel"
Möchtest Du gern als Austauschschülerin eine deutsche Schule besuchen und in einer deutschen Familie leben?
Bewirb dich jetzt auf einen der 20 Plätze!
Schreib uns:

- *Was möchtest du in einer deutschen Familie erleben?*
- *Was möchtest du mit deinen Partnerschülern unternehmen?*
- *Warum bist du der/die Richtige für den Austausch?*

Schreib einen Bewerbungsbrief.
Schreib 120–180 Wörter.

ES – Intercambio de estudiantes

Has visto este anuncio en un periódico.
¡Estudiar en España!
Viajes de intercambio de estudiantes organizados por la agencia "Cosmoeducación".
¿Te gustaría formar parte de un intercambio con un colegio español y vivir con una familia española?
Solicita una de las 20 plazas que quedan libres.
Escribe una carta en la que cuentes:

- *qué te gustaría aprender de una familia española*
- *qué te gustaría hacer con tus compañeros de clase*
- *por qué crees que puedes ser la persona indicada*

Escribe una carta de solicitud.
Escribe 120–180 palabras.

IT – Studiare in Italia

Hai letto in un giornale il seguente annuncio:
Vivi l'Italia!
Programma di scambio studenti organizzato dall'agenzia "Studenti del mondo"
Vorresti partecipare ad un programma di scambio studenti presso scuole e famiglie italiane?
Iscriviti ora: ci sono solo 20 posti disponibili!
Scrivici per dirci:

- *che cosa vorresti imparare vivendo in una famiglia italiana*
- *che cosa ti piacerebbe fare con i tuoi nuovi compagni italiani*
- *perché sei tu la persona giusta*

Scrivi una lettera in risposta all'annuncio.
Scrivi 120–180 parole.

Example performances – English

Task	Achieves A1	Pre-A1
A1 Holiday photo	"Hi! I living in Hotel Bellevue and this is nice, We have swimming pool and a nice resturant. The weather is very good, its sunny and very hot. And the people play vollyball and they are nice. Good bye!"	They play voleyball. The namn of the hotel is Belleevue. Have a greates tree.
	Achieves A2	**Still at A1**
A2 New hobby	Dear Lynda, How are you? I want to tell you something. I have a new hobby, my new hobby is playing playstation. I started a month ago. I like it because you have different games for it and, it is just so much fun. You have to come and play with me sometime Lots of love, Maria	Halo! I have new hobby and this is listen to the music. For this hobby I started when I will 13 years old. This hobby I like so much, because I like music and I like sing.
	Achieves B1	**Still at A2**
B1 Favourite family member	Hallo, My family is great and I love it, but I love my mother the most. We always going shoping together, or do some funny different stuff. I love when we watching a scary movie. We making so much popcoen and laughing all the time. My mother is always with me, that is why I love her so much. She is the strongest person in the world. It is so funny with her. We love singing and my father goes crazy. In the winter we always go skiing and that is one of the best things in the year. I love my family, but my mother is at the top, she is the best.	Dear John, Thanks for your email. In my family I like a lot Marie. It's my sister. I have 3 sisters but I'm going to talk you about Sophie. Sometimes we go shopping together and we kocht a lot of clothes. Marie is very friendly. We talk a lot together about our personnal life: about boys friends, school. It's funny. Last week I wend in her flat in Brussel. She's a student in chemistery, The day we went shopping for find a dress for her. We finded it and she's very beautiful. See you soon Isabelle
	Achieves B2	**Still at B1**
B2 Student exchange	To StudentWorld agency 19th March My name is Nicola Marinova, I'm sixteen years old and I live in Varna, Bulgaria. I saw an advertisment in the newspaper about exchange trips organised by your agency and I want to live with an English family and to be a student in an English school.	My name is Anna Kowalska and I will like to be an exchange student in an Englisch school. I will love to live with an Englisch family and share my life with them. I really want to learn all about Englisch cultur, the food and

Example performances – English (continued)

	Achieves B2	Still at B1
	It's very interesting for me to learn about the life in an ordinary English family. I want to drink English tea with milk and to feel England at all. It will be a pleasure to me when I meet my English classmatess, too. I really want to learn how the students in your country spend their free time and their holidays. I think that England is great country with a variety of enterteiments for young peoples like me. And at the end I think this opportunity should be given to me because I'm really interest about England at all and I think that will be a great chance for me to give a start in my life as an adult."	the language. People say that there is the place of work and money and I really want to know is this thrue. I'll always wanted to be an exchange student and meet new people, make friendz, and have one different life with adventures and who knows what else. I think I'm gut for this and everybody needs to have one chance. I diserve this opportunity

Example performances – French

Task	Achieves A1	Pre-A1
A1 Photo de vacances	"Bounjour Anna. Ça va ? Je suis en vacances avec ma famille. C'est très bien ici ! L'hôtel est supèr, le mange est bon, . . . ! Le temps ici est genial. Tous les jours, il fait du soleil. Je trouve des amis, est nous nageons dans la mer où nous jouons au foot, volleyball, . . . ! À prochaine samedi. Jeanne"	"Ça-va Mathilde ? J'aime la Hotel Bellevue parceque est très belle, les activités sont joer footbol et voleibol, est très bonne. Salut Mathilde !"
	Achieves A2	**Still at A1**
A2 Nouveau passe-temps	Salut ! J'ai commencé une nouvelle activité ! J'ai réalisée une activité de la lecture en semaine passé et j'ai aimé parce que je peux étudier les languages. Bisous !	Salut me ami. Je as commencé une nouvelle activité, est football. Je commencé cette activité en septembre. Je adore fait cette activité pourque je adore sport. Adeus mi ami.
	Achieves B1	**Still at A2**
B1 Membre de la famille	Le membre de ma famille qui j'aime beaucoup c'est mon père. Il est sociable, un vraiment ami, amusant et sympathique. Ensemble, nous jouons au football, volleyball. . . Nous allons au théâtre, au cinéma et nous allons	Je suis Beata Schmidt, j'ai 16 ans, le membre de ma famille que j'aime très beacoup est ma cousine Magda. Je l'aime beacoup parce-que je et elle nos entendons très bien, parce-

Example performances – French (continued)

	Achieves B1	Still at A2
	vu le SLBenfica, au stadium. Il est du FCPorto et je suis du SLBenfica, et quand existe un Porto-Benfica, nous allons au stadium. Psicologiquement, nous sommes passives, amustants et intelligent. Nous nous entendons très bien parce que, simplement, nous sommes père et fills."	que elle m'aime et je l'aime. Nous aime etudié groupé et nous sortons a promener. En fin nous aime beacoup entre nous. Il y a 19 ans et elle etude anasthesie en ovida, en la université. Fin, j'aime ma cousine"
	Achieves B2	**Still at B1**
B2 Échanges scolaires	Bonjour, Je voudrais me presenter à la candidature de votre places en France. Je pense que je suis très bonne studiante et que je pouvait apprendre beaucoup avec notre course. Si je suis avec une famille française je pense que je apprenderais beaucoup de choses et nouvaux mots et expressions. Je seulement apprendre français dans l'école donc votre course est un chose très bonne pour moi. Je voudrait aller à la plage, connaitre nouvelles personnes, aller au cinéma et faire beaucoup de sport. Je voudrait parler avec mes partenaires et apprendre pour ils aussi. Je pourrais decouvrir un nouvaux culture et je pense que ce course m'aiderais à madurer et à vivre pendant quelques jours sans ma famille et mes amis. Ce course pourrait être une bonne opportunité pour moi parce que j'aime beaucoup la culture français et la France. Je pense que tu va choisir moi, j'espere votre reponse. Marie	"Bonjour, je suis Andrzej Belinski et je veux être un candidat. J'aimerais aller en France parce que je ne suis pas vraiment bon en français et la langue de français me semble très chouette. Je aimerais aller à France aussi parce que je veux apprendre de vivre quand un français vivre. J'aimerais que mes partenaires m'aider de parler très bien français. Je voudrais aussi de jouer au football avec eux parce que j'aime très bien le sport football. C'est ma vie. Je pense que cette expérience serait une bonne opportunité pour moi parce que je veux devenir un docteur et si je parle très bien français et l'anglais ça sera un peu plus facile. P.S. Je suis une bonne personne !"

