



COUNCIL OF EUROPE    CONSEIL DE L'EUROPE

Language Policy Division  
Division des Politiques linguistiques

January 2009

## **Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)**

### **A Manual**

Language Policy Division, Strasbourg  
[www.coe.int/lang](http://www.coe.int/lang)

# Contents

<b>List of Figures</b>	Page iii
<b>List of Tables</b>	Page v
<b>List of Forms</b>	Page vii
<b>Preface</b>	Page ix
<b>Chapter 1:</b> The CEFR and the Manual	Page 1
<b>Chapter 2:</b> The Linking Process	Page 7
<b>Chapter 3:</b> Familiarisation	Page 17
<b>Chapter 4:</b> Specification	Page 25
<b>Chapter 5:</b> Standardisation Training and Benchmarking	Page 35
<b>Chapter 6:</b> Standard Setting Procedures	Page 57
<b>Chapter 7:</b> Validation	Page 89
<b>References</b>	Page 119
<b>Appendix A</b> Forms and Scales for Description and Specification (Ch. 1 & 4)	Page 122
A1: Salient Characteristics of CEFR Levels (Ch. 1)	Page 123
A2: Forms for Describing the Examination (Ch. 4)	Page 126
A3: Specification: Communicative Language Activities (Ch. 4)	Page 132
A4: Specification: Communicative Language Competence (Ch. 4)	Page 142
A5: Specification: Outcome of the Analysis (Ch. 4)	Page 152
<b>Appendix B</b> Content Analysis Grids (Ch.4)	
B1: CEFR Content Analysis Grid for Listening & Reading	Page 153
B2: CEFR Content Analysis Grids for Writing and Speaking Tasks	Page 159
<b>Appendix C</b> Forms and Scales for Standardisation & Benchmarking (Ch. 5)	Page 181
<b>Reference Supplement:</b>	
<b>Section A:</b> Summary of the Linking Process	
<b>Section B:</b> Standard Setting	
<b>Section C:</b> Classical Test Theory	
<b>Section D:</b> Qualitative Analysis Methods	
<b>Section E:</b> Generalisability Theory	
<b>Section F:</b> Factor Analysis	
<b>Section G:</b> Item Response Theory	
<b>Section H:</b> Test Equating	

## List of Figures

Figure 2.1:	Validity Evidence of Linkage of Examination/Test Results to the CEFR	Page 8
Figure 2.2:	Visual Representation of Procedures to Relate Examinations to the CEFR	Page 15
Figure 6.1:	Frequency Distributions of Test Scores in Two Contrasting Groups	Page 67
Figure 6.2:	Logistic Regression	Page 73
Figure 6.3:	Panel Member Recording Form for Bookmark Method	Page 78
Figure 6.4:	Items with Unequal Discrimination	Page 82
Figure 6.5:	Item Map, Indicating Difficulty and Discrimination	Page 83
Figure 7.1:	Empirical Item Characteristic Curve for a Problematic Item	Page 101
Figure 7.2:	A Test Characteristic Curve	Page 105
Figure 7.3:	Bivariate Decision Table Using Nine Levels	Page 112
Figure 7.4:	Bivariate Decision Table Using Five Levels	Page 113
Figure 7.5:	Item Map with Test Items and “Can Do” Statements	Page 116

## List of Tables

Table 3.1:	Time Management for Familiarisation Activities	Page 23
Table 3.2:	Documents to be Prepared for Familiarisation Activities	Page 23
Table 4.1:	Forms and Scales for Communicative Language Activities	Page 32
Table 4.2:	CEFR Scales for Aspects of Communicative Language Competence	Page 32
Table 5.1:	Time Management for Assessing Oral Performance Samples	Page 45
Table 5.2:	Time Management for Assessing Written Performance Samples	Page 46
Table 5.3:	Documents and Tools to be prepared for Rating Writing	Page 47
Table 5.4:	Reference Sources in the CEFR	Page 49
Table 5.5:	Standardisation and Benchmarking: Summary	Page 55
Table 6.1:	Overview of the Methods Discussed	Page 61
Table 6.2:	Basic Data in the Tucker-Angoff method	Page 62
Table 6.3:	Computing the Expected Score of 100 Borderline Persons	Page 66
Table 6.4:	Frequency Distribution Corresponding to Figure 6.1	Page 68
Table 6.5:	Decision Tables for Five Cut-off Scores	Page 68
Table 6.6:	Summary of the Rangefinding Round	Page 71
Table 6.7:	Results of the Pinpointing Round (partially)	Page 73
Table 6.8:	Example of an ID Matching Response Form (abridged)	Page 75
Table 6.9:	Bookmarks and Achievement Levels	Page 80
Table 6.10:	Estimated Theta	Page 81
Table 7.1:	Balanced Incomplete Block Design with Three Blocks	Page 93
Table 7.2:	Balanced Incomplete Block Design with Seven Blocks	Page 93
Table 7.3:	Example of High Consistency and Total Disagreement	Page 98
Table 7.4:	Bivariate Frequency Table using Four Levels	Page 99
Table 7.5:	Frequencies of Allocation of a Single Item to Different CEFR Levels	Page 101
Table 7.6:	Summary of Disagreement per Item	Page 102
Table 7.7:	Outcome of a Tucker-Angoff Procedure	Page 102

Table 7.8:	Variance Decomposition	Page 103
Table 7.9:	Decision Accuracy	Page 107
Table 7.10:	Decision Consistency	Page 108
Table 7.11:	Marginal Distributions Across Levels (Frequencies)	Page 110
Table 7.12:	Marginal Distributions Across Levels (Percentages)	Page 111
Table 7.13:	Design for a Paired Standard Setting	Page 115
Table A1:	Salient Characteristics: Interaction & Production	Page 123
Table A2:	Salient Characteristics: Reception	Page 124
Table A3:	Relevant Qualitative Factors for Reception	Page 143
Table A4:	Relevant Qualitative Factors for Spoken Interaction	Page 148
Table A5:	Relevant Qualitative Factors for Production	Page 149
Table C1:	Global Oral Assessment Scale	Page 184
Table C2:	Oral Assessment Criteria Grid	Page 185
Table C3:	Supplementary Criteria Grid: “Plus levels”	Page 186
Table C4:	Written Assessment Criteria Grid	Page 187

## List of Forms

Form A1:	General Examination Description	Page 126
Form A2:	Test Development	Page 127
Form A3:	Marking	Page 129
Form A4:	Grading	Page 130
Form A5:	Reporting Results	Page 130
Form A6:	Data Analysis	Page 131
Form A7:	Rationale for Decisions	Page 131
Form A8:	Initial Estimation of Overall Examination Level	Page 28 / 132
Form A9:	Listening Comprehension	Page 132
Form A10:	Reading Comprehension	Page 133
Form A11:	Spoken Interaction	Page 134
Form A12:	Written Interaction	Page 136
Form A13:	Spoken Production	Page 137
Form A14:	Written Production	Page 138
Form A15:	Integrated Skills Combinations	Page 139
Form A16:	Integrated Skills	Page 139
Form A17:	Spoken Mediation	Page 140
Form A18:	Written Mediation	Page 141
Form A19:	Aspects of Language Competence in Reception	Page 142
Form A20:	Aspects of Language Competence in Interaction	Page 145
Form A21:	Aspects of Language Competence in Production	Page 146
Form A22:	Aspects of Language Competence in Mediation	Page 150
Form A23:	Graphic Profile of the Relationship of the Examination to CEFR Levels	Page 33 / 152
Form A24:	Confirmed Estimation of Overall Examination Level	Page 34 / 152
Form C1:	Training Record Form	Page 181
Form C2:	Analytic Rating Form (Swiss Project)	Page 182
Form C3:	Holistic Rating Form (DIALANG)	Page 182
Form C4:	Collation Global Rating Form (DIALANG)	Page 183
Form C5:	Item Rating Form (DIALANG)	Page 183

These forms are also available on the website [www.coe.int/lang](http://www.coe.int/lang)

## Preface

The Council of Europe wishes to acknowledge with gratitude all those who have made it possible to develop this Manual, and in particular the contributions by:

- The Finnish authorities who provided the forum in Helsinki to launch the initiative in July 2002.
- The “Sounding Board” of consultants for the pilot edition (Prof. Charles Alderson, Dr Gergely A. David, Dr John de Jong, Dr Felianka Kaftandjieva, Dr Michael Makosch, Dr Michael Milanovic, Professor Günther Nold, Professor Mats Oscarson, Prof. Günther Schneider, Dr Claude Springer and also Mr Josef Biro, Ms Erna van Hest, Mr Peter Lenz, Ms Jana Pernicová, Dr Vladimir Kondrat Shleg, Ms Christine Tagliante and Dr John Trim) for their important feedback in the early stage of the project.
- The Authoring Group, under the leadership of Dr Brian North:
  - Dr Neus Figueras                      – Departament d’Educació, Generalitat de Catalunya, Spain
  - Dr Brian North                        – Eurocentres Foundation, Switzerland
  - Professor Sauli Takala                – University of Jyväskylä, Finland (emeritus)
  - Dr Piet van Avermaet                 – Centre for Diversity and Learning, University of Ghent, Belgium
  - Association of Language Testers in Europe (ALTE)
  - Dr Norman Verhelst                  – Cito, The Netherlands
- Dr Jay Banerjee (University of Lancaster) and Dr Felianka Kaftandjieva (University of Sofia) for their contributions to the Reference Supplement to the Manual.
- The institutions who made available illustrative performance samples and sample test items that have been circulated on DVD/CD ROM and made available on the Council of Europe’s website in order to assist in standardisation training (especially: Eurocentres; Cambridge ESOL; the CIEP; the University for Foreigners, Perugia; the Goethe-Institut; the Finnish authorities; DIALANG; the Generalitat de Catalunya and CAPLE).
- ALTE (especially Nick Saville) and the members of the “Dutch CEFR project group” (Charles Alderson, Neus Figueras, Günther Nold, Henk Kuijper, Sauli Takala, Claire Tardieu) for contributing to the “Toolkit” related to this Manual with the Content Analysis Grids which they developed for Speaking and Writing, and for Listening and Reading respectively.
- The many individuals and institutions who gave detailed feedback on the pilot version, especially: the members of ALTE; Asset Languages (Cambridge ESOL); Budapest Business School; Cito; Claudia Harsch; the Goethe-Institut; the Polish Ministry of Education; the Taiwan Ministry of Education; TestDaF; Trinity College London; and the University for Foreigners, Perugia.

*Language Policy Division*

Directorate of Education and Languages (DG IV)

F – 67075 STRASBOURG Cedex

[www.coe.int/lang](http://www.coe.int/lang)

[www.coe.int/portfolio](http://www.coe.int/portfolio)

# Chapter 1

## The CEFR and the Manual

### 1.1. The Aims of the Manual

### 1.2. The Context of the Manual

### 1.1. The Aims of the Manual

The primary aim of this Manual is to help the providers of examinations to develop, apply and report transparent, practical procedures in a cumulative process of continuing improvement in order to situate their examination(s) in relation to the Common European Framework (CEFR). The Manual is not the sole guide to linking a test to the CEFR and there is no compulsion on any institution to undertake such linking. However, institutions wishing to make claims about the relationship of their examinations to the levels of the CEFR may find the procedures helpful to demonstrate the validity of those claims.

The approach developed in the Manual offers guidance to users to:

- describe the examination coverage, administration and analysis procedures;
- relate results reported from the examination to the CEFR Common Reference Levels;
- provide supporting evidence that reports the procedures followed to do so.

Following the traditions of Council of Europe action in promoting language education, however, the Manual has wider aims to actively promote and facilitate cooperation among relevant institutions and experts in member countries. The Manual aims to:

- contribute to competence building in the area of linking assessments to the CEFR;
- encourage increased transparency on the part of examination providers;
- encourage the development of both formal and informal national and international networks of institutions and experts.

The Council of Europe's Language Policy Division recommends that examination providers who use the suggested procedures, or other procedures achieving the same ends, write up their experience in a report. Such reports should describe the use of procedures, discuss successes and difficulties and provide evidence for the claims being made for the examination. Users are encouraged to write these reports in order to:

- increase the transparency of the content of examinations (theoretical rationale, aims of examination, etc.);
- increase the transparency of the intended level of examinations;
- give test takers, test users and teaching and testing professionals the opportunity to analyse the quality of an examination and of the claimed relation with the CEFR;
- provide an argumentation why some of the recommended procedures may not have been followed;
- provide future researchers with a wider range of techniques to supplement those outlined in this Manual.

It is important to note that while the Manual covers a broad range of activities, its aim is limited:

- It provides a guide specifically focused on procedures involved in the justification of a claim that a certain examination or test is linked to the CEFR.

- It does not provide a general guide how to construct good language tests or examinations. There are several useful guides that do this, as mentioned in Chapter 4, and they should be consulted.
- It does not prescribe any single approach to constructing language tests. While the CEFR espouses an action-oriented approach to language learning, being comprehensive, it accepts that different examinations reflect various goals (“constructs”).
- It does not require the test(s) to be specifically designed to assess proficiency in relation to the CEFR, though clearly exploitation of the CEFR during the process of training, task design, item writing and rating scale development strengthens the content-related claim to linkage.
- It does not provide a label, statement of validity or accreditation that any examination is linked to the CEFR. Any such claims and statements are the responsibility of the institution making them. There are professional associations concerned with standards and codes of practice (e.g. the AERA: American Educational Research Association (AERA/APA/NCME 1999); EALTA [www.ealta.org](http://www.ealta.org); ALTE [www.ALTE.org](http://www.ALTE.org)) which are a source of further support and advice on language testing and linking procedures.

Despite the above, the pilot Manual has in fact been consulted by examination authorities in many different ways:

- to apply to an existing test that predates the CEFR and therefore without any clear link to it, in order to be able to report scores on the test in relation to CEFR levels;
- to corroborate the relationship of an existing test that predates the CEFR to the construct represented by the CEFR and to the levels of the CEFR; this applies to tests developed in relation to the series of content specifications developed by the Council of Europe since the 1970s now associated with CEFR levels: Breakthrough:A1, Waystage: A2, Threshold: B1, Vantage: B2 (van Ek and Trim 2001a–c);
- to corroborate the relationship to the CEFR of an existing test developed after the appearance of the CEFR but preceding the appearance of the Manual itself; this applies to some tests produced between 1997 and 2004;
- to inform the revision of an existing examination in order to relate it more closely to the CEFR construct and levels;
- to assist schools to develop procedures to relate their assessments to the CEFR.

The Manual was not conceived as a tool for linking existing frameworks or scales to the CEFR, but the sets of procedures proposed may be useful in doing so. For an existing framework, the relationship could be mapped from the point of view of content and coverage using the Specification stage. Performance samples benchmarked to the framework under study could be used in a cross-benchmarking exercise after Standardisation training: CEFR illustrative samples could be rated with the criteria used in the framework under study and benchmark samples from the framework under study could be rated with the CEFR criteria for spoken and written performance provided in this Manual. Finally, tests from the framework under study could be investigated in an External Validation study.

In order to help users assess the relevance and the implications of the procedures for their own context, “Reflection Boxes” that summarise some of the main points and issues are included at the end of each chapter (*Users of the Manual may wish to consider ...*), after the model used in the CEFR itself.

## 1.2. The Context of the Manual

The Common European Framework of Reference for Languages has a very broad aim to provide:

“... a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. The description also covers the cultural context in which language is set. The Framework also defines levels of proficiency which allow learners’ progress to be measured at each stage of learning and on a life-long basis” (Council of Europe 2001a: 1).

But the CEFR is also specifically concerned with testing and examinations, and it is here that the Manual is intended to provide support:

“One of the aims of the Framework is to help partners to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications. For this purpose the Descriptive Scheme and the Common Reference Levels have been developed. Between them they provide a conceptual grid which users can exploit to describe their system” (Council of Europe 2001a: 21).

The aim of the CEFR is to facilitate reflection, communication and networking in language education. The aim of any local strategy ought to be to meet needs in context. The key to linking the two into a coherent system is flexibility. The CEFR is a concertina-like reference tool that provides categories, levels and descriptors that educational professionals can merge or sub-divide, elaborate or summarise – whilst still relating to the common hierarchical structure. CEFR users are encouraged to adopt language activities, competences and proficiency stepping-stones that are appropriate to their local context, yet can be related to the greater scheme of things and thus communicated more easily to colleagues in other educational institutions and to other stakeholders like learners, parents and employers.

Thus 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.

The CEFR is already serving this function flexibly in its implementation through the European Language Portfolio. The portfolio is a new educational tool and it has been developed through intensive and extensive international cooperation. Thus the conditions for its implementation in a sufficiently uniform manner are relatively good, even if there have been and are a variety of constraints impacting the portfolio project.

By contrast the mutual recognition of language qualifications awarded by all relevant bodies is a much more complicated matter. The language assessment profession in Europe has very different traditions. At the one extreme there are examination providers who operate in the classical tradition of yearly examinations set by a board of experts and marked in relation to an intuitive understanding of the required standard. There are many contexts in which the examination or test leading to a significant qualification is set by the teacher or school staff rather than an external body, usually but not always under the supervision of a visiting expert. Then again there are many examinations that focus on the operationalisation of task specifications, with written criteria, marking schemes and examiner training to aid consistency, sometimes including and sometimes excluding some form of pretesting or empirical validation. Finally, at the other extreme, there are highly centralised examination systems in which primarily selected-response items measuring receptive skills drawn from item banks, sometimes supplemented by a productive (usually written) task, are used to determine competence and award qualifications. National policies, traditions and evaluation cultures as well as the policies, cultures and legitimate interests of language testing and examination bodies are factors that can constrain the common interest of mutual recognition of qualifications. However it is in everybody’s best interests that good practices are applied in testing.

Apart from the question of tradition, there is the question of competence and resources. Well-established institutions have, or can be expected to have, both the material and human resources to be able to develop and apply procedures reflecting best practice and to have proper training, quality assurance and control systems. In some contexts there is less experience and a less-informed assessment culture. There may be only limited familiarity with the networking and assessor-training techniques associated with standards-oriented educational assessment, which are a prerequisite for consistent performance assessment. On the

other hand there may be only limited familiarity with the qualitative and psychometric approaches that are a pre-requisite for adequate test validation. Above all there may be only limited familiarity with techniques for linking assessments, since most assessment communities are accustomed to working in isolation.

Therefore it is not surprising that following the publication of the CEFR, there were calls for the Council of Europe to take a more active role in assisting examination providers in their efforts to validate the relationship of their examinations to the Common European Framework of Reference. The topic was the theme of a seminar kindly hosted by the Finnish authorities in Helsinki in July 2002 (Council of Europe 2002), at the conclusion of which the Language Policy Division of the Council of Europe in Strasbourg set up the project to develop this Manual.

This Manual is a continuation of the work of the Council of Europe's Language Policy Division in developing planning tools which provide reference points and common objectives as the basis for a coherent and transparent structure for effective teaching/learning and assessment relevant to the needs of learners as well as society, and that can facilitate personal mobility. This work first became widely known in the 1970s with the publication of "The Threshold Level" (van Ek 1976; van Ek and Trim 2001b) and the development of versions of it for different languages. The 1990s saw the research and development for the CEFR, first circulated in two pilot editions before full publication in English and French in 2001, the European Year of Languages, (Council of Europe 2001a, 2001b) and now published in over 30 languages. The initial main impact of the CEFR was the "Common Reference Levels" (A1–C2) that it introduced. The CEFR is now, however, itself inspiring a new generation of sets of objectives for curriculum developers, further elaborated from the CEFR descriptors (see Section 4.3.3.). This current Manual, with its emphasis on relating *assessments* to one another through the mediation of the CEFR, is a logical complement to these developments on *levels* and *objectives*.

There is no suggestion that different examinations that have been linked to the CEFR by following procedures such as those proposed in this Manual could be considered to be in some way equivalent. Examinations vary in content and style, according to the needs of their context and the traditions of the pedagogic culture in which they are developed, so two examinations may both be "at Level B2" and yet differ considerably. Learners in two different contexts might gain very different scores on (a) an examination whose style and content they are familiar with and (b) an examination at the same level developed for a different context. Secondly, the fact that several examinations may be claimed to be "at Level B2" as a result of following procedures to link them to the CEFR, such as those suggested in this Manual, does not mean that those examinations are claimed to be exactly the same level. B2 represents a band of language proficiency that is quite wide; the "pass" cut-off level for the different examinations may be pitched at different points within that range, not all coinciding at exactly the same borderline between B1 and B2.

Both curricula and examinations for language learning need to be developed for and adapted to the context in which they are to be used. The authors of the CEFR make it clear that the CEFR is in no way to be interpreted as a harmonisation project. It is not the intention of the CEFR to tell language professionals what their objectives should be:

"We have NOT set out to tell practitioners what to do or how to do it. We are raising questions not answering them. It is not the function of the CEF to lay down the objectives that users should pursue or the methods they should employ" (Council of Europe 2001a: xi Note to the User).

Neither is it the intention of this Manual to tell language professionals what their standards should be, or how they should prove linkage to them. Both the CEFR and this Manual share the aims of encouraging reflection, facilitating communication (between language professionals and between educational sectors) and providing a reference work listing processes and techniques. Member states and institutions concerned with language teaching and learning operate and cooperate autonomously; it is their privilege and responsibility to choose approaches appropriate to their purpose and context.