Example performances – German

Task	Achieves A1	Pre-A1
A1 Urlaubsfoto	Hallo Sonja, Wie gehst-du? Was machst-du in dein Urlaub? Ich bin in Hawaï. Ich schlafe in dem „Hotel Bellevue“. Das Wetter ist super. Der Son ist immer das! Ich habe viele Freunde und wir spielen oft Volley. Ich bin glücklich. Bye bye Jan Kowalski	Das Hotel heißst Bellevue. Hotel ist in Adria. Hotel habt viel Windov, and das Auto. Weter ist a wunderschön. Sommer ist, and wunderschön tag. Leute trage t-shirt and Hand. Leute spilen Vollyball in Adria.
	Achieves A2	**Still at A1**
A2 Neues Hobby	Liebe Rose, Ich liebe Kino! Das ist mein neue Hobby! Ich gehe ins Kino wenn gibt es Gute Film zu sehen. Mein Lieblings-Film ist „Harry Potter“ oder „Some like it hot“ mit Marilyn Monroe! Meine Lieblingsschauspielerin sind Marilyn Monroe, Rose McGowen und Shannen Doherty. Ich mage Kino, weil du kannst Film sehen. Angela	An: Andrzej Von: Maria Ich habe ein neues Hobby. Meine neues Hobby ist spielen Volleyball. Ich spielen Volleyball seit drei Jahre. Ich mag spielen Volleyball. Lieben Grußer
	Achieves B1	**Still at A2**
B1 Familien-mitglied	Hallo, wie geht es? Ich schreibe über meinen kleinen Bruder Tadek. Er ist 6 Jahre alt. Er ist sehr komisch und magt Spongebot Swammkopf. Er sieht ihn auf Deutsch. So hat er Deutsch gelernt. Er geht in den Kindergarten und hat Freunde. Wir haben nicht sehr viel gemeinsammes veil wir nicht die selbe Generation sind, aber wir verstehen uns sehr gut. Er errinert mich (an) auf mich wan ich klein war. Nicht mit dem aussehen, aber mit gedanken. Ich mag meinen kleinen Bruder und mag es Zeit mit ihnen zu verbringen.	‘Liebe Darin! Ich ferbringe viele Zeit mit meiner Familie. Meine Mutter und ich spielen viele spiele und sie erzelt mir geshiste uber seine kindertage. Meine Oma ist immer auf der vardets und ich helfe sie. Mit meine Tante ferbringe ich nicht viele zeit aber unser Zeit zuzamen ist lustisch. Mit meinen Onkel ferbringe ich die ganze vochenende. Wir turnen und spielen fußball. Unsere Zeit zuzamen ist lustich aber argern aber der streight ist nicht groß. Wir versteht uns gut veil wir volen z u verstanden uns. Wir sind eine lustige Familie. Liebe gruse, Filip

Example performances – German (continued)

	Achieves B2	Still at B1
B2 Austausch-schülerin	Sehr geehrte Damen und Herren, in der Zeitschrift habe ich diese Anzeige gelesen und möchte mich um einen Platz bewerben. Ich habe den Wunsch, als Austauschschüler eine deutsche Schule zu besuchen. Sehr wichtig für mich ist das leben in einer deutschen Familie. Ich interessiere mich für die deutschen Traditionen. Mich interessiert auch, was die Deutschen gerne essen. Mit meinem Partnerschüler möchte ich in die Schule gehen und einen Unterricht in einer deutschen Schule beobachten. Ich möchte wissen, wie der Schulalltag in einer deutschen Schule aussieht. Ich möchte Mathe und Sport Stunden besuchen. Ich bin der Richtige für den Austausch, weil ich sehr gut Deutsch spreche. Ich möchte die deutschen Kultur und Tradition lernen. Milka Elzinga	Hallo, Ich bin Katarina, ich bin 15 Jahre alt und Ich wohne in Stockholm. Ich möchte als Austauschschüler eine deutsche Schule besuchen und in einer deutschen Familie leben, weil ich deutsch in Deutschland sprechen möchte und ich möchte mehr Deutsch lernen. Ich bin die Richtige für den Austausch, weil ich mag Deutsch und Deutschland. Katarina

Example performances – Italian

Task	Achieves A1	Pre-A1
A1	Not tested	
	Achieves A2	**Still at A1**
A2 Nuovo hobby	Caro Bobby, Come stai? Io sono molto bene perché ho un nuovo hobby. Volevo questo hobby da bambino ma non potevo, perché di darmi avere sedici anni per pratticare questo hobby. Il hobby è guidando un "gokart". Mi piace tanto perché quando guido è vincere mi piace tanto vedere tutti i persone gridano mio nome! Per favore parliami del tuo hobby.	Caro Glenn, Io ho un nuovo hobby. Il nuovo hobby e il calcio. Sono cominciato due anni e sono contento. Mi piace tanto perché e un hobby di fisica. Tuo amico Matthew
	Achieves B1	**Still at A2**
B1 Familiare preferito	Caro Cristoph, Mi hai fatto una domanda nell'ultima email: mi hai detto che vuoi che ti parlo di una persona della mia famiglia che mi piace. Ho pensato un po', e ho deciso che la persona che	Caro Claudio, io sono scrivere quest'email per parlare di una persona della mia famiglia che mi piace molto. Questa persona e il mio padre perché noi

Example performances – Italian (continued)

	Achieves B1	Still at A2
	mi piace di più e mia mamma. Mia mamma mi aiuta quando ho bisognio dell'aiuto. Guardiamo la TV insieme e facciamo molte altre cose. Andiamo molto d'accordo. Penso che questo è perché noi amiamo fare le stesse cose, e allora le facciamo insieme. A presto, Macek	facciamo tante cose insieme. Noi andiamo a giocare calcio o guardare calcio allo stadio. Noi anche andare a pescare insieme e anche cucinare insieme. Noi andiamo così d'accordo perché noi abbiamo molte cose in comune e allora faccio queste cose insieme. Io sono molto cuntento di avere un pardre che ama listessi cose che io ama. Tuo amico Marco
	Achieves B2	**Still at B1**
B2 Studiare in Italia	Caro Signore/a, ho letto nel giornale l'annuncio per lo scambio studenti organizzato da voi, "Studenti del mondo". Dato che è una cosa che m'interessa davvero desidero partecipare in questo programma. Voglio vivere con delle famiglie italiane per seguire la loro vita italiana. Così, posso comparare la loro vita con la nostra, particolarmente il loro modo di fare, tra famiglia e anche con delle persone per loro sconosciute. Voglio esserci anche perché ascoltando l'italiano parlato tutto il giorno mi aiutera veramente tanto. Con i miei nuovi compagni, desidererei visitare dei posti più meravigliosi d'Italia come la Fontana di Trevi e il Colosseo. Credo che sono la persona giusta per questo programma perché sono una persona molto avventurosa e non ho paura di conoscere gente nuova o di essere lontana da casa perché sono indipendente. Grazie per il vostro tempo! Tanti saluti, ..	Studenti del mondo, scrivo quella lettera perche io vorrei partecipare al programma di scambio studenti presso scuole e famiglie italiane. Voglio cominciare con che cosa vorrei imparare vivendo in una famiglia italiana. Io vorrei imparare la cultura italiana, come cosa si mangiano gli italiani, come si vestino gli italiani e la storia italiana. Con i miei amici io piacerei imparare, giocare sport e vivere con loro come i miei figli. Penso che io sono la persona giusta perche sono sincero, responsabile e ho una grande idea di simpatia e generosità. Henryk

Example performances – Spanish

Task	Achieves A1	Pre-A1
A1 **Foto de vacaciones**	Hola, estoy en Hotel Bellevue en Español. Es un Hotel muy grande y bien. Tienes un piscina, un plan de voleybol y más guapo chicas. Hace sol y calor, tengo 30 grados. Español es un país muy impresionante. ¡Ciao! Alejandro	Holla amigo, estoy en vacaciones, estoy en el hotel bellevue y el tiempo es bueno y estoy con sus amigos.
A2 **Nuevo hobby**	**Achieves A2**	**Still at A1**
	Hola! Tengo un muy interesante nuevo hobby. Me gusta montar a caballo. Porque es siempre una aventura muy divertido. Por la mañana montar a caballo con mi amiga. Saludos! Angela	Hola tengo un nuevo hobby, mi nuevo hobby es bandy de sala, es muy divertido.
	Achieves B1	**Still at A2**
B1 **Miembro de la familia**	¡Hola! La persona de mí familia que me gusta mucho es mi hermana. Se llama Agata y tiene veinte años. Me gusta ella porque es muy amable y puedo hablar de todo con ella. No vive en mí casa, pero encuentamos más ó menos cinco veces a mes. La próxima fin de semana hemos ido a un café y un museo de photas. ¡Ha hecho muy divertido! Durante los veranos estamos en una isla juntos. Nos bañamos y tomamos el sol. ¿Y tú tienes alguien en tu familia que te gusta mucho? ¡Escríbeme! Bianca	Yo y mi hermano queremos ir a un luego donde podemos bañar. Hacemos esto aproximadamente una vez cada dos semanas. Está muy divertido. En el verano queremos ir a bici en el pueblo de nosotros, y hacia el mal o unas tiendas. Tambien queremos que sólo estamos en el casa y hablar con nosotros o ver una película. Estamos muy tan mi hermano y yo. Pensamos que cada cosas está divertido. me guste mucho.

Example performances – Spanish (continued)

	Achieves B2	Still at B1
B2 Intercambio de estudiantes	Muy señor mío, Me dirijo a usted en respuesta al anuncio que he visto ayer en la revista de mi instituto, en el que proponen un intercambio con un colegio español. Tengo 16 años y yo soy muy interesada en este anuncio y creo que puedo ser la persona indicada para la beca porque me gusta mucho España. En efecto me gustaría mucho aprender las costumbres de los españoles y por eso quiero vivir en una familia española de manera que vea como es la vida y como pasan sus días los españoles. Me gustaría también ir en el colegio y aprender lo que estudian los chicos de mi edad. Si voy a clases de español me ayudará mejorar mi español y aprender de su manera de hablar. Atentamente, Claudia Schmidt	Me llamo Clément y pienso que soy la persona indicada porque me gusta aprender y soy muy interesado para España. Me gustarío aprender cómo vive una familia española, la cocina española, y perfecionar mi (maîtrise) de la lengua española. Con mis compañeros de clase, me gustarío visitar los monumentos los más conoces de españa, las más grandes (villes) y los paysajes. Con ellos, quiero ver films y hacer varias actividades.

References

Alderson, J C (1991) Bands and scores, in Alderson, J C and North, B (Eds) *Language Testing in the 1990s: The Communicative Legacy*, London: Macmillan, 71–86.

Alderson, J C (2007) The CEFR and the need for more research, *The Modern Language Journal* 91 (4), 659–663.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2004) *The Development of Specifications for Item Development and Classification Within the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Reading and Listening*, Final report of the Dutch CEF construct project, unpublished document.

Alderson, J C, Figueras, N, Kuijper, H, Nold, G, Takala, S and Tardieu, C (2006) Analysing tests of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project, *Language Assessment Quarterly* 3 (1), 3–30.

Alexander, R (2012) *Visions of Education, Roads to Reform: The Global Educational Race and the Cambridge Primary Review*, available online: www.robinalexanderorguk/wp-content/uploads/2012/04/Alexander-Chile-GMU-lecture-1-121030.pdf

American Educational Research Association, American Psychological Association and National Council on Measurement in Education (1999) *Standards for Educational and Psychological Testing*, Washington, DC: American Educational Research Association.

Andersen, Ø E (2011) Semi-automatic ESOL error annotation, *English Profile Journal* 2, available online: journals.cambridge.org/abstract_s2041536211000018.

Andrich, D (1988) *Rasch Models for Measurement*, Newbury Park: Sage Publications.

Angoff, W H (1974) Criterion-referencing, norm-referencing, and the SAT, *College Board Review* 92 (3–5), 21.

Ashton, K (2008) *Comparing proficiency levels in an assessment context: The construct of reading for secondary school learners of German, Japanese and Urdu in England*, doctoral thesis submitted to the Cambridge University Faculty of Education.

Ashton, K (2013) Reflections on the European Survey on Language Competences: Looking back, looking forwards, *Research Notes* 52, 20–23.

Assessment Reform Group (1999) *Assessment for Learning: Beyond the Black Box*, available online: www.nuffieldfoundationorg/sites/default/files/files/beyond_blackbox.pdf

Assessment Reform Group (2002) *Assessment for Learning: 10 Principles*, available online: www.aaiaorguk/content/uploads/2010/06/Assessment-for-Learning-10-principles.pdf

Assessment Systems Corporation (2013) *Assessment Systems Corporation*, available online: http://www.assesscom/index.html

Association for Language Learning (2012) *GCSE results day 2012 – updates, news and comment*, available online: www.all-languagesorguk/news/news_list/gcse_results_day_2012_updates_news_and_comment

Association of Language Testers in Europe (1998) *ALTE News* 7 (2), available online: eventsalteorg/downloads/indexphp?docid=141

Bachman, L F (1990) *Fundamental Considerations in Language Testing,* Oxford: Oxford University Press.

Bachman, L F (1991) What does language testing have to offer? *TESOL Quarterly* 25, 671–704.

Bachman, L F (2005) Building and supporting a case for test use, *Language Assessment Quarterly* 2, 1–24.

Bachman, L F and Palmer, A S (1996) *Language Testing in Practice: Designing and Developing Useful Language Tests,* Oxford: Oxford University Press.

Bachman, L F, Davidson, F, Ryan, K and Choi, I-C (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*, Studies in Language Testing volume 1, Cambridge: UCLES/Cambridge University Press.

Banks, C (1999) *A Review of Candidates Taking More than One Examination*, internal report, Cambridge: Cambridge ESOL.

Beacco, J-C, Byram, M, Cavalli, M, Coste, D, Cuenat, M E, Goullier, F and Panthier, J (2010) *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education*, Strasbourg: Language Policy Division, Directorate of Education and Languages, DGIV Council of Europe.

Beeston, S (2000) The UCLES EFL item banking system, *Research Notes* 1, 5–7.

Bennett, R E (2012) Consequences that cannot be avoided: A response to Paul Newton, *Measurement: Interdisciplinary Research and Perspectives* 10 (1–2).

Bernstein, B (1971) *Class, Codes and Control: Volume 1*, London: Routledge and Kegan Paul.

Black, P (2004) *Raising Standards Through Formative Assessment*, paper presented at GTC conference – New Relationships: Teaching, Learning and Accountability, London.

Bond, T G and Fox, C M (2001) *Applying the Rasch Model*, New Jersey: Lawrence Erlbaum Associates.

Boud, D (2006) Foreword in Carless, D, Joughlin, G, Liu, N F and Associates, *How Assessment Supports Learning: Learning-oriented Assessment in Action*, Hong Kong: Hong Kong University Press, ix–x.

Bramley, T (2005) A rank-ordering method for equating tests by expert judgment, *Journal of Applied Measurement* 6 (2), 202–223.

Breton, G (2008) *Cross-language benchmarking seminar to calibrate examples of spoken production in English, French, German, Italian and Spanish with regard to the six levels of the Common European Framework of Reference for Languages (CEFR)*, le Centre international d'études pédagogiques, Sèvres, 23–25 June 2008.

Bung, K (1973) *The Foreign Language Needs of Hotel Waiters and Staff*, Strasbourg: Council of Europe.

Byram, M and Parmenter, L (Eds) (2012) *The Common European Framework of Reference: The Globalisation of Language Education Policy*, Bristol: Multilingual Matters.

Cambridge Assessment Network (2010) *Successful Assessment Reform: A Farewell to the Assessment Reform Group*, Cambridge: Cambridge Assessment Network.

Cambridge ESOL (2007) *European Survey on Language Competences Tender Submission,* unpublished document, Cambridge ESOL.

Cambridge ESOL (2011) *Using the CEFR: Principles of Good Practice,* available online: www.cambridgeenglishorg/images/126011-using-cefr-principles-of-good-practice.pdf

Cambridge ESOL (2012) *Research Notes* 50.

Carless, D (2007) Learning-oriented assessment: conceptual issues and practical implication, *Innovations in Education and Teaching International* 44 (1), 57–66.

Cizek, G J (2001) *Setting Performance Standards: Concepts, Methods and Perspectives*, New Jersey: Lawrence Erlbaum Publishers.

Cizek, J and Bunch, M B (2007) *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests,* London: Sage Publications.

Coleman, J A (1996) *Studying Languages: A Survey of British and European Students*, London: The National Centre for Languages.

Coleman, J A, Galaczi, A and Astruc, L (2007) Motivation of UK school pupils towards foreign languages: a large-scale survey at Key Stage 3, *The Language Learning Journal* 35 (2), 245–281.

Coste, D (1976) *Un niveau-seuil: Systèmes d'apprentissage des langues vivantes par les adultes*, Strasbourg: Conseil de la coopération culturelle du Conseil de l'Europe.

Council of Europe (1998) *Modern Languages: Learning, Teaching, Assessment – A Common European Framework of Reference*, Strasbourg: Language Policy Division.

Council of Europe (2001) *Common European Framework of Reference for Languages: Learning, Teaching, Assessment,* Cambridge: Cambridge University Press.

Council of Europe (2003) *Relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Manual, Preliminary Pilot Version*, Strasbourg: Language Policy Division, available online: www.coeint/t/dg4/linguistic/Manuel1_ENasp

Council of Europe (2006) *The Common European Framework of Reference for Languages (CEFR) Report to Intergovernmental Policy Forum,* Strasbourg: Council of Europe.

Council of Europe (2007) *The Common European Framework of Reference for Languages (CEFR) and the Development of Language Policies: Challenges and Responsibilities,* paper presented at the Intergovernmental Policy Forum, Strasbourg, 6–8 February 2007, available online: www.coeint/lang

Council of Europe (2009) *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR) A Manual,* available online: www.coeint/t/dg4/linguistic/manuel1_enasp

Council of Europe (2010) *Guide for the Development and Implementation of Curricula for Plurilingual and Intercultural Education*, Strasbourg: Council of Europe.

Council of Europe (2011) *Manual for Language Test Development and Examining*, ALTE on behalf of the Language Policy Division, Council of Europe, available online: www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Council of Europe (2012) *The CEFR and Language Examinations: A Toolkit*, available online: www.coeint/t/dg4/linguistic/Manuel1_ENasp#Illustrations

Cronbach, L J (1971) Test validation, in Thorndike, R L (Ed) *Educational Measurement* (2nd edition), Washington, DC: American Council on Education, 443–507.

Culej, J B (2013) The European Survey on Language Competences in Croatia: Results and implications, *Research Notes* 52, 16–20.

Cummins, J (1979) Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters, *Working Papers on Bilingualism* 19, 121–129.

Cummins, J (1984) *Bilingualism and Special Education: Issues in Assessment and Pedagogy,* Clevedon: Multilingual Matters.

Datta, M (2000) *Bilinguality and Literacy: Principles and Practice*, London: Continuum.

Dearing, R and King, L (2006) *Languages Review: Consultation Report*, London: Department for Education and Skills.

Department for Education (2007) *The Languages Ladder – Steps to Success,* Manchester: Department for Education.

Department for Education (2013) *Modern Languages: GCSE Subject Content and Assessment Objectives*, Manchester: Department for Education, available online: www.education.gov.uk/consultations/downloadableDocs/GCSE%20Modern%20Language_final.pdf

Department for Education and Skills (2002) *Languages for All, Languages for Life – A Strategy for England,* available online: webarchivenationalarchivesgovuk/20130401151715/https://www.educationgovuk/publications/eOrderingDownload/DfESLanguagesStrategypdf

Department for Education and Skills (2004) *14–19 curriculum and qualifications reform: final report of the Working Group on 14–19 Reform,* available online: http://newsbbccouk/1/shared/bsp/hi/pdfs/15_02_05_tomlinson.pdf

Department for Education and Skills (2007) *Languages Review,* available online: webarchivenationalarchivesgovuk/20070506100225/http://teachernetgovuk/_doc/11124/LanguageReview.pdf

Department for Education and Skills and the Qualifications and Curriculum Authority (1999) *The National Curriculum for England: Modern Foreign Languages*, London: Department for Education and Skills and the Qualifications Curriculum Authority.