A pilot version of this Manual was published in September 2003 (Council of Europe 2003) and presented at a seminar in Strasbourg in April 2004. The appearance of the Manual in September 2003, shortly after the full publication of the CEFR itself in English and French (2001), had a considerable impact. To a great

extent, the scale of the impact of both the CEFR itself and the Manual can be regarded as fortunate timing. At precisely the point at which examination providers were looking for ways to increase the transparency of their examinations and make them more relevant in a European context, the CEFR and Manual appeared to offer a principled way of doing so. As a result, the methodology of many CEFR linking projects was influenced by the approach suggested in the Manual. At the same time those approaches were criticised and further elaborated during the more than 20 case studies of such pilot linking projects that were carried out.

Many of these projects were presented at a meeting in Cambridge in December 2007 and at the EALTA pre-conference colloquium in Athens, 2008. Feedback both from institutions involved in piloting and from a wide range of other interested parties in and beyond Europe has greatly assisted in the preparation of this revised edition, which, whilst naturally not definitive, is more comprehensive. The papers from the Cambridge meeting are being published in a compendium of case studies in the “Studies in Language Testing” series by Cambridge University Press; the papers from the Athens meeting are being published in a compendium of case studies by Cito, in association with the Council of Europe and EALTA. It is hoped that these studies, together with this Manual and the growing range of tools accompanying the CEFR, will contribute to the development of expertise in the linking of language examinations to the CEFR and to the discussion of the issues that arise in this process.

*Users of the Manual may wish to consider:*

- *the relevance of the CEFR in their assessment and testing context*
- *the reasons for and aims of their application of the Manual*
- *the requirements that their specific context sets for the application of the Manual*
- *the parts of the Manual that are likely to be most relevant for them*
- *how they might report their results so as to contribute to the building of expertise in the area of linking*

## Chapter 2

### The Linking Process

#### 2.1. Approach Adopted

#### 2.2. Quality Concerns

#### 2.3. Stages of the Process

#### 2.4. Use of the CEFR

#### 2.5. Use of the Manual

#### 2.1. Approach Adopted

Relating an examination or test to the CEFR is a complex endeavour. The existence of a relationship between the examination and the CEFR is not a simple observable fact, but is an assertion for which the examination provider needs to provide both theoretical and empirical evidence. The procedure by which such evidence is obtained is in fact the “validation of the claim”.

Relating (linking) examinations or tests to the CEFR presupposes standard setting, which can be defined as a process of establishing one or more cut scores on examinations. These cut scores divide the distribution of examinees’ test performances into two or more CEFR levels.

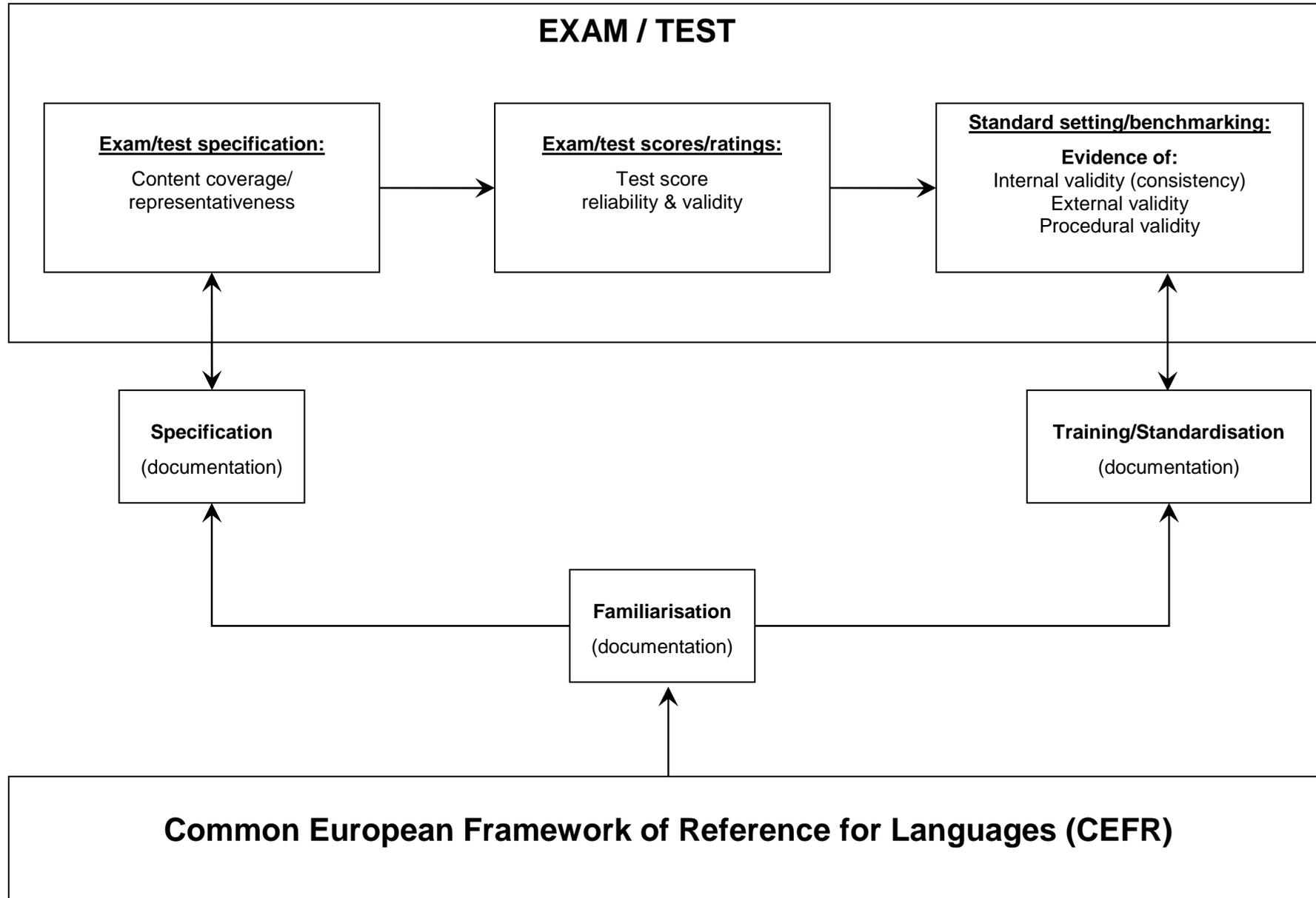
Appropriate standards can be best guaranteed if the due process of standard setting is attended to from the beginning. Standard setting involves decision making which requires high-quality data and rigorous work. As these decisions may have important consequences, they need to be fair, open, valid, efficient and defensible. This can be facilitated by the use of well-tried systematic processes and explicit criteria.

In standard setting, it is usual to refer to content standards and performance standards. Content standards describe the content domain from which the examination can be or has been constructed. Very frequently this description refers to performance levels. Such descriptions are by necessity general and usually formulated in qualitative terms. In standard setting literature they are referred to as “Performance Level Descriptors” (PLDs: See Section 6.7.) and act as a general reference system against which particular examinations can be described. Performance standards refer to specific examinations and express the minimum performance on that specific test or examination; in this sense they are synonymous to “cut scores”.

There is, however, one major point which needs to be stressed. The Common European Framework of Reference for Languages (CEFR) provides the content and Performance Level Descriptors. The PLDs are given, unlike the situation in most standard setting in other contexts, where the PLDs first need to be defined.

This means that the CEFR needs to be referred to at all stages of the linking process as illustrated in Figure 2.1. The approach adopted in this Manual is such that thorough familiarity with the CEFR is a fundamental requirement.

Figure 2.1: Validity Evidence of Linkage of Examination/Test Results to the CEFR



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 inter-related 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.

The project needs to start with Familiarisation, described in Chapter 3. Only after such familiarisation is it possible to describe the examination/test concerned through the Specification procedures (Chapter 4). Those procedures start with checks and reports on the evidence for the quality of the examination (reliability and validity); demonstration of such examination quality is a pre-requisite for the linking to proceed.

Because standard setting requires judgments of items and performances, the data obtained need to be of high quality. Therefore rigorous training of those involved in the process is needed and this is dealt with in Chapter 5.

There are a number of methods to set standards and the ones deemed to be the most relevant in the current context are described in Chapter 6. The quality of standard setting can vary and therefore it is important to show evidence of how defensible the standards are. Various types of validity evidence of standard setting, which need to be provided, are presented in Chapter 7.

Users of the Manual are asked to identify, from the range of techniques and options offered and similar techniques in the literature, those most appropriate and feasible for their context. The approach adopted is an inclusive one. The Manual aims to encourage the application of principles of best practice even in situations in which resources and expertise are limited. First steps may be modest, but the aim is to help examination providers to work within a structure, so that later work can build on what has been done before. The common structure advocated by the Manual may also offer the possibility for institutions to more easily pool efforts and seek synergies in certain areas.

It is important to emphasise that the five sets of procedures (or “stages”) are not just steps in a linear process undertaken in isolation one after another. It is vital to check at the conclusion of each stage that the endeavour is “on track”: that the interpretation of levels in the project does reflect the common interpretation illustrated by the illustrative samples. In the case of the revision or development of an examination, it is advisable to embed procedures recommended in the Manual at each stage of the test development/reform process so that the linking to the CEFR develops in an organic, cyclical way as the team becomes more and more familiar with the CEFR – and is not left to an external project undertaken by another department or external consultant before and after the main project is finished.

Although they should not be seen as linear steps, the five sets of procedures follow a logical order. At all stages it is recommended that users start with the productive skills (speaking and writing) since these can be more directly related to the rich description in the CEFR, thus providing a clear basis for training, judgments and discussion.

## **2.2. Quality Concerns**

Linking of a test to the CEFR cannot be valid unless the examination or test that is the subject of the linking can demonstrate validity in its own right. A test that is not appropriate to context will not be made more appropriate by linking to the CEFR; an examination that has no procedures for ensuring that standards applied by interviewers or markers are equivalent in severity, or that successive forms of tests administered in different sessions are equivalent, cannot make credible claims of any linkage of its standard(s) to the CEFR because it cannot demonstrate internal consistency in the operationalisation of its standard(s).

There are several good reference works which provide guidance for good practice in test development. This Manual will not discuss such good practice as its main aim is to provide guidance for standard setting. Chapter 7 addresses issues related to test development, piloting and analysis, the Reference Supplement offers additional information, especially on analysis techniques; but the reader is referred to the extensive literature on test development e.g. Alderson et al (1995), Davidson & Lynch (2002), Ebel & Frisbee (1986), Downing & Haladyna (2006), Milanovic (2002), Weir (1993), the collection of publications and materials produced in the “Into Europe” project under the auspices of the British Council Hungary ([www.examsreform.hu/Pages/Exams.html](http://www.examsreform.hu/Pages/Exams.html)).