Ebbinghaus, H (1885) *Memory: A Contribution to Experimental Psychology,* English translation, available online: psyedasuedu/~classics/Ebbinghaus/index.htm

European Commission (2005) *Commission Communication of 1 August 2005 – The European Indicator of Language Competence, COM(2005) 356 final – Not published in the Official Journal,* available online: europaeu/legislation_summaries/education_training_youth/lifelong_learning/c11083_enhtm

European Commission (2012a) *First European Survey on Language Competences: Final Report,* Luxembourg: Publications Office of the European Union, available online: eceuropaeu/languages/eslc/index.html

European Commission (2012b) *First European Survey on Language Competences: Technical Report*, Luxembourg: Publications Office of the European Union, available online: eceuropaeu/languages/eslc/index.html

European Commission (2013) *European Survey on Language Competences and European Benchmark*, Luxembourg: Publications Office of the European Union, available online: eceuropaeu/languages/languages-of-europe/language-competence_en.htm

Figueras, N and Noijons, J (Eds) (2009) *Linking to the CEFR Levels: Research Perspectives*, Arnhem: Cito/EALTA.

Flux, T (2001) *KET / YLE Link Project 2001*, internal report, Cambridge: Cambridge ESOL.

Frederiksen, N, Mislevy, R J and Bejar, I I (1993) *Test Theory for a New Generation of Tests*, London: Routledge.

Fulcher, G (2008) Testing times ahead? *Liaison Magazine* 1, 20–24.

Galaczi, E D (2010) Face-to-face and computer-based assessment of speaking: Challenges and opportunities, in Araújo, L (Ed) *Computer-based Assessment of Foreign Language Speaking Skills,* Luxemburg: European Union, 29–51.

Galaczi, E D and Khalifa, H (2009) *Project overview: Examples of speaking performances at CEFR levels A2 to C2,* available online: www.cambridgeenglishorg/research-and-validation/fitness-for-purpose/

Galaczi E D and Weir, C J (Eds) (2013) *Exploring Language Frameworks: Proceedings from the ALTE Kraków Conference July 2011*, Studies in Language Testing volume 36, Cambridge: UCLES/Cambridge University Press.

Galaczi, E D, ffrench, A, Hubbard, C and Green, A (2011) Developing assessment scales for large-scale speaking tests: a multiple-method approach, *Assessment in Education: Principles, Policy & Practice* 18 (3), 217–237.

Gass, SM, Doughty, C J and Long, M H (2007) *Input and Interaction: The Handbook of Second Language Acquisition,* Malden: Blackwell Publishing.

Geach, J (1996) Community languages, in Hawkins, E (Ed) *30 Years of Language Teaching,* London: CILT, 141–152.

Geranpayeh A, and Taylor L (Eds) (2013) *Examining Listening: Research and Practice in Assessing Second Language Listening*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Goldstein, H (2012) Francis Galton, measurement, psychometrics and social progress, *Assessment in Education: Principles, Policy & Practice* 19 (2), 147–158.

Green, A (2012) *Language Functions Revisited: Theoretical and Empirical Bases For Language Construct Definition Across the Ability Range,* English Profile Studies volume 2, Cambridge: UCLES/Cambridge University Press.

Hambleton, R K, Swaminathan, H and Rogers, H J (1991) *Fundamentals of Item Response Theory Volume 2,* Newbury Park: Sage.

Hammersley, M (2005) Is the evidence-informed practice movement doing more good than harm? Reflections on Iain Chalmers' case for research-based policy making and practice, *Evidence & Policy: A Journal of Research, Debate and Practice* 1 (1), 85–100.

Hawkey, R and Barker, F (2004) Developing a common scale for the assessment of writing, *Assessing Writing* 9 (3), 122–159.

Hawkey, R and Milanovic, M (2013) *Cambridge English Exams–The First Hundred Years, A History of English Language Assessment from the University of Cambridge 1913–2013*, Studies in Language Testing volume 38, Cambridge: UCLES/Cambridge University Press.

Hawkins, E W (1999) Foreign language study and language awareness, *Language Awareness* 8, 124–142.

Hawkins, J A and Filipović, L (2012) *Criterial Features in L2 English: Specifying the Reference Levels of the Common European Framework*, English Profile Studies volume 1, Cambridge: Cambridge: UCLES/Cambridge University Press.

Henning, G (1987) *A Guide to Language Testing: Development, Evaluation, Research,* Boston: Thomson Heinle.

Hudson, T (2005) Trends in assessment scales and criterion-referenced language assessment, *Annual Review of Applied Linguistics* 25, 205–227.

Huhta, A, Luoma, S, Oscarson, M, Sajavaara, K, Takala, S and Teasdale, A (2002) *A diagnostic language assessment system for adult learners*, in Alderson, J C (Ed) *Case Studies in the Use of the Common European Framework*, Strasbourg: Council of Europe, 130–146.

Hulstijn, J H (2007) The Shaky Ground Beneath the CEFR: Quantitative and Qualitative Dimensions of Language Proficiency, *The Modern Language Journal* 91 (4), 663–667.

International Association for the Evaluation of Educational Achievement (2012) *About TIMSS 2011*, available online: timssandpirlsbcedu/timss2011/indexhtml

Jones, N (1992) *An item bank for testing English language proficiency: Using the Rasch model to construct an objective measure,* unpublished PhD, Edinburgh: University of Edinburgh.

Jones, N (1993) *The Singapore Telecom / UCLES Computer-Adaptive Testing Project*, internal document, Cambridge: Cambridge ESOL.

Jones, N (2000a) BULATS: A case study comparing computer based and paper-and pencil tests, *Research Notes* 3, 10–13.

Jones, N (2000b) *Describing levels*, unpublished paper.

Jones, N (2001) The ALTE 'Can Do' Project and the role of measurement in constructing a proficiency framework, *Research Notes* 5, 5–8.

Jones, N (2002) *Relating the ALTE framework to the Common European Framework of Reference*, in Alderson, J C (Ed) *Case Studies in the Use of the Common European Framework*, Strasbourg: Council of Europe, 167–183.

Jones, N (2004) *Construct Work for LLAS 10-04*, internal document, Cambridge: Cambridge ESOL.

Jones, N (2005a) Raising the Languages Ladder: constructing a new framework for accrediting foreign language skills, *Research Notes* 19, 15–19.

Jones, N (2005b) *High level Task Types: Proposed Feature Set for Describing Task Types on LIBS – Final Draft*, internal document, Cambridge: Cambridge ESOL.

Jones, N (2006) *The impact of the CEFR on language testing in Europe*, paper presented at the symposium 'A new direction in foreign language education: The potential of the Common European Framework of Reference for Languages', Osaka University of Foreign Studies, Japan, March 2006.

Jones, N (2007) Assessment and the National Languages Strategy, *Cambridge Journal of Education* 37 (1), 17–33.

Jones, N (2009a) A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, in Figueras, N and Noijons, J (Eds) *Linking to the CEFR Levels: Research Perspectives*, Arnhem: Cito, Council of Europe, EALTA, available online: www.coeint/t/dg4/linguistic/Proceedings_CITO_EN.pdf

Jones, N (2009b) A comparative approach to constructing a multilingual proficiency framework: constraining the role of standard setting, *Research Notes* 37, 6–9.

Jones, N (2012) Reliability and dependability, in Fulcher, D and Davidson, F (Eds) *The Routledge Handbook of Language Testing*, London: Routledge, 350–262.

Jones, N (2013a) Defining an inclusive framework for languages, in Galaczi E D and Weir, C J (Eds) *Exploring Language Frameworks: Proceedings from the ALTE Kraków Conference July 2011*, Studies in Language Testing volume 36, Cambridge: UCLES/Cambridge University Press, 105–117.

Jones, N (2013b) The European Survey on Language Competences and its significance for Cambridge English Language Assessment, *Research Notes* 52, 2–7.

Jones, N and Benton, T (2012) *GCSE – CEFR Comparison From European Survey Data*, unpublished paper.

Jones, N and Saville, N (2007) Scales and frameworks, in Spolsky, B and Hult, F M (Eds) *The Handbook of Educational Linguistics*, London: Wiley-Blackwell, 495–509.

Jones, N and Thighe, D (2005) *CB BULATS Plurilingual Study: An examination of the comparability of CB BULATS English and Foreign Language Versions,* internal report, Cambridge: Cambridge ESOL.

Jones, N Ashton, K and Chen S Y (2005) Rising to the challenge of Asset Languages, *Research Notes* 19, 2–4.

Jones N, Ashton K and Walker, T (2010) Asset Languages: a case study of piloting the CEFR Manual, in Martyniuk, W (Ed) *Aligning Tests to the CEFR: Reflections on the Use of the Council of Europe's Draft Manual,* Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press, 227–246.

Jones, N, Saville, N and Hamilton, M (2013) *A systemic view of assessment within an educational context*, paper presented at ILTA/AAAL symposium, Dallas, 2013.

Kane, M T (1992) An argument-based approach to validity, *Psychological Bulletin* 12, 527–535.

Kane, M T (2004) Certification testing as an illustration of argument-based validation, *Measurement: Interdisciplinary Research and perspectives* 2 (3), 135–170.

Kane, M T (2012) All validity is construct validity or is it? *Measurement: Interdisciplinary Research and Perspectives* 10 (1–2), 66–70.

Keenan, B (2000) National curriculum, *Community Languages Bulletin* 6.

Keenan E L and Comrie, B (1977) Noun Phrase Accessibility and Universal Grammar, *Linguistic Inquiry* 8 (1), 63–99.

Kenyon, D and Malone, M (2010) Investigating examinee autonomy in a computerized test of oral proficiency, in Araujo, L (Ed), *Computer-based Assessment of Foreign Language Speaking Skills*, Luxembourg: European Union, 1–28.