The concern for good quality in language test development is also present in the following standards for good practice:

- EALTA (European Association of Language Testing and Assessment, [www.ealta.eu.org](http://www.ealta.eu.org)). The EALTA Guidelines for Good Practice in Language Testing and Assessment provide an accessible list of the most important questions that all those involved in assessment and testing practices (whether individuals or institutions) need to take into account before, during and after test development.
- ALTE (Association of Language Testers in Europe, [www.alte.org](http://www.alte.org)). The ALTE Code of Practice and ALTE Minimum Standards for Establishing Quality Profiles in Language Assessment provide a set of 17 detailed minimum standards that help examination providers to structure and evaluate their test development and administration processes.
- AERA (American Educational Research Association, [www.aera.net](http://www.aera.net)). AERA (1999) provides a comprehensive, authoritative collection of theory-based standards for educational and psychological testing.
- ILTA (International Language Testing Association, [www.ilta.org](http://www.ilta.org)). In addition, drawing on the AERA work and other authorities, ILTA has collated and summarised in the ILTA Code of Practice for language testers the most crucial principles in language testing theory and practice.

### 2.3. Stages of the Process

As mentioned earlier in this chapter, the process of linking a test to the CEFR consists of a set of procedures that need to be carried out at different stages:

**Familiarisation** (Chapter 3): A selection of training activities designed to ensure that participants in the linking process have a detailed knowledge of the CEFR, its levels and illustrative descriptors. This Familiarisation stage is necessary at the start of both the Specification and the Standardisation procedures. Familiarisation with the CEFR is equally a logical pre-requisite for effective linking. It is a good practice to assess and report how successful the familiarisation training has been.

**Specification** (Chapter 4): A self-audit of the coverage of the examination (content and tasks types) profiled in relation to the categories presented in CEFR Chapter 4 “Language use and the language learner” and CEFR Chapter 5 “The user/learner’s competences”. As well as serving a reporting function, these procedures also have a certain awareness-raising function that may assist in further improving the quality of the examination concerned. Forms A2 and A8–A20 in Chapter 4 focus on content analysis and the relationship of content to the CEFR. Specification can be seen as a primarily qualitative method: providing evidence through “content-based arguments”. There are also quantitative methods for content validation that can be considered (see, e.g. Kaftandjieva 2007).

**Standardisation Training and Benchmarking** (Chapter 5): The suggested procedures facilitate the implementation of a common understanding of the “Common Reference Levels”, exploiting CEFR illustrative samples for spoken and written performance. These procedures deepen the familiarity with the CEFR levels obtained through the kinds of activities outlined in Chapter 3 (Familiarisation) and assure that judgments taken in rating performances reflect the constructs described in the CEFR. It is logical to

standardise – through sufficient training – the interpretation of the levels in this way before moving on to (a) benchmarking local performance samples and tasks/items (Section 5.7), and (b) Standard setting (Chapter 6). Successful benchmarking of local samples may be used to corroborate a claim based on Specification. If the result of the benchmarking process is that performance samples from the test are successfully benchmarked to the levels that were intended in designing the test, this corroborates the claim based on Specification.

**Standard Setting** (Chapter 6): The crucial point in the process of linking an examination to the CEFR is the establishment of a decision rule to allocate students to one of the CEFR levels on the basis of their performance in the examination. Usually this takes the form of deciding on cut-off scores, borderline performances. The preceding stages of Familiarisation, Specification and Standardisation can be seen as preparatory activities to lead to valid and rational decisions. Chapter 6 describes procedures to arrive at the final decision of setting cut scores. The material presented there draws on an extensive literature on standard setting, and the procedures presented in Chapter 6 are a selection from the many available procedures deemed to be suitable in the context of language testing. Additional procedures based on the exploitation of teacher judgments and IRT to incorporate an external criterion (e.g. CEFR illustrative items, or teacher assessments with CEFR illustrative descriptors) into a linking study are presented in Extra Material provided by Brian North and Neil Jones.

**Validation** (Chapter 7): While the preceding stages of Familiarisation, Specification, Standardisation and Standard Setting can be conceived roughly to represent a chronological order of activities, it would be naïve to postpone validation activities until everything has been done, and to conceive it as an ultimate verdict on the quality of the linking process. Validation must rather be seen as a continuous process of quality monitoring, giving an answer to the general question: “Did we reach the aims set for this activity?” A simple, but nevertheless important example has already been referred to: it is important to provide CEFR familiarisation and standardisation training, but it is equally important to check if such activities have been successful; this is precisely what is meant by validation. Aspects of validity and procedures to collect validity evidence are described in this final chapter.

Aspects of validity and procedures on how to collect validity evidence have been put together in the final chapter (Chapter 7) of this Manual.

## 2.4. Use of the CEFR

A common framework of reference enables different examinations be related to each other indirectly without any claim that two examinations are exactly equivalent. The focus of examinations may vary but their coverage can be profiled with the categories and levels of the framework. In the same way that no two learners at Level B2 are at Level B2 for the same reason, no two examinations at Level B2 have completely identical profiles.

The parts of the CEFR most relevant for linking examinations are:

- Chapter 3 “The Common Reference Levels”;
- Chapter 4 “Language Use and the Language User” – with scales for Communicative Language Activities and for Communicative Language Strategies;
- Chapter 5 “The User/Learner’s Competences”, particularly Section 5.2 “Communicative Language Competences” with the illustrative scales for aspects of linguistic, pragmatic and socio-linguistic competence.

Users of this Manual will find the full text of the CEFR and related documents, plus a number of useful tools on the Council of Europe website, including the following:

### Documents

- The CEFR in English and French, including appendices.
- Links to other language versions on the Council of Europe website ([www.coe.int/lang](http://www.coe.int/lang); [www.coe.int/portfolio](http://www.coe.int/portfolio))
- The Manual, including appendices.
- The forms and reference Grids included in the Manual.

- The Reference Supplement.

### Content Analysis Grids

- CEFR Content Analysis Grid for listening and reading (sometimes referred to as “the Dutch CEFR Grid”): Appendix B1.
- CEFR Content Analysis Grids for speaking and writing, developed by ALTE: Appendix B2.

### Illustrative Descriptors ([www.coe.int/portfolio](http://www.coe.int/portfolio))

- The descriptors from the CEFR (in English).
- The descriptor bank from the European Language Portfolio, demonstrating the relationship between those descriptors and the original CEFR descriptors.
- A collation of C1/C2 descriptors (in English) from the CEFR and related projects that marks which descriptors were calibrated to CEFR levels and which were not.

### Illustrative Samples

- Documentation for DVDs of illustrative samples of spoken performance by adults, available at the time of writing for English, French, Italian and Portuguese<sup>1</sup>.
- Illustrative samples of written performance, available at the time of writing for English, French, German, Portuguese and Italian.
- Illustrative items for listening and reading for English, French, German, Italian and Spanish.

Other related resources will be added to this CEFR “Toolkit” listed on [www.coe.int/lang](http://www.coe.int/lang) and [www.coe.int/portfolio](http://www.coe.int/portfolio) as they become available.

### Especially Relevant Parts of the CEFR

In considering which specific resources in the CEFR to consult, the user may find the following scales and descriptions of the levels especially useful from a global perspective:

	<b>English Version</b>	<b>French Version</b>
<b>Overviews of the Common Reference Levels</b>		
• Table 1 “Common Reference Levels” in Chapter 3	Page 24	Page 25
• Section 3.6. “Content Coherence in Common Reference Levels”	Pages 33–36	Pages 32–34
• Document B5 “Coherence in Descriptor Calibration”	Pages 223–224	Pages 159–160
• “Levels of Proficiency in the ALTE Framework”	Pages 249-50	Pages 176–177
<b>Overviews of Communicative Activities</b>		
• Table 2, Portfolio Self-assessment Grid	Pages 26–27	Pages 26–27
• DIALANG Document C3 “Elaborated Descriptive Scales ...”	Pages 238–243	Pages 170–172
• ALTE Document D1: Skills Summaries	Page 251	Page 178
• Overall Listening Comprehension: scale	Page 66	Page 55
• Overall Reading Comprehension: scale	Page 69	Page 57
• Overall Spoken Interaction: scale	Page 74	Page 61
• Overall Written Interaction: scale	Page 83	Page 68
• Overall Spoken Production: scale	Page 58	Page 49
• Overall Written Production: scale	Page 61	Page 51
<b>Overviews of Aspects of Communicative Language Competence</b>		
• Table 3 “Qualitative Aspects of Spoken Language Use”	Pages 28–29	Page 28
• General Range: scale	Page 110	Page 87
• Grammatical Accuracy: scale	Page 114	Page 90
• Socio-linguistic Appropriateness	Page 122	Page 95
• Spoken Fluency	Page 129	Page 100

<sup>1</sup> The DVD for German is published with its documentation by Langenscheidt as Bolton et al (2008).

DVDs of spoken performance by teenage learners of English, French, German, Spanish and Italian, calibrated at the Cross-language Benchmarking Seminar in Sèvres in June 2008, will be available by early 2009.

In relation to examinations intended for the world of work or for university entrance, users may in addition find the following scales particularly relevant, since they cover the more specialised functional demands of such contexts.

### **Communicative Activities Particularly Relevant to the Occupational and Educational Domains**

	<b>English Version</b>	<b>French Version</b>
• Listening as a Member of a Live Audience	Page 67	Page 56
• Note-taking (Lectures, Seminars etc.)	Page 96	Page 77
• Reading for Orientation	Page 70	Page 58
• Reading for Information and Argument	Page 70	Page 58
• Reading Instructions	Page 71	Page 59
• Processing Text	Page 96	Page 77
• Information Exchange	Page 81	Page 67
• Formal Discussion and Meetings	Page 78	Page 64
• Understanding Conversation between Native Speakers	Page 66	Page 55
• Sustained Monologue: putting a case (e.g. in debate)	Page 59	Page 50
• Addressing Audiences	Page 60	Page 50
• Reports and Essays	Page 62	Page 52

The calibration of the CEFR illustrative descriptors is described in CEFR Appendix A, in North (2000a), North and Schneider (1998) and Schneider and North (2000).

## **2.5. Use of the Manual**

The chapters that follow address the different stages in the linking process and for each stage a series of procedures are presented from which users can select those most relevant or adequate for their context.

The Manual is *not* intended as a blueprint for the development of a new examination. However, it is intended to encourage reflection about good practice. Indeed, several users who piloted the preliminary edition commented that going through the procedures in the Manual was a good way to critically review and evaluate the content and the statistical characteristics of an examination – and that the outcome of this process was as important as the claim to linkage.

The Manual presents a principled set of procedures and techniques that provides support in what is a technically complicated and demanding process. Informed judgment is called for at several stages of the process. The responsibility for designing a coherent and appropriate linking process lies with the examination provider concerned. This responsibility involves:

- reflection on the needs, resources, expertise and priorities in the context concerned;
- selection of appropriate procedures from those explained – or others reported in the literature;
- realistic project planning in a modular, staged approach that will ensure results;
- collaboration and networking with colleagues in other sectors and countries;
- coordination of the participants in the local linking process;
- thoughtful application of the procedures;
- reliable recording of results;
- accurate, transparent and detailed reporting of conclusions.