Khalifa, H and Weir, C J (2009) *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

King, L (2011) *Overview of The National Languages Strategy 2003–2011*, available online: www.languagescompanycom/images/stories/docs/news/national_language_strategy.pdf

Kolen, M J and Brennan, R L (2004) *Test Equating: Methods and Practices* (2nd edition), New York: Springer-Verlag.

Lacey, C and Lawton, D (1981) *Issues in Evaluation and Accountability,* London: Methuen.

Laming, D (2004) *Human Judgment: The Eye of the Beholder,* London: Thomson.
Lantolf, J P and Frawley, W (1992) Rejecting the OPI – again: A response to Hagen, *ADFL Bulletin* 23 (2), 34–37.
Larsen-Freeman, D and Cameron, L (2008) *Complex Systems and Applied Linguistics,* Oxford: Oxford University Press.
Leung, C (2004) Developing formative teacher assessment, *Language Assessment Quarterly* 1 (1), 19–41.
Lim, G S (2012) Developing and validating a mark scheme for writing, *Research Notes* 49, 6–10.
Linacre, J M (2006) Rasch Analysis of Rank-Ordered Data, *Journal of Applied Measurement* 7 (11), 129–139.
Linacre, J M (2011) *FACETS*, available online: www.winstepscom/facets.htm
Little, D (2006) The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact, *Language Teaching* 39, 167–190.
Little, D, Lazenby Simpson, B and O'Connor, F (2002) Meeting the English language needs of refugees in Ireland, in Alderson, J C (Ed) *Case Studies in the Use of the Common European Framework*, Strasbourg: Council of Europe, 53–67.
Long, M (1996) The role of the linguistic environment in second language acquisition, in Ritchie, W and Bhatia, T (Eds) *Handbook of Second Language Acquisition*, San Diego: Academic Press, 413–68.
Macaro, E (2001) *Learning Strategies in Foreign and Second Language Classrooms*, London: Continuum.
Maris, G (2009) Standard setting from a psychometric point of view, in Figueras, N and Noijons, J (Eds) (2009) *Linking to the CEFR Levels: Research Perspectives*, Arnhem: Cito/EALTA, 59–65.
Martyniuk, W (Ed) (2010) *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual,* Studies in Language Testing volume 33, Cambridge: UCLES/Cambridge University Press.
McKenna, S (2013) The European Survey on Language Competences and the media, *Research Notes* 52, 23–24.
Messick, S (1989) Validity, in Linn, R L (Ed) *Educational Measurement*, New York: American Council on Education and Macmillan, 3rd edn, 1–103.
Milanovic, M (1996) Series editor's note in Bachman, L F, Davidson, F, Ryan, K and Choi, I-C (1995) *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study*, Cambridge: Cambridge University Press.
Milanovic, M (2009) Cambridge ESOL and the CEFR, *Research Notes* 37, 2–5.
Mislevy, R J (1996) Test theory reconceived, *Journal of Educational Measurement* 33 (4), 379–416.
Mislevy, R J, Steinberg, L S and Almond R G (2002) Design and analysis in task-based language assessment, *Language Testing* 19 (4), 477–496.
Mislevy, R J, Steinberg, L S and Almond, R G (2003) On the structure of educational assessments, *Measurement: Interdisciplinary Research and Perspectives* 1 (1), 3–62.
Mitchell, R (2003) Rethinking the concept of progression in the national curriculum for modern foreign languages: A research perspective, *Language Learning* 27, 15–23.

Morrow, K (Ed) (2004) *Insights from the Common European Framework*, Oxford: Oxford University Press.
National Research Council (2003) *Understanding Others, Educating Ourselves: Getting More from International Comparative Studies in Education*, Washington, DC: The National Academies Press, available online: www.napedu/openbookphp?isbn=0309088550.
NESTA Futurelab (2002) *Literature Review in Languages, Technology and Learning*, available online: archivefuturelaborguk/resources/documents/lit_reviews/Languages_Reviewpdf
Newton, P E (2012) Clarifying the consensus definition of validity, *Measurement: Interdisciplinary Research and Perspectives* 10 (1–2), 1–29.
North, B (1995) The development of a common framework scale of descriptors of language proficiency based on a theory of measurement, *System* 23, 445–465.
North, B (1996) *The development of a common framework scale of language proficiency,* PhD thesis, New York: Peter Lang, Thames Valley University.
North, B (2000) *The Development of a Common Framework Scale of Language Proficiency*, New York: Peter Lang.
North, B and Jones, N (2009) *Further Material on Maintaining Standards Across Languages, Contexts and Administrations by Exploiting Teacher Judgment and IRT Scaling*, Strasbourg: Council of Europe.
Nuffield Foundation (2000) *Languages: The Next Generation*, available online: languagesnuffieldfoundationorg/filelibrary/pdf/languages_finalreport.pdf
Nuffield Foundation (2002) *A Learning Ladder for Languages: Possibilities, Risks and Benefits*, available online: languagesnuffieldfoundationorg/filelibrary/pdf/learning_ladderpdf
O'Malley, J M and Chamot, A (1990) *Learning Strategies in Second Language Acquisition*, Cambridge: Cambridge University Press.
Organisation for Economic Co-operation and Development (2012) *Strong Performers and Successful Reformers in Education: Lessons from PISA for Japan OECD*, available online: http://www.oecdorg/
Oxford, Cambridge and RSA Examinations (2007) *Report to Centres on External Assessment*, Cambridge: Oxford, Cambridge and RSA Examinations.
Oxford, Cambridge and RSA Examinations/Cambridge ESOL (2005a) *Asset Languages Breakthrough Stage Listening, Speaking, Reading and Writing for Assessment from September 2007*, Cambridge: Oxford, Cambridge and RSA Examinations.
Oxford, Cambridge and RSA Examinations/Cambridge ESOL (2005b) *OCR Asset Languages Scheme (Pilot) Preliminary Stage,* Cambridge: Oxford, Cambridge and RSA Examinations.
Panayides, P, Robinson, C and Tymms, P (2009) The assessment revolution that has passed England by: Rasch measurement*, British Educational Research Journal* 36 (4), 611–626.
Pellegrino, J (2003) *The 'second transformation' – issues in combining advances in the learning sciences with IT capabilities*, paper presented to the NRC ILIT Committee, January 2003.
Pellegrino, J W, Chudowsky, N and Glaser, R (2001) *Knowing What Students Know: The Science and Design of Educational Assessment*, Washington, DC: National Academy Press.
Perlmann-Balme, M (2013) European Survey on Language Competences – comparability of A1 level competences across five languages, in Galaczi, E D

and Weir, C J (Eds) *Exploring Language Frameworks: Proceedings from the ALTE Kraków Conference July 2011*, Studies in Language Testing volume 36, Cambridge: UCLES/Cambridge University Press, 85–102.

Pollitt, A and Murray, N L (1996) What raters really pay attention to, in *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium (LTRC), Cambridge and Arnhem* 3, 74–91.

Purpura, J (2004) *Assessing Grammar*, Cambridge: Cambridge University Press.

Purpura, J (2009) The impact of large-scale and classroom-based language assessments on the individual, *Studies in Language Testing Matters* 31 (17), 301–221.

Qualifications and Curriculum Authority (2008) *Grade Standards in GCSE Modern Foreign Languages*, London: Qualifications and Curriculum Authority.

Robinson, M (2013) Innovation in language test development, *Research Notes* 52, 7–13.

Royal-Dawson, L (1994) *Candidates taking 2 exams June 94*, unpublished internal document.

Sahlberg, P (2011) *Finnish Lessons: What Can the World Learn from Educational Change in Finland?* New York: Teachers College Press.

Saville, N (1995) *The ALTE Framework Project – Phase Three: The validation and grouping at critical levels of the 'Can Do' statements*, unpublished internal document.

Saville, N (2003) The process of test development and revision within UCLES EFL, in Milanovic, M and Weir, C (Eds), *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002* volume 15, Cambridge: UCLES/Cambridge University Press, 57–120.

Saville, N (2005) An interview with John Trim at 80, *Language Assessment Quarterly* 2 (4), 263–288.

Saville, N (2009) *Developing a model for investigating the impact of language assessment within educational contexts by a public examination provider*, unpublished PhD thesis, University of Bedfordshire.

Saville, N (2011) *An interview with John Trim and Nick Saville*, available online: www.youtube.com/

Saville, N (2012) Applying a model for investigating the impact of language assessment within educational contexts: The Cambridge ESOL approach, *Research Notes* 50, 4–8.

Sayer, A (1992) *Method in Social Science: A Realist Approach,* London and New York: Routledge.

Schleicher, A (2012) The role and value of international datasets, *Research Intelligence* 119, 10–11.

Shaw, S and Weir, C J (2007) *Examining Second Language Writing: Research and Practice in Assessing Second Language Writing,* Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Shohamy, E (1996) Competence and performance in language testing, in Brown, G, Malmkjaer, K and Williams, J (Eds) *Performance and Competence in Second Language Acquisition*, Cambridge: Cambridge University Press, 136–151.

Smith, P C and Kendall, J M (1963) Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales, *Journal of Applied Psychology* 47 (2), 149–155.

Sneddon, R (2000) Language and literacy: Children's experiences in multilingual environments, *International Journal of Bilingual Education and Bilingualism*, 3 (4), 265–282.
Spolsky, B (1986) A Multiple Choice for Language Testers, *Language Testing* 3 (2), 147–158.
Sturman, L (2012) *Making Best Use of International Comparison Data*, Research Intelligence 119, London: British Educational Research Association.
SurveyLang (2008) *Inception Report for the European Survey on Language Competences,* Luxembourg: Publications Office of the European Union.
Swain, M (1985) Communicative competence: Some roles of comprehensible input and comprehensible output in its development, in Gass, S and Madden, C (Eds) *Input in Second Language Acquisition,* Rowley: Newbury House, 235–253.
Szpotowicz, M (2013) The European Survey on Language Competences – the Polish experience, *Research Notes* 52, 13–16.
Tattersall, K (2004) *Teacher assessment: realistic objective or impossible dream?* paper presented at AEA Europe Conference, Budapest, November 2004.
Taylor, L (Ed) (2011) *Examining Speaking: Research and Practice in Assessing Second Language Speaking,* Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.
Taylor, L and Jones, N (2006) ESOL exams and the Common European Framework of Reference (CEFR) *Research Notes* 24, May 2–5.
The British Academy (2013) *Languages: The State of the Nation: Demand and Supply of Language Skills in the UK Summary Report,* London: The British Academy.
The Languages Company (2013) *The Languages Company*, available online: www.languagescompany.com/
Thorndike, E L (1920) A constant error in psychological ratings, *Journal of Applied Psychology* 4 (1), 25–29.
Thurstone, L L (1927) A law of comparative judgment, *Psychological Review* 3, 273–86.
University of Cambridge Local Examinations Syndicate (2000) EFL Research at UCLES, *Research Notes* 1.
University of Cambridge Local Examinations Syndicate (2008) *Research Notes* 33.
Van Ek, J A (1975) *Systems Development in Adult Learning: The Threshold Level in a European Unit/Credit System for Modern Language Learning by Adults,* Strasbourg: Council of Europe Press.
Van Ek, J A and Alexander, L G (1977) *Waystage,* Strasbourg: Council of Europe.
Van Ek, J A and Trim, J L M (1990a/1998a) *Threshold 1990*, Cambridge: Cambridge University Press.
Van Ek, J A and Trim, J L M (1990b/1998b) *Waystage 1990*, Cambridge: Cambridge University Press.
Van Ek, J A and Trim, J L M (2001) *Vantage*, Cambridge: Cambridge University Press.
Van Moere, A (2010) Automated spoken language testing: test construction and scoring model development, in Araujo, L (Ed) *Computer-based Assessment (CBA) of Foreign Language Speaking Skills*, Brussels: European Union, 84–99.
Weir, C J (2005a) *Language Testing and Validation: An Evidence-Based Approach*, Oxford: Palgrave.

Weir, C J (2005b) Limitations of the Council of Europe's Framework of reference (CEFR) in developing comparable examinations and tests, *Language Testing* 22 (3), 281–300.
Weir, C J (2013a) An overview of the influences on English language teaching in the United Kingdom 1913-2012, in Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 1-102.
Weir, C J (2013b) Case study: A quantitative analysis of the context validity of the CPE reading passages used in translation tasks (1913–88), summary tasks (1930–2010) and comprehension question (MCQ/SAQ) tasks (1940–2010), in Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012*, Studies in Language Testing volume 37, Cambridge: UCLES/Cambridge University Press, 472–537.
Weir, C J and Milanovic, M (Eds) (2003) *Continuity and Innovation: Revising the Cambridge Proficiency in English Examination 1913–2002*, Studies in Language Testing volume 15, Cambridge: UCLES/Cambridge University Press.
Weir, C J and Taylor, L (2011) Conclusions and recommendations, in Taylor, L (Eds) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press, 293-313.
Weir, C J, Huizhong, Y and Yan, J (Eds) (2000) *An Empirical Investigation of the Componentiality of L2 Reading in English for Academic Purposes*, Studies in Language Testing volume 12, Cambridge: UCLES/Cambridge University Press.
Weir, C J, Vidaković, I and Galaczi, E D (2013) *Measured Constructs: A History of Cambridge English Language Examinations 1913–2012* Studies in Language Testing volume 37, Cambridge: UCLES Cambridge University Press.
WIDA (2012) *English Language Proficiency (ELP) Standards 2012 edition*, available online: www.widaus/standards/elpaspx
Widdowson, H G (1978) *Teaching Language as Communication*, Oxford: Oxford University Press.
Wiliam, D (2007) Once you know what they've learned, what do you do next? Designing curriculum and assessment for growth, in Lissitz, R (Ed) *Assessing and Modeling Cognitive Development in School*, Maple Grove: JAM Press, 241–270.
Wilkins, D A (1976) *Notional Syllabuses*, Oxford: Oxford University Press.
Winsteps (2013) *Winsteps*, available online: http://www.winsteps.com
Wolfe E W, (2004) Identifying rater effects using latent trait models, *Psychology Science* 46 (1), 35–51.
Wood, R (1991) *Assessment and Testing: A Survey of Research*, Cambridge: Cambridge University Press.
Wright, B D and Stone, M H (1979) *Best Test Design,* Chicago: MESA Press.

Author index

A
Alderson, J C 8, 11, 40, 56, 93, 94
Alexander, L G 88,
Alexander, R 199, 212
Almond, R G 10
American Educational Research Association 46, 47
American Psychological Association 46, 47
Andrich, D 37
Angoff, W H 29
Ashton, K 44, 138–140, 156, 202
Assessment Reform Group 112, 141, 151
Assessment Systems Corporation 70
Association for Language Learning 152
Associations of Language Testers in Europe 2, 3, 6, 10, 27, 45, 61, 62, 64, 65, 75-84, 87, 89, 93, 154, 155, 158, 175, 176, 203, 220
Astruc, L 139

B
Bachman, L F 1, 2, 7, 9, 55, 109, 203, 214
Banks, C 73
Barker, F 39, 40, 89
Beacco, J-C 30
Beeston, S 85
Bejar, I I 15
Bennett, R E 47
Benton, T 150
Bernstein, B 208
Black, P 106
Bond, T G 17
Boud, D 211
Bramley, T 41
Brennan, R L 36
Breton, G 41, 42, 55
Bunch, M B 43
Byram, M 30, 96

C
Cambridge Assessment Network 151
Cambridge ESOL 7, 35, 45, 58, 61, 69–71, 74, 82, 84, 85, 87–89, 91, 96, 115, 128, 138, 139, 143, 158, 159, 176
Cameron, L 213
Carless, D 211
Cavalli, M 30
Chamot, A 124
Chen, S Y 138
Choi, I-C 1
Cizek, G J 43, 56
Coleman, J A 110, 139
Comrie, B 21, 22
Coste, D 30, 88
Council of Europe 2, 7, 10, 28, 31, 33, 35, 39, 40, 42–45, 50–53, 57, 60, 77, 87–90, 93–98, 101, 139, 156, 158, 159, 176, 178, 184, 207, 208, 210
Cronbach, L J 49
Cuenat, M E 30
Culej, J B 202
Cummins, J 54, 207

D
Datta, M 111
Davidson, F 1
Dearing, R 111, 114
Department for Education 115, 152
Department for Education and Skills 100, 103, 104, 105, 114, 118, 135, 137, 139–141
Department for Education and Schools and the Qualifications Curriculum Authority 139
Doughty, C J 214

E
Ebbinghaus, H 74
European Commission 4, 154, 155, 157, 159, 161, 179, 181, 182, 184, 186, 188, 190, 192, 193, 195, 196, 199, 200, 201

F
ffrench, A 38
Figueras, N xiii, 8, 43, 93
Filipović, L 22, 27, 28, 55
Flux, T 37
Fox, C M 17
Frawley, W 6
Frederiksen, N 15, 212, 213
Fulcher, G 93

G
Galaczi, A 139
Galaczi, E D xv, 1, 38–40, 89, 157
Gass, S M 214

Geach, J 111
Geranpayeh, A xiv, xviii, 2, 9, 55, 203
Glaser, R 213
Goldstein, H 12–14
Goullier, F 30
Green, A 38, 56, 96

H
Hambleton, R K 17
Hamilton, M 211
Hammersley, M 199
Hawkey, R 1, 39, 40, 89
Hawkins, E W 210
Hawkins, J A 22, 27, 28, 55
Henning, G 215
Hubbard, C 38
Hudson, T 111
Huhta, A 8
Huizhong, Y 15
Hulstijn, J H 94

I
International Association for the Evaluation of Educational Achievement 198

J
Jones, N xiii, xvii, 8, 14, 16, 21, 26, 28, 40, 41, 43, 44, 53, 56, 57, 62, 65–67, 75, 80, 82, 87, 89, 90–92, 106, 110, 117, 122, 138, 139, 141, 142, 150, 156, 182, 200, 202, 204, 211

K
Kane, M T 9, 49
Keenan, E L 21, 22
Keenan, B122
Kendall, J M 94
Kenyon, D 157
Khalifa, H xiv, xviii, 2, 9, 55, 89, 158, 203
King, L 111, 150–152
Kolen, M J 36
Kuijper, H 8, 93

L
Lacey, C 14
Laming, D 27
Lantolf, J P 6
Larsen-Freeman, D 213
Lawton, D 14
Lazenby Simpson, B 8
Leung, C 141
Lim, G S 40
Linacre, J M 24, 41, 67
Little, D 8, 111
Long, M 214
Long, M H 214
Luoma, S 8

M
Macaro, E 124
Malone, M 157
Maris, G xiii
Martyniuk, W xii, 43, 44
McKenna, S 202
Messick, S 9, 46, 47, 49, 54
Milanovic, M 1, 45, 61, 62, 65, 89, 156
Mislevy, R J 9, 10, 15, 212
Mitchell, R 111, 115
Morrow, K 8
Murray, N L 11

N
National Council on Measurement in Education 46, 47
National Research Council 199
NESTA Futurelab 100
Newton, P E 46, 47, 48
Noijons, J xiii, 43
Nold, G 8, 93
North, B xiii, 5, 6, 30, 40, 56, 81, 88–92, 94, 95, 121, 128
Nuffield Foundation 99–102, 105, 106, 116, 121, 122, 140