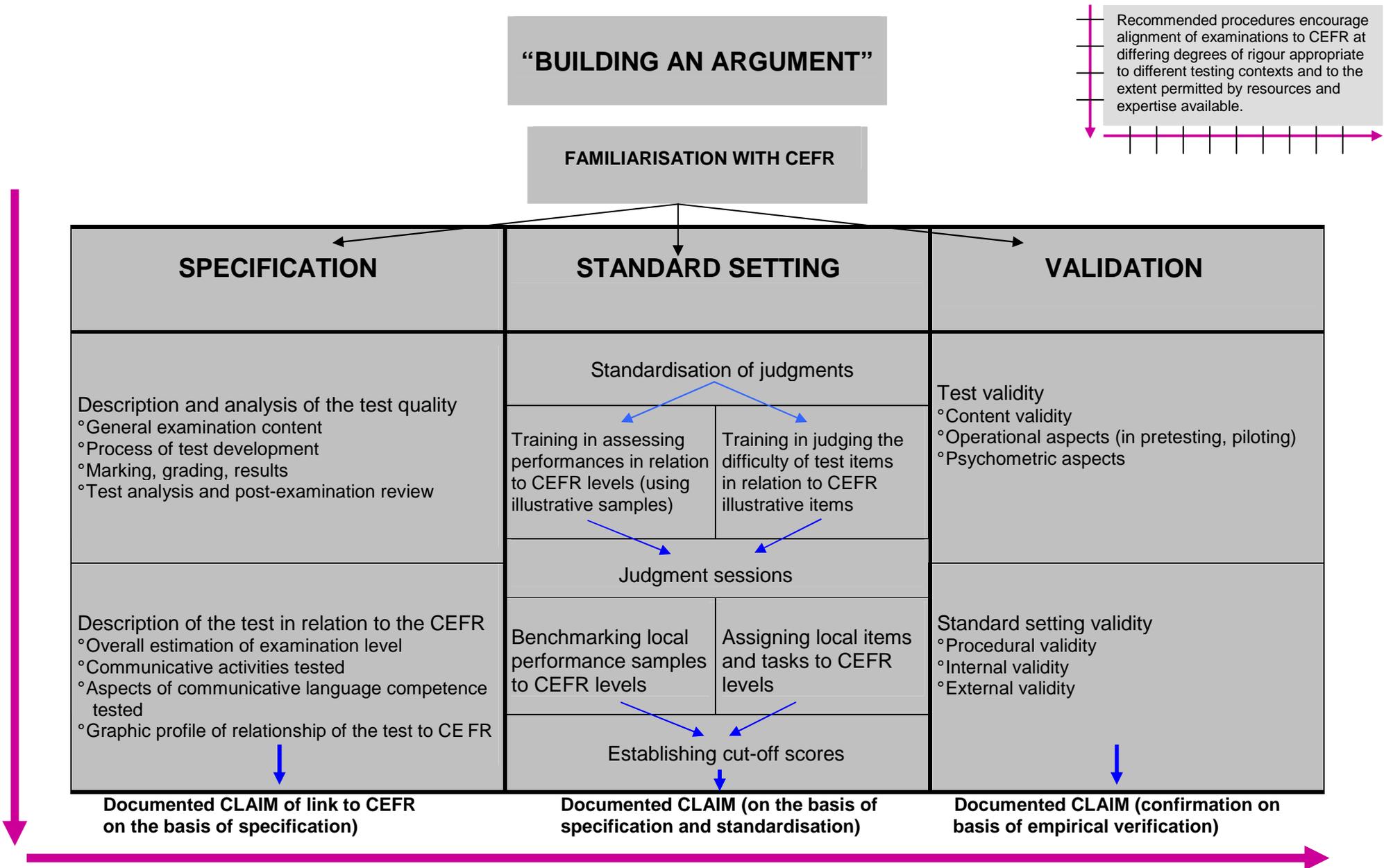
Figure 2.2 is a visual representation of the stages in the process of relating examinations to the CEFR. It highlights how linking an examination or a test can be seen as the development of a line of argument, making claims about different aspects of linkage and providing corroborating evidence of their validity as the process unfolds. Not all examination providers may consider they can undertake studies in all of the areas outlined in the Manual. However, even less well-resourced examination providers should select techniques from all areas. A claim that a qualification is linked to the CEFR can only be taken seriously if evidence

exists that claims based on specifications (content standards) and standard setting (performance standards) are corroborated through validation.

*Users of the Manual may wish to consider, before embarking on the linking project:*

- *what the approach proposed means in their context in general terms*
- *what the approach means in their context in more specific terms (time, resources,...)*
- *how feasible the different sets of procedures are in their context*
- *whether to focus in depth on one or two sets of procedures, or apply the principles of all five sets of procedures in a limited way, especially if resources are limited*
- *how they will justify their conclusions to the public and to their professional colleagues*

Figure 2.2: Visual Representation of Procedures to Relate Examinations to the CEFR



## Chapter 3

### Familiarisation

- 3.1. Introduction
- 3.2. Preparatory Activities before the Seminar
- 3.3. Introductory Activities at the Seminar
- 3.4. Qualitative Analysis of the CEFR Scales
- 3.5. Preparation for Rating

#### 3.1. Introduction

Before undertaking Specification and Standardisation, it is necessary to organise Familiarisation tasks to ensure that all those who will be involved in the process of relating an examination to the CEFR have an in-depth knowledge of it. Experience in case studies piloting the Manual and in benchmarking seminars producing DVDs has underlined the fact that many of the language professionals who take part in a linking project start with a considerably lower level of familiarity with the CEFR than they think they have. In particular, while most are familiar with the more global CEFR Tables 1 (Global scale) and Table 2 (Portfolio self-assessment grid), many do not have a clear picture of the salient features of the language proficiency in different skills of learners at the different levels.

In discussing Familiarisation, one needs to make a distinction between familiarisation with the CEFR itself, with the rating instruments to be used, and with the activities to be undertaken. There is no absolute boundary between the end of Familiarisation and the beginning of Specification or Standardisation; in each case the first activities in the main task are in effect a continuation of the Familiarisation process.

Another point to keep in mind has to do with the task at hand. One needs to bear in mind whether one is talking about a selected panel of experts or a full implementation of the CEFR in a team or in a whole institution, and what precise linking activities the particular Familiarisation session serves as an introduction to. The time that individuals will take to complete any familiarisation activity depends greatly on the level of familiarity they already have with the CEFR. The time the entire Familiarisation processes take (as repeated before Specification and Standardisation activities) will depend upon the aim and scope of the linking project concerned.

Panellists also tend to be much influenced by local institutional standards intended to be at CEFR levels and criterion descriptors for them or locally produced variants of CEFR descriptors. In addition, they are often unaware that there is a distinction between the level of descriptors for CEFR criterion levels (in all subscales plus the summary Tables 1, 2 and 3) and the CEFR “plus levels” (only found on subscales). It is important that those involved in the linking process focus on the CEFR criterion descriptors – and do not let their view of a CEFR level be over-influenced by descriptors that represent an exceptional performance at that level (a “plus level”).

Bearing these points in mind, this chapter proposes familiarisation activities in the four groups outlined below. In the rest of this chapter, these techniques are explained in more detail. Users are strongly advised to select activities from each group at the start of both the Specification process and of the Standardisation process.

### **Preparatory Activities before the Familiarisation Seminar**

Before a Familiarisation workshop, members of the project team should be asked individually to undertake several focused activities in order to reflect on important aspects of the CEFR levels.

- a) Reading Section 3.6 in the CEFR (English pp. 33–36) that describes the salient features of the levels, as made up by the illustrative descriptors.
- b) Considering a selection of the question boxes printed at the end of relevant sections of CEFR Chapter 3 (Common Reference Levels), Chapter 4 (Language Use and the Language User/Learner) and Chapter 5 (The User/Learner's Competences).
- c) Accessing the website CEFTTrain ([www.CEFtrain.net](http://www.CEFtrain.net)), which focuses on the salient characteristics of the levels and provides, for English, video examples, scripts and sample test items intended for the primary, secondary and adult sectors.

### **Introductory Activities at the Seminar**

- d) Sorting the text for the different levels in Table A1 in this Manual, which summarises the salient characteristics of the Common Reference Levels (CEFR 3.6).
- e) Self-assessment of own language level in a foreign language – using CEFR Table 2 (Portfolio self-assessment grid) – and discussion with peers.

### **Qualitative Analysis of the CEFR Scales**

- f) Sorting into piles by level or rank order the individual descriptors from a CEFR scale related to the skills that will be worked on. For example, for Speaking, one might use Fluency or two to three related CEFR scales for types of Spoken Production and/or Interaction (e.g. Conversation, Informal Discussion Turn-taking). The scale is chopped into its constituent descriptors for this task.
- g) Reconstructing CEFR Table 2 from the individual descriptors for each cell of the table.

### **Preparation for Rating the Productive Skills**

- h) Reconstructing the CEFR-based rating Grid that is going to be used, in which some of the cells have been emptied. If the seminar starts with Speaking, this will be CEFR Table 3 (Manual Table C2). If the seminar starts with Writing, this will be Table C4 in this Manual (or alternative).
- i) Illustrating the CEFR levels with videoed learner performances from the DVD for the language concerned.

## **3.2. Preparatory Activities before the Seminar**

Organisers of familiarisation activities should be aware of the clear difference between a presentation of the CEFR and a Familiarisation seminar/workshop. Whereas the former aims at a general introduction of the scope and content of the CEFR for a variety of purposes, the latter is expected to provide participants with sufficient awareness of the CEFR levels to analyse and assess test tasks and performances in relation to the CEFR levels.

In order to make the seminar as useful and successful as possible, it is highly recommended that the coordinator of the Familiarisation seminar prepares the necessary documents and information that can

allow participants to prepare for it, and sends a “pre-task pack” (by post or by e-mail) 2–3 weeks before the seminar. Those participants who have already attended a presentation on the CEFR will be able to “refresh” their knowledge, and those who have not will be able to study introductory materials about the CEFR. Whatever the participants’ degree of familiarity with the CEFR, the coordinator will need to inform them that preparing for the workshop individually may take a minimum of 3–5 hours if all three activities are included.

After the initial input on the CEFR, either of the following activities can be used as an introduction to the seminar proper and as a way of contributing to the cohesion of the group.

**a) Reading Section 3.6 in the CEFR (including Table A1)**

This activity is recommended when the organisers do not know for sure whether the participants are familiar with the CEFR levels, although it can also work as a “refresher” for those who already are. The task that participants are given is to read the levels in Table A1 and the text in Section 3.6., in order to be able to identify the salient features for each level and in order to ascertain in which level they would place the learners they work with (the work done individually before the seminar can be taken up at the seminar as an introductory activity and/or as an ice breaker, providing a useful link with pre-seminar work).

**b) Consideration of a selection of CEFR question boxes**

This activity is more appropriate when the majority of professionals involved are expected to be already somewhat familiar with the CEFR levels (for example, have worked with the CEFR or know the levels). The objective of the exercise is to make the participants aware of the many possible facets to consider when developing and analysing test tasks and also of the comprehensiveness of the scope of the CEFR.

There are a number of ways in which this activity can be prepared:

- A boxed checklist like the one focusing on speaking, which is presented below, might be photocopied so that participants are led to reflect on the different facets in assessing speaking.

<p><i>Users of the Framework for the purpose of analysing and assessing speaking performances may wish to consider and, where appropriate, state:</i></p> <ul style="list-style-type: none"> <li>– <i>how the physical conditions under which the learner will have to communicate will affect what he/she is required to do;</i></li> <li>– <i>how the number and nature of the interlocutors will affect what the learner is required to do;</i></li> <li>– <i>under what time pressure the learner will have to operate;</i></li> <li>– <i>to what extent the learners will need to adjust to the interlocutor’s mental context;</i></li> <li>– <i>how the perceived level of difficulty of a task might be taken into account in the evaluation of successful task completion and in (self) assessment of the learner’s communicative competence.</i></li> </ul>	<p>Relevant / Why?</p>
--	------------------------

- The coordinator(s) might themselves make a selection of CEFR Question Boxes that seem particularly relevant and make up a different checklist, depending on what skills are to be focused on during the seminar.

- The coordinators may draw on the work done by the participants in this activity when discussing the sorting exercises (f–g) in Section 3.4.

### **c) Accessing the CEFTTrain website**

The CEFTTrain project<sup>2</sup> developed a selection of activities to familiarise teachers with the CEFR levels. It contains exercises with the CEFR scales and tasks and performances (for primary, secondary and adult sectors) analysed and discussed in relation to the CEFR levels on the basis of the agreed ratings of the project members. Accessing this website is very useful to provide participants with a hands-on example of what will take place during the seminar. Participants should be advised to concentrate on the sector most relevant for them, and to focus on the skills that will be dealt with during the seminar.

### **3.3. Introductory Activities at the Seminar**

After welcoming the participants, the coordinator will ensure that they all have a good understanding of what the seminar will be about and its timetable.

The first activity of the seminar will be a brief input session on the relevance of the CEFR in the field of testing. After this, the coordinator will proceed with one or both of the activities below, making sure that participants can draw on the work they have done prior to the seminar.

#### **d) Sorting the text for the different levels in Table A1**

This is a good activity to relate the seminar to the work done individually before the seminar.

- Participants are presented with a sorting exercise based on the Salient characteristics cells in Table A1 of this Manual, which simplifies CEFR Section 3.6. Level references should be eliminated so that participants need to read the descriptors really carefully. The coordinator presents the participants with a sheet containing the 10 descriptors in a jumbled order. The task is to assign the descriptors to levels A1–C2.
- Once the participants have finished this task, and in order to provide the “answer key”, the full Table A1 will be distributed.
- The coordinator then asks the participants to share – in pairs or small groups – their views on the salient features of each of the CEFR levels, on the basis of their individual study of CEFR Table 1 and Section 3.6 (activity a), and the sorting exercise they just undertook. One way to do this in a concrete fashion is to ask the participants to highlight key elements in colour.
- As a follow up, participants could be asked which level they feel might be the most relevant to their professional work. Then they can be grouped with others interested in the same level, and given a checklist of CEFR descriptors for that level, such as can be found in the Swiss prototype ELP available at [www.sprachenportfolio.ch/esp\\_e/esp15plus/index.htm](http://www.sprachenportfolio.ch/esp_e/esp15plus/index.htm) (select from left menu: ELP Model 15+; Learners; Downloads).

#### **e) Self-assessment using CEFR Table 2**

This is a particularly good starting point for groups of participants who are already familiar with the European Language Portfolio. CEFR Table 2 is an important part of the ELP and often referred to as the ELP grid.

---

<sup>2</sup> The CEFTTrain project was an EU Socrates funded project coordinated by the University of Helsinki with partners from four other countries: Italy, Austria, Germany and Spain, including the involvement of Neus Figueras, one of the authors of this Manual.

- Participants are asked to self-assess their ability in two foreign languages with the ELP grid (CEFR Table 2). They then discuss this with neighbours. The amount of discussion so generated should not be underestimated. It is important to guide the discussion in such a way that participants become aware of the existence of uneven language profiles and the session leader can explain how the CEFR takes into account their existence and fosters their recognition.
- It is a good idea to supplement this initial self-assessment (as ELP users are advised to do) by consulting a checklist of CEFR descriptors for the level concerned, such as can be found in the Swiss prototype ELP already mentioned.
- As an alternative, or in addition, participants could be asked to self-assess the *quality* of their foreign language(s): how well they can do what they can do. For this task one could use either:
  - a) CEFR Table 3 (Table C2) defining each level for Linguistic Range, Grammatical Accuracy, Fluency, Coherence and Interaction
  - b) The CEFR Fluency Scale (English page 129), and Accuracy Scale (English page 114).

### 3.4. Qualitative Analysis of the CEFR Scales

Once the introductory activities phase has been completed, the Familiarisation phase should proceed with more in-depth work and discussion of CEFR levels in relation to the descriptors for the specific skill concerned. The coordinator should select at least one of the two options presented.

#### f) *Sorting the individual descriptors from a CEFR scale*

Descriptor sorting was used extensively in the Swiss project which developed the CEFR descriptors, in Portfolio development in several contexts, and in several Finnish projects. This activity is effective because it forces participants to consider the descriptors in isolation from each other as independent criteria.

However, the activity requires some preparation and it is best to keep this activity relatively simple.

- The coordinator prepares in advance envelopes for each person or pair. Each envelope contains a scale or several scales chopped up into their constituent descriptors like strips of ticker tape. If related scales are mixed, (e.g. Conversation, Informal Discussion, Turn-taking) it is best to ensure that the total number of individual descriptors does not exceed 40! If scissors are used for the chopping, it is best to cut twice between each pair of adjacent descriptors, discarding the empty middle strip, in order to eliminate “clues” caused by one’s skill at cutting straight! It is also a good idea to ask the participants not to write on the descriptors – so they can be used again.
- Participants, either individually or in pairs, then sort the descriptors into levels. They may start with “A”, “B” and “C” and then sub-divide, or go straight to the six levels, as they wish.
- Then they discuss with neighbouring participants/pairs and reach a consensus.
- Then they compare their solutions with the right answer.

It is to be expected that some descriptors will get reversed in the process, but generally, provided a consensus-building phase has taken place, the order will normally more or less repeat that in the CEFR scales.

#### g) *Reconstructing CEFR Table 2*

This activity is a variant of the previous one, but using CEFR Table 2 (the ELP grid) – itself constructed from CEFR descriptors – rather than CEFR scales themselves. There is a multi-scale variant (6 language activities x 6 levels = 36 descriptors) or a simpler version (one column = 6 descriptors). The chopped up cells of the table are again best presented in an envelope.

- One can provide an A3 piece of paper, blown up from the ELP grid, but with all the cells empty. Participants can then be asked to place the descriptors into the correct cells.
- Symbols for the different skills can be put on the descriptors to save participants from wasting time in finding out that “I can use simple phrases and sentences to describe where I live and people I know” is intended as a Spoken descriptor – (Spoken Production).
- This activity can also be done with only half the cells in the table deleted. This is advisable with big groups and also with rooms without big tables.

A combination of this reconstruction activity with self-assessment of own language level (c: above) has been found to be particularly effective if done as follows:

- Participants, in small groups, carefully read and discuss each descriptor to reconstruct the table. The coordinator supervises group work and helps to clarify doubts about the interpretation of the different descriptors.
- The coordinator distributes a copy of the completed and “whole” CEFR Table 2 for participants to check their reconstruction exercise and to facilitate discussion.
- Participants are asked to self-assess their own knowledge of foreign languages (first individually) and then to discuss it with their group in terms of CEFR levels and skills, as these are described in CEFR Table 2.

### **3.5. Preparation for Rating**

Having assured a thorough familiarisation with the CEFR levels, the last phase of the familiarisation can start. This involves preparing the participants in more detail for the rating of tasks and performances in the relevant skill(s). If the work to follow is to assess reading and listening tasks only, the coordinator may decide not to carry out activity (i). Activity (h) is – on the contrary – mandatory for each skill before the rating starts.

#### ***h) Reconstructing CEFR-based rating Grid to be used***

The coordinator will prepare this activity on the basis of the scale to be used to rate the tasks/performances.

The exercise is done in exactly the same way as described in (f) (sorting CEFR descriptors) above.

An alternative to the sorting technique with chopped descriptors in an envelope is to use a checklist-type form with the levels of the skill descriptors jumbled. The participants then have to label each descriptor with its correct CEFR level (as described in (d) above).

The coordinator prepares an “answer key” checklist to give to the participants after a good number of descriptors have been discussed and “corrected” with the whole group.

#### ***i) Illustrating the CEFR levels with student videoed performances***

This activity provides a very good, tangible grasp of the CEFR levels and is relevant even if the participants are not going to be working on speaking.

The activity can only be carried out if the coordinator has access to the published CEFR illustrative sample performances. Care should be taken in selecting those performances which are most relevant to the participants in terms of level and age group. The suggested procedure is as follows:

- The coordinator plays the selected performance(s) once and asks the participants to assign a level to it according to Table A1.

- Before discussion participants are given CEFR Table 3 (Table C2) and are asked to confirm their initial level assignment individually.
- The coordinator then fosters discussion in groups of the level(s) assigned in relation to the descriptors in CEFR Table 3 (Table C2).
- The coordinator gives the participants the level assigned to the performance in the published video and distributes the documentation for it, which states why this is the case, with reference to the descriptors of CEFR Table 3 (Table C2).