O
O'Connor, F 8
O'Malley, J M 124
Organisation for Economic Co-operation and Development 198
Oscarson, M 8
Oxford, Cambridge and RSA Examinations 3, 85, 99, 120, 132, 133–134, 137, 140, 145–150
Oxford, Cambridge and RSA Examinations / Cambridge ESOL 115, 143

P
Palmer, A S 9
Panayides, P 14
Panthier, J 30
Parmenter, L 96
Pellegrino, J 141
Pellegrino, J W 213
Perlmann-Balme, M 163
Pollitt, A 11, 62
Purpura, J 211

Q
Qualifications and Curriculum Authority 110, 117, 119, 135, 139

R
Robinson, C 14
Robinson, M 156, 159
Rogers, H J 17

Royal-Dawson, L 72
Ryan, K 1

S
Sahlberg, P 199, 211, 212
Sajavaara, K 8
Saville, N xviii, 1, 13, 16, 26, 28, 39, 77–79, 90, 95, 98, 203, 211
Sayer, A 2, 13
Schleicher, A 199
Shaw, S xiv, xviii, 2, 9, 40, 55, 158, 203
Shohamy, E 93, 94
Smith, P C 94
Sneddon, R 112
Spolsky, B 31, 59
Steinberg, L S 10
Stone, M H 17
Sturman, L 198
SurveyLang 154–156, 158, 159, 161, 164, 167, 168, 170–173, 176, 178, 185, 199
Swain, M 214
Swaminathan, H 17
Szpotowicz, M 202

T
Takala, S 8, 93
Tardieu, C 8, 93
Tattersall, K 141
Taylor, L xii–xiii, xiv, xviii, 2, 9, 55, 87, 89, 203
Teasdale, A 8
The British Academy 149
The Languages Company 150, 152
Thighe, D 67
Thorndike, E L 33
Thurstone, L L 41
Trim, J L M 88, 89, 90, 95, 96, 97
Tymms, P 14

U
UCLES 1, 61, 63, 65, 69, 84

V
Van Ek, J A 88, 89
Van Moere, A 157
Vidaković, I 1

W
Walker, T 44, 139
Weir, C J xiv, xv, xvi, xviii, 1, 2, 7, 8, 9, 15, 40, 46, 55, 62, 69, 89, 93, 95, 111, 112, 126, 157, 158, 160, 203
WIDA 207
Widdowson, H G 55, 214
Wiliam, D 109
Wilkins, D A 88, 89, 93
Winsteps 70, 86
Wolfe, E W 25
Wood, R 15
Wright, B D 17

Y
Yan, J 15

Subject index

A
Ability 9, 14, 15, 16, 17, 18, 19, 24, 25, 29, 30, 36, 38, 44, 67, 73, 77, 111, 122, 177, 179, 185, 188, 189,
  Communicative language ability 6, 8, 31, 32, 41, 53, 90, 92, 113
Accountability 141, 149, 151, 212
Accuracy 24, 33, 35, 50, 55, 157, 217, 218
Algorithms 62, 63
Alignment
  across languages 5, 27, 40, 41, 42, 43, 44, 45, 49, 57, 59, 77, 178,
  across skills 3, 27, 28, 29, 36, 59, 60, 181, 183–186,
  across contexts and levels 30, 31, 35, 37, 38, 39, 40, 44, 45, 49, 50, 51, 52, 53, 59
Alliance Francaise 64, 65
ALTE 2, 3, 45, 62, 64, 65, 77, 78, 81, 82, 83, 89, 154, 155, 158, 176
  Manual for language test development 45
  Can Do project 27, 61, 62, 76, 79, 81, 89, 93, 175, 176, 203, 220
  Framework 3, 6, 10, 65, 75, 76, 77, 78, 79, 176
American Council on the Teaching of Foreign Languages (ACTFL) 6
Anchor tests 26, 37, 71, 127
Anchoring 18, 36, 37, 63, 66, 71, 73, 85, 86, 121, 128, 175
Audio tracks 166
Authoring tool 165, 166
Automated Language Teaching and Assessment (ALTA) 214

B
Behavioural scaling 94
Benchmarking 33, 41, 44, 50–52, 89, 171
Bias 38, 45, 81, 157
BICS (Basic Interpersonal Communication Skills) 54, 207
BIGSTEPS 70
Body of work *see* Standard setting
Bookmark method 176, 178
Borderline 20, 21, 180
Borderline learner 20
BULATS – the Business Language Testing Service 65, 66, 67, 68, 76, 178
Business 28, 64–66, 201, 212
Business English Certificate (BEC) 39, 85

C
Calibration 18, 26, 30, 36, 52, 63, 64, 65, 70, 79, 85, 86, 94, 128, 129, 130, 187
CALP (Cognitive Academic Language Proficiency) 54, 207
Cambridge English 1, 2, 3, 4, 5, 7, 9, 11, 12, 14, 17, 19, 24, 25, 26, 27, 29, 30, 35, 37, 38, 39, 40, 45, 57, 59, 60, 61, 62, 69, 71, 75, 76, 87, 89, 96, 99, 113, 116, 119, 125, 134, 140, 150, 154, 155, 156, 170, 175, 176, 192, 199, 203, 204, 209, 211, 213, 214, 216, 223, 224
  common scale 35, 38, 58, 61, 62, 69, 70, 71, 74, 75, 76, 89, 90, 91, 92, 127, 128, 138, 139, 175, 223, 224
Cambridge Proficiency 69
Cambridge-TOEFL comparability study 1, 203, 205
Candidates Taking More Than One Exam 71
CEFR
  Levels 9, 10, 27, 29, 30, 34, 41, 43, 44, 51, 62, 71, 72, 78, 80, 89, 90, 94, 116, 117, 119, 133, 150, 164, 176, 181, 190, 191, 192, 233, 235
  action-oriented model 9, 10, 59, 90, 156, 157, 159, 161
  descriptor scales 9, 30, 54, 76, 77, 81, 82, 90, 93, 94, 185, 204, 207
  policing 10
  survey on its use 93–96
CELS suite of modular exams 71
Certificate of Advanced English (CAE) 63, 71, 72, 73, 74, 82, 84, 89, 119
Certificate of Proficiency in English (CPE) 1, 63, 71, 72, 74, 82, 84, 89, 91, 119, 206, 223, 224
CLIL (Content and Language Integrated Learning) 53, 54, 59, 196, 207
Cloze passage *see* Task types
Coherence 33, 55, 56, 88, 94, 217, 218, 224, 234
Commissioning 160, 162, 163
Common European Framework of Reference *see* CEFR

Common Scale 35–41, 58, 61–63, 69–76, 89–93, 127–130, 138–140, 175–176, 223–225
CommuniCAT 65, 66
Community languages 30, 101, 106, 108, 109, 111, 116, 122, 124, 134
Comparability
  between teacher and external assessment 142, 143
Comparative judgement 27, 40, 41, 46, 69, 166
  marking writing 166
Computer-adaptive testing 62, 63, 64, 65, 205
Consequential validity *see* Validity
Construct definition 3, 4, 5, 11, 14, 28, 31, 32, 59
Construct validity *see* validity
Construct
  coverage 162, 165
Constructed-response *see* Task types
Contingency table 71
Correlation 66, 67, 68, 83, 174
Criterion-related validity *see* Validity
Cross-validation 51
Customisation of questionnaires 168, 1669, 170, 175, 193

D
Decision rule 51
Demographic information 168, 169, 203
Department for Education and Skills (DfES) 100, 103, 104, 105, 114, 115, 118, 135, 137, 139, 140, 144
Design effect 174
Domains of language use 9, 80, 124, 159, 160
Dutch 56, 64, 65, 79

E
Effective sample size 174
English Profile Programme 22, 28, 55–59, 75, 214
European Academic Software Award 65
European Language Portfolio 10
European Survey on Language Competences (ESLC) 2, 4, 10, 14, 28, 29, 31, 33, 42, 44, 49, 57, 69, 97, 149, 150, 151, 152, 154–202, 208, 209, 211, 216, 233
European Year of Languages 77
Examination levels 5, 39, 89
Exemplar performances 131

F
FACETS 24, 25, 67
Facility 16, 17, 21, 38, 163
Falsifiability 13
Familiarisation material 161
Familiarity 35, 115, 157, 161
First Certificate in English (FCE) 63, 71, 72, 73, 74, 82, 84, 89, 119
Fluency 33, 55, 56, 81, 82, 83, 217, 218
Formative 15, 16, 99, 106, 109, 111, 112, 125, 132, 138, 140, 142, 144, 151, 204, 211
Formative assessment 16, 105, 113, 141, 142, 148, 151, 211
Frameworks 1, 2, 3, 4, 5, 6, 10, 28, 31, 34, 45, 49, 54, 76, 77, 78, 98, 102, 103, 106, 109, 117, 119, 139, 154, 168, 203, 204, 206, 208, 215, 216
French 4, 42, 63, 64, 65, 66, 68, 79, 81, 88, 108, 109, 120, 130, 134, 136, 137, 139, 143, 148, 150, 157, 192, 193, 201, 232, 240, 241

G
Gallup Europe 154, 155, 159, 165
General notions 58
German 4, 42, 64, 65, 66, 68, 79, 108, 109, 130, 134, 135, 136, 137, 139, 140, 148, 150, 188, 189, 192, 193, 200, 201, 232, 242, 243
Goethe Institute 64, 65, 154
Graded objectives 102, 112, 142, 146, 151
Guessing 66, 67
  correction for 67

H
High-stakes assessment 58, 110, 206, 212

I
IELTS 62, 76, 85, 205, 206
Impact studies 203, 208, 209, 210
INES (International Indicators of Education Systems) 169, 170
Interaction 14, 19, 33, 55, 56, 137, 213, 214, 215, 217, 218
Interactional authenticity 19, 55, 138, 214
Internal consistency 14, 15
Interpretation 6, 11, 13, 15, 16, 18, 28, 29, 38, 46, 47, 48, 59, 118, 188, 190, 203, 215
ISCED levels 155, 170, 171
Italian 4, 42, 65, 79, 134, 148, 157, 165, 178, 180, 183, 187, 188, 192, 232, 236, 237, 238, 243, 244
item banking 3, 11, 15, 25, 26, 27, 35, 61, 62, 65, 70, 74, 89, 99, 113, 125, 126., 132, 136, 165, 175, 214
item writer guidelines 76, 161
item writing 59, 134, 162, 203
ITEMAN 70

J
Japanese 120, 122, 131, 138, 139, 140, 148, 215, 232
Judgement
  Comparative 27, 40, 41, 46, 69, 166
  in standard setting 46, 57, 62, 183, 191
  multilingual 178, 183
  of performance skills 24, 46, 74,

K
Kanji 131
Key English Test (KET) 37, 71, 73, 82, 84, 89, 91, 119, 224, 225
KoBaLT 65

L
Language education 2, 3, 11, 32, 87, 95, 96, 98, 99, 101, 140, 150, 155, 192, 196, 202, 204, 208, 209, 210, 216
Language functions 157, 160, 227, 230, 231
Language policy 77, 95, 154–202, 204, 207, 209
Language Policy Division 45, 88
Languages Ladder 105, 107, 108, 109, 113–123, 125–127, 131, 133, 138, 139, 142, 143, 144, 146, 148, 151, 226
Latent trait 11, 17, 215
Learner characteristics 112
Learning curve 73, 74
Learning gains 74, 128
Learning ladder 34, 69, 89, 90, 92, 99, 102, 120, 121, 122, 131, 140
Learning Oriented Assessment (LOA) 11, 16, 75, 154, 203, 211, 212, 213, 214, 215, 216
Leniency 24, 133
LIBS (Local Item Banking System) 25, 74, 85, 86
Lingua Programme 77, 78
Linguaskill 63–65
Listening 9, 21, 29–30, 35, 38, 42, 55, 58, 69, 78–79, 95, 102–107, 111–113, 115, 121, 126–131, 133, 144, 155, 158, 173, 176–183, 186–190, 207, 220, 221, 226
Localisation 169–170
Logistic regression 182, 183
Logits 18, 22, 68, 82, 189
Low-stakes assessment 62, 65, 67, 142

M
Manpower 63–65
Mastery 6, 20–22, 23, 33, 83–84, 105–107, 119, 133–136, 179–180, 223–224, 232
  Minimum 21
Matching *see* Task types
Measurement
  optimum level of difficulty 22, 29
  trait-based 12–16, 213–214
Mixed-methods 139–140
Monitoring of Exam Difficulty 71
Motivation 48, 97, 101–104, 110–112, 136, 139, 142–143, 148, 201, 213
Multimedia 64
Multiple-choice *see* Task types

N
National Curriculum 6, 115–118, 122, 126, 130, 139–140, 146, 152
National Language Standards 115
National Languages Strategy (2002) 99–100, 103–107, 110, 113, 121, 140, 146, 149–151
National Qualifications Framework (NQF) 106, 117–119
Native speaker 34, 104, 110, 111, 119–120, 152, 207–208, 222
Natural levels 57, 89, 91–92
Needs analysis 31, 59, 93, 97
Normative information 74, 181

O
Objectively marked tests 12, 14, 22, 29–30, 35–38, 42, 46, 58, 86, 91, 121, 125, 216, 130–131, 139. 178
Optical mark recognition 125
Oxford, Cambridge and RSA Examinations (OCR) 3, 85, 99, 120, 132, 133, 137, 140, 145–150

P
Pass mark 16–17, 72, 144
Pass rate 16–17, 74
Performance level 9, 19–23, 59, 69, 165, 167
Performance skills 10, 12, 19–21, 24–25, 29–30, 35, 41, 46, 58, 74, 130, 176, 183
Performance, expected 29, 131–132
Plurilingual 49, 52, 67, 68, 69, 98, 158, 178, 210
Plurilingualism 208
Portfolio 10, 88, 143, 198
Practical adequacy 13
Pragmatic competence 160
Preliminary English Test (PET) 71, 73, 82, 84, 89, 119
Pretesting 26, 37, 578, 61–62, 70–74, 125–130, 135, 161–162
Probability 17–25, 83–84, 172
Probability proportional to size (PPS) 172
Proficiency tests 5–11, 30, 50, 63–64, 98, 108, 113–114
  compared to achievement tests 6, 108, 114
Proficiency
  versus learning stage 115–118, 133
Profiled reporting 30

Progression
context-specific 94
in topic range 56, 59
linguistic 58–59
Purposeful language use 52–55

Q
Quality of data 173–174
Questionnaires 27, 64, 80–81, 155–156, 158, 168–170, 173, 175, 178, 185–186, 190–196, 201
Quick Placement Test 65

R
Range 33, 55–56, 217–219
Rasch model 11–19, 59, 70, 175–176
criticised 12–14
estimation 66
goodness of fit 18
multi-faceted 24–25, 191–192
Ratings 11, 41–42, 55, 64, 67, 76, 80–83, 126–127, 139, 178, 181, 188–190
supervisor 27, 64–65
Rating scales 5–6, 24–28, 38–40, 64–65, 93–94, 126–128, 131–132
Reading 9, 14–15, 20–21, 30, 38, 42, 55, 58, 69, 78–79, 93–95, 102–107, 111–115, 120–121, 126–130, 132, 139–140, 155, 158, 173, 176, 178, 181–183, 186–189, 207, 221
Receptive skills 58, 81, 104, 123, 125, 176
Regression 83, 84, 182, 183, 193, 195, 196
Reliability 14–15, 30, 51, 58, 64, 66, 113, 138–141, 170, 171, 203, 216
test-retest 66
Response probability 20–21
Rote learning 113, 125, 146, 151
Routing test 164, 172

S
Salient features 15, 27, 53, 55–56, 95, 122, 125, 230
Sampling 155, 170–175
Scaffolding 21–24, 69, 112, 214
Scoring validity *see* Validity
Script acquisition 120, 140
Second language acquisition 6, 93, 115
Security 26, 85, 136, 144, 216
Self-evaluation 185–186, 214
Severity 24–25, 133, 167, 191
Singapore Telecom 62–63
Skills constructs *see* Reading, Writing, Listening, Speaking, Use of English
Socio-cognitive model 2, 6–10, 19, 31, 54–55, 62, 95, 157, 159–160, 213, 214
Spanish 4, 42, 64–68, 79, 130, 134, 136, 137, 148, 192, 195, 200, 201, 232, 245–246
Speaking 9, 14, 20–21, 29–30, 33–35, 38–44, 55, 58, 74, 78–79, 95, 102–108, 112–115, 126, 130–133, 139, 157, 186–188, 190, 207, 220, 224
not tested in ESLC 157, 186
Stages of attainment 6, 116–118
Standard setting
Body of Work 180–181, 183
learner-centred 43–44, 58, 176
task-centred 42–44, 56, 176, 178
premises of 43
task or examinee-centred 56
using teacher judgements 27, 91, 177
Standardisation
cross-language 131–135
Standards
differing across countries 167–168
of attainment 69, 115–118, 141–143
Stratification 171–174
Subjectivity 24, 27, 38–40, 46, 55, 58, 130–131
Successful language learner 201
Summative 4, 16, 58, 106, 109, 212, 215, 216
Summative assessment 15–16, 40, 112–113, 140–142, 204, 216
SurveyLang 154–161, 164, 167, 168, 170–173, 176, 178, 185, 199

T
Targeted testing 163–164
Task, difficulty 11, 16–27, 29–31, 35–36, 57, 59, 69, 71–74, 83, 131, 176–177
Task order effect 165
Task types
cloze passages 63, 163
constructed-response 63
matching 37–38, 160–161
multiple-choice 21, 43, 55, 63, 160–161, 163, 233–234
selected response 160–161
Teacher assessment 57, 105–106, 112, 135–137, 140–148
The National Institute for Educational Measurement (Cito) 154–156, 168, 176, 178
The Nuffield Foundation 100, 102
Theta scale 17
Threshold 31, 87, 89, 223, 225
Training 10, 24, 50–51, 61, 74, 75, 85, 86, 108, 116, 131–135, 145–147, 151, 155, 167169, 177, 191, 196–198, 204

U
UK National Curriculum 6, 115–118, 122, 126, 130, 139, 140, 146, 152
Unidimensionality 15, 37
Università per Stranieri di Perugia 65, 154

University of Louvain 64
Urdu 108, 112, 139–140, 148, 232
US Foreign Service Institute (FSI) 5–6
USB sticks 165
Use of English 12, 29

V
Validation
  external 50–52, 184
  of standard setting 50, 184
Validity
  consequential 8, 9, 113
  construct 8–10, 47–48, 113, 203
  criterion-related 8, 9, 113, 131, 209
  scoring 1, 8, 9, 62, 126, 203
Vetting 162–163
Vocabulary 40, 113–115, 123, 124, 137, 138, 141, 146, 152, 218, 222–224, 229, 233

W
WINSTEPS 70, 86
Writing 9, 10, 14, 20–23, 29–30, 35, 38–42, 55, 58, 74, 78–79, 95, 102–107, 111–113, 115, 126, 130–133, 139, 152, 155, 173, 176–188, 195–196, 208, 222, 223

Y
Young learners 37, 53–54, 91, 207