**Table 3.1: Time Management for Familiarisation Activities**

<b><u>Familiarisation</u></b>	
• can be organised independently from any other training activity, and can be recycled at the start of the Specification and the Standardisation activities.	
• takes about 3 hours:	
– Brief presentation of CEFR Familiarisation seminar by the coordinator	(30 mins)
– Introductory activity (d–e) and discussion	(45 mins)
– Qualitative activity (f–g) including group work	(45 mins)
– Preparation for rating (h–i)	(45 mins)
– Concluding	(15 mins)

**Table 3.2: Documents to be Prepared for Familiarisation Activities**

- Preparatory pack (to be sent to participants by post or email), with instruction sheet:
  - Table 1 in the CEFR
  - Section 3.6
  - Question checklists based on those CEFR reflection boxes (at end of each chapter) best suited to context
- Copies of jumbled edited descriptors of salient characteristics on Table 2.1 for all participants
- Copies of Table A1 in the Manual for all participants
- Copies of CEFR Table 2 for all the participants (all contexts)
- Cut up versions of CEFR Table 2 for group work (all contexts, one set per envelope, one envelope per working group)
- Cut up CEFR scales, as appropriate to the assessment in question (more detailed briefing on a particular skill: one chopped scale per envelope, envelopes for each working group), e.g.
  - for Speaking: (1) Overall Spoken Interaction; (2) Spoken Fluency; (3) General Linguistic Range
  - for Listening: (1) Overall Listening Comprehension, (2) Understanding Conversation Between Native speakers, (3) Listening to Audio Media and Recordings
- Copies of ELP checklists of descriptors<sup>3</sup> for one or two particular levels across the whole set of CEFR scales (more detailed briefing on a particular level)

<sup>3</sup> For this purpose, only descriptors from a validated ELP should be used; it should be possible to trace each ELP adapted wording back to an original CEFR descriptor – as for example in the descriptor bank prepared by Günther Schneider and Peter Lenz on [www.coe.int/portfolio](http://www.coe.int/portfolio)

### **Table 3.2: Documents to be Prepared for Familiarisation Activities (continued)**

- Copies of CEFR Table 3 (Table C2) when applicable
- Selection of two student videoed performances from published illustrative samples
- Documentation for the performance samples to be used

*Users of the Manual may wish to consider:*

- *how well the overall aims and functions of the CEFR are familiar to the panel*
- *what strategy is needed to consolidate familiarisation with the CEFR*
- *whether panellists should be asked to (re-)read certain chapters/sections in addition to CEFR 3.6*
- *which CEFR Question Boxes might be most useful*
- *whether a CEFR “pre-task” should be collected and analysed, or done informally*
- *which CEFR scales would be best to use for sorting exercises*
- *whether to use CEFR illustrative samples on DVD at this stage*
- *a method of knowing whether more familiarisation is needed – e.g. a CEFR quiz?*
- *whether the outcomes of the Familiarisation phase suggest any changes to the planning*

# **Chapter 4**

## **Specification**

### **4.1. Introduction**

### **4.2. General Description of the Examination**

### **4.3. Available Specification Tools**

#### **4.3.1. Manual Tables and Forms**

#### **4.3.2. Content Analysis Grids**

##### **4.3.2.1. The CEFR Content Analysis Grid for Listening & Reading**

##### **4.3.2.2. The CEFR Content Analysis Grids for Speaking & Writing**

#### **4.3.3. Reference Works**

### **4.4. Procedures**

### **4.5. Making the Claim: Graphical Profiling of Relationship of the Examination to the CEFR**

## 4.1. Introduction

This chapter deals with the content analysis of an examination or test in order to relate it to the CEFR from the point of view of coverage. This might be done by discussion, or by individual analysis followed by discussion. The end product is a claim by the institution concerned of a degree of linking to the CEFR based on Specification, profiling their examination in relation to CEFR categories and levels.

However, as pointed out in Chapter 2, such a claim makes little sense unless it is accompanied by evidence of good practice, internal validity and adequate quality procedures for all the steps of the test development and administration cycle.

The chapter has three aims:

To contribute to increasing awareness:

- of the importance of good content analysis of language examinations;
- of the CEFR, especially its descriptor scales;
- of the rationale for relating language examinations to an international framework like the CEFR;
- of ways in which the CEFR can be exploited in planning and describing language examinations.

To define minimum standards in terms of:

- the quality of content specification in language examinations;
- the process of linking examinations to the CEFR.

To provide practical support to help users to:

- complete the proposed content analysis and linking process;
- provide evidence of internal consistency and construct validity;
- report a claim that makes the results of the examination in question more transparent to both the users of examination results and to test takers themselves.

The specification procedures outlined in the chapter involve four steps:

- assuring adequate familiarisation with the CEFR (Chapter 3);
- analysing the content of the examination or test in question in relation to the relevant categories of the CEFR; should an area tested not be covered by the CEFR, the user is asked to describe it;
- profiling the examination or test in relation to the relevant descriptor scales of the CEFR on the basis of this content analysis;
- making a first claim on the basis of this content analysis that an examination or test in question is related to a particular level of the CEFR.

The procedures involve three types of activity:

- familiarisation activities as described in Chapter 3;
- filling in a number of checklists with details about the content of the language examination;
- using relevant CEFR descriptors to relate the language examination to the levels and categories of the CEFR.

This Specification process gives examination providers the opportunity to:

- increase the awareness of the importance of a good content analysis of examinations;
- become familiar with and use the CEFR in planning and describing language examinations;
- describe and analyse in a detailed way the content of an examination or test;
- provide evidence of the quality of the examination or test;
- provide evidence of the relation between examinations/tests and the CEFR;

- provide guidance for item writers;
- increase the transparency for teachers, testers, examination users and test takers about the content and quality of the examination or test and its relationship to the CEFR. The forms to be filled in have an awareness-raising function (process) and are also sources of evidence to support the claim made (product).

The procedures that are proposed in this chapter are not the only ones that exist. They have been designed for the current purpose. Users may wish to consult other procedures published in the literature for relating an examination to a framework through descriptive analysis (e.g. Alderson et al (1995: Chapter 2), Davidson and Lynch (1993; 2002), Lynch and Davidson (1994; 1998).

## 4.2. General Description of the Examination

The first step in embarking on a linking project is to define and describe clearly the test that is going to be linked to the CEFR. Does the test have sufficient internal validity? Are there areas in which further development work would be advisable in order to increase or confirm the quality of the test and thus the meaningfulness of the result of the subsequent linking project? Experience in the case studies which piloted the preliminary version of this Manual showed that this process offered an excellent opportunity to step back from operational concerns and reflect on the extent to which the examination, and procedures associated with it, was meeting its aims. This is an awareness-raising process which cannot be undertaken by a single researcher or team member. Sometimes, this exercise throws up a lack of coherence between official test specifications, which may not have been revised for some years, and the test in practice – as represented by forms of the test administered in recent sessions. The exercise is certainly easier to complete if formal test specifications exist. If they do not exist, the process of completing the forms associated with this chapter will help the user to consider aspects that should be included in such a specification.

Section A2 in the Appendix contains the following forms:

- A1:** General Description of the Examination
- A2:** Test Development
- A3:** Marking
- A4:** Grading
- A5:** Reporting Results
- A6:** Data Analysis
- A7:** Rationale for Decisions

To complete the forms, users should have available both the specification and copies of the last three administered test forms. If the subject of the linking project is a suite of examinations at different levels, the forms should ideally be completed for each individual exam.

Form A1 asks for definition of the examination purpose and objectives, and target population, plus an overview of the communicative activities tested, the different subtests and subsections and the information and results made available to test users.

Forms A2–A6 describe the most important steps in the examination cycle. They require information about the development of the examination, marking, grading, reporting results and data analysis, as described below:

- the test development process (Form A2);
- the marking schemes and scoring rules for different subtests (Form A3);
- the grading and standard setting procedures for different subtests (Form A4);
- the reporting of results (Form A5);
- the analysis and review procedures (Form A6).

Form A7 (Rationale) is the opportunity for the examination provider to explain and justify decisions. For example: why are these areas tested and not others? Why is this particular weighting used? Why is double marking only used in exceptional cases? If no profile of results across subtests (or skills) is provided, why is this? Is it a reliability issue or a policy decision?

Form A8 then records the institution's initial estimation of the overall CEFR level at which the examination or test is situated.

<b><i>Initial Estimation of Overall CEFR Level</i></b>		
<input type="checkbox"/> A1	<input type="checkbox"/> B1	<input type="checkbox"/> C1
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> A2	<input type="checkbox"/> B2	<input type="checkbox"/> C2
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<b>Short rationale, reference to documentation</b>		

***Form A8: Initial Estimation of Overall Examination Level***

The detailed specification process is reported in Forms A9–A22. (Please see Appendix A, Sections A2–A5). Form A23 presents the outcome of this specification process as a graphic profile of the examination coverage in relation to the most relevant categories and levels of the CEFR. This form is discussed with an example in Section 4.4.

The procedures described here have been designed for the current publication. They are, of course, not the only ones that have been developed with the aim of specifying examination test tasks and users may wish to consult other procedures published in the literature for relating an examination to a framework through descriptive analysis (e.g. Alderson et al 1995; Davidson and Lynch 2002), or other instruments that exist for the content analysis of an examination.

The procedures should be followed in the same way for both general purpose examinations and examinations for specific purposes. The CEFR itself discusses domains (public, personal, educational, occupational) and the main reason for the grouping of communicative language activities under Reception, Interaction, Production, Mediation rather than the traditional four skills is precisely because this categorisation encompasses educational and occupational activities more effectively.

**4.3. Available Specification Tools**

There are three different types of CEFR-based tools that are available – in addition to the CEFR publication itself, available in 36 languages at the time of writing:

- the tables and forms in the Appendices to this Manual;
- Content Analysis Grids that offer the possibility to work at the more detailed level of individual test tasks, classifying them by standard criteria;
- reference works for different languages: especially useful for linguistic specifications.

### **4.3.1. Manual Tables and Forms**

This chapter refers to a set of tables derived from the CEFR descriptor scales, with related forms to fill in. Since the CEFR is designed to be comprehensive, the number of forms in this chapter is quite extensive. The forms in this chapter can be found in Sections A2–A5 in the Appendix, but are also available for downloading on the website [www.coe.int/lang](http://www.coe.int/lang)

Forms and related tables are provided for each of the Communicative Language Activities (CEFR Chapter 4) and for the Aspects of Communicative Language Competence (CEFR Chapter 5). The forms are tools to provide a detailed analysis of the examination or test in question and to relate that examination/test to the relevant subscales of the CEFR. In most of the forms, a short description, reference and/or justification is asked for.

In case studies piloting this Manual, several users commented that completing these forms was a very good way to review and evaluate the coverage of the examination and to re-assess its fitness for its stated purpose.

### **4.3.2. Content Analysis Grids**

The CEFR Content Analysis Grid for Listening & Reading and the CEFR Content Analysis Grids for Speaking & Writing have been developed to offer users of the Manual an opportunity to operate at a greater level of detail than that provided solely by the CEFR subscales, and the associated tables in Appendix A referred to in Section 4.2. above, since individual test tasks are categorised.

In case studies piloting this Manual, some users exploited these Grids and found them more useful to their purposes than the actual forms referred to above. Users who are interested in using the Manual to assist in the development of a new examination or to make a formal critical review of an existing examination or test may find them particularly useful.

The most recent copies of the Grids plus illustrative samples using completed Grids can be downloaded from [www.coe.int/portfolio](http://www.coe.int/portfolio)

#### **4.3.2.1. The CEFR Content Analysis Grid for Listening & Reading**

The CEFR Content Analysis Grid for Listening & Reading is an on-line tool that allows test developers to analyse tests of Reading and Listening, in order to relate them to the CEFR<sup>4</sup>. Information about each task, text and item in the test is entered into the Grid by specifying their characteristics (e.g. text source, discourse type, estimated difficulty level, etc.) from a set of options derived directly or indirectly from the CEFR. The analyst must, however, be fully familiar with the CEFR in order to use the Grid effectively. For further guidance the system therefore also includes a familiarisation component.

A link to the on-line version of the Grid is also available on [www.coe.int/portfolio](http://www.coe.int/portfolio). The direct link is [www.lancs.ac.uk/fss/projects/grid](http://www.lancs.ac.uk/fss/projects/grid)

A paper version of the Grid is included in Appendix B.

---

<sup>4</sup> With the financial support of the Dutch Ministry of Education, a working group consisting of J. Charles Alderson (Project Coordinator), Neus Figueras, Henk Kuijpers, Günther Nold, Sauli Takala and Claire Tardieu developed an instrument for describing and rating listening and reading tasks following the CEFR as closely as possible. With further funding from the Dutch Ministry of Education the group developed a computerized version which is available at [www.lancs.ac.uk/fss/projects/grid](http://www.lancs.ac.uk/fss/projects/grid). This tool was originally referred to informally as “The Dutch Grid.”

– For more information see Section B1 in the Appendix and Alderson et al (2006). A full report on the project is available on request from the Project Coordinator at [c.alderson@lancaster.ac.uk](mailto:c.alderson@lancaster.ac.uk)

It is possible to supplement the Grid with new categories (e.g. related to the curriculum/syllabus) in the paper version.

While the Grid was developed to analyse tests of reading and listening, it can also be used as a tool in planning such tests. In certain case studies during the piloting of this Manual, it was also used in Standardisation training (See Chapter 5).

#### 4.3.2.2. The CEFR Content Analysis Grids for Speaking & Writing

*The CEFR Grids for the Analysis of Speaking and Writing Tasks* have also been designed to help users to describe the features of their test tasks in relation to the CEFR in a standardised way. The Grids<sup>5</sup>, to be modified as the need arises, are each available on the Council of Europe website. There are two modes for each of the two Grids: one for analysis and one for presentation/reporting. For more information on the Grids, please see Appendix B2.

**The Analysis (“Input”) Grids:** These two Grids are suitable for use in workshops in which participants complete the Grid(s) for a given set of test tasks. The aim is to profile the features of tasks, expected performances (answer length, discourse types, register etc.) rating instruments and feedback given to candidates. An example task is accompanied by this analysis plus a sample answer complete with score allocated and a commentary. The Grids are useful for training task developers for standardising test tasks presented for different languages at the same level.

Completing the Grids can also form a useful bridge between Specification and Standardisation of the interpretation of CEFR levels with illustrative samples (See Chapter 5). They can also be used to select local samples that are going to be used for benchmarking (See Chapter 5).

**The Presentation (“Output”) Grids:** This simpler form of the Grids is intended to report the description of the test tasks created with the Analysis Grid (for Speaking or Writing) discussed above. They provide the detailed information which, when supplemented by appropriate references to the CEFR qualitative criteria (e.g. CEFR Table 3; Manual Table C.2) for each benchmarked sample, can provide the basis for good documentation and examination user guides.

#### 4.3.3. Reference Works

The content analysis in the Specification procedures takes as its main reference point the CEFR itself. However, as a common framework the CEFR is by definition language-independent. For detailed content specifications for specific languages, the following supplementary reference works may be useful:

- The series of content specifications related to the CEFR, which were developed in association with the Council of Europe in the 1970s–1990s *before* the development of the CEFR. For English the list of specifications is as follows: A1: *Breakthrough*<sup>6</sup>; A2: *Waystage* (van Ek and Trim 2001a); B1: *Threshold Level* (van Ek 1976; van Ek and Trim 2001b); B2: *Vantage Level* (van Ek and Trim 2001c).

---

<sup>5</sup> The Speaking and the Writing Grids were each produced by the ALTE Manual Special Interest Group in cooperation with the Council of Europe. The history of the Grids dates back to the *ALTE Content Analysis Checklists*. Developed with LINGUA funding (93-09/1326/UK-III) in 1993, the aim was to facilitate systematic comparison of examination materials across various languages. In the development of the Grids described here, attention was also paid to the work of the Dutch Construct Project – which produced the Listening and Reading Grid.

<sup>6</sup> *Breakthrough* has not been published, but is available from the Council of Europe and ALTE Secretariats.

- The series of CEFR-related “Reference Level Descriptions” that have been developed for different languages *since* the development of the CEFR. An up-to-date list can be found at [www.coe.int/lang](http://www.coe.int/lang) and includes the following:
  - for German: Glaboniat, M., Müller, M., Schmitz, H., Rusch, P., Wertenschlag, L. (2002/5) *Profile Deutsch (A1–A2, B1–B2, C1–C2)*, Berlin: Langenscheidt;
  - for French: Beacco et al (2004, 2006, 2007, 2008) *Niveau B2/A2/A1/A1.1 pour le français: un référentiel*;
  - for Spanish: Instituto Cervantes (2007) *Niveles de referencia para el español – Plan Curricular del Instituto Cervantes: A1, A2–B1, B2–C1, C2*;
  - for Italian: Parizzi, F. and Spinelli, B. (forthcoming) *Profilo della Lingua Italiana*, Firenze: La Nuova Italia.

#### 4.4. Procedures

The procedures involve consulting the CEFR, the Appendices to this Manual and other sources referred to above, before systematically completing the series of forms provided in Appendix A and available electronically from [www.coe.int/lang](http://www.coe.int/lang)

1. **Selecting the Panel:** A first step is the setting up of a panel of experts from within and (if possible) from outside the organisation or institute and to designate a coordinator. The group of internal and external experts should consist of representatives of the different key stages in language testing development.
2. **Familiarisation:** Before starting the Specification procedures, it is essential that the panel becomes familiar with the CEFR itself. Therefore the place to start is with the Familiarisation Activities in Chapter 3.
3. **Selection of Approach:** Afterwards, the group needs to become familiar with the different forms and the related tables, plus the other specification tools outlined in Section 4.2 and take decisions on the approach to be taken and the forms and/or Grids to be completed. It is not intended that *all* the forms in Appendix A should be completed. It must be stressed that only those forms relevant to the content of the examination should be completed; the group selects those forms that are relevant for the analysis of the examination in question. To give two examples: if an examination consists of only vocabulary tasks, then only the relevant forms should be filled in and only the relevant vocabulary range scale should be looked at. If an examination measures several linguistic competences for different skills, more forms should be filled in and more scales should be looked at.

*The minimum standard* is that the following forms should be completed:

- the forms in Phase 1 (General Description: A1–A7);
  - Form A8 (Initial Estimation of Overall Examination Level);
  - those forms – ranging from A9–A22 – that are relevant to the examination or test tasks in question;
  - Form A23 (Form A23: Graphic Profile of the Relationship of the Examination to CEFR Levels);
  - Form A24 (Confirmed Estimation of Overall Examination Level);
  - relevant evidence to support the claim made.
4. **Communicative Language Activities:** The forms for Communicative Language Activities (Forms A9–A18) are normally completed first. As has been said before, each of these forms can be filled in by the appropriate person in the institution involved. However, a more interactive procedure for filling in the forms may be desirable. The information provided in the forms will be

more reliable when more than one person has been involved. So each member of the panel fills in one or more of the selected forms. After having filled in the forms the panel meets and comes to agreement on what has been filled in.

Table 4.1 gives an overview of the forms and related CEFR scales that are provided. At the end of most of the forms users are asked for a comparison of the subtest concerned with a relevant CEFR subscale.

Form	Communicative Language Activity	Form	Scale
A9	Listening Comprehension	✓	✓
A10	Reading Comprehension	✓	✓
A11	Spoken Interaction	✓	✓
A12	Written Interaction	✓	✓
A13	Spoken Production	✓	✓
A14	Written Production	✓	✓
A15	Integrated Skill Combinations	✓	
A16	Integrated Skills	✓	✓
A17	Spoken Mediation	✓	
A18	Written Mediation	✓	

	RECEPTION		INTERACTION		PRODUCTION		MEDIATION	
	Listening	Reading	Spoken Interaction	Written Interaction	Spoken Production	Written Production	Spoken Mediation	Written Mediation
<b>Linguistic Competence</b>								
▪ General Linguistic Range	✓	✓	✓	✓	✓	✓	✓	✓
▪ Vocabulary Range	✓	✓	✓	✓	✓	✓	✓	✓
▪ Vocabulary Control			✓	✓	✓	✓	✓	✓
▪ Grammatical Accuracy			✓	✓	✓	✓	✓	✓
▪ Phonological Control			✓		✓		✓	
▪ Orthographic Control				✓		✓		✓
<b>Socio-linguistic Competence</b>								
▪ Socio-linguistic Appropriateness	✓	✓	✓	✓	✓	✓	✓	✓
<b>Pragmatic Competence</b>								
▪ Flexibility			✓	✓			✓	✓
▪ Turntaking			✓					
▪ Thematic Development	✓	✓		✓	✓	✓	✓	✓
▪ Cohesion and Coherence	✓	✓			✓	✓	✓	✓
▪ Spoken Fluency			✓		✓		✓	
▪ Propositional Precision	✓	✓			✓	✓	✓	✓
<b>Strategic Competence</b>								
▪ Identifying cues/infering	✓	✓					✓	✓
▪ Turntaking (repeated)			✓					
▪ Cooperating			✓	✓				
▪ Asking for clarification			✓	✓				
▪ Planning					✓	✓		✓
▪ Compensating			✓	✓	✓	✓	✓	✓
▪ Monitoring and Repair			✓	✓	✓	✓	✓	✓

5. **Communicative Language Competence:** Next, the forms for Aspects of Communicative Language Competence should be completed (Forms A19–A22). Table 4.2. gives an overview of the different communicative competences for which information can be provided. This section is organised in a different way. First the CEFR descriptors are provided in a tabular form. Secondly, users are asked to fill in the relevant form on the basis of an analysis of the examination or test tasks in question. At the end of each form users are asked to compare the examination with the relevant CEFR scale presented beforehand. A description and an indication of the level should be given for each of the aspects of competences distinguished in the CEFR that are relevant. The same group of experts can complete the forms in an interactive way.

The forms are provided in the following order:

- Reception (Form A19);
- Interaction (Form A20);
- Production (Form A21);
- Mediation (Form A22).

For Mediation no CEFR scale is provided. Users are asked to refer to the descriptors for Reception and Production.

#### 4.5. Making the Claim: Graphical Profiling of Relationship of the Examination to the CEFR

Once the examination has been analysed in terms of the categories of the CEFR, the result of the content linking should be profiled graphically. This graphical presentation profiles the content of the examination in question in terms of the relevant CEFR subscales for Communicative Language Activities and for Aspects of Language Competence (see the example of a completed Form A23 below).

<b>C2</b>								
<b>C1</b>								
<b>B2.2</b>								
<b>B2</b>								
<b>B1.2</b>								
<b>B1</b>								
<b>A2.2</b>								
<b>A2</b>								
<b>A1</b>								
<b>Overall</b>	<b>Listening</b>	<b>Reading</b>	<b>Social Conversation</b>	<b>Information Exchange</b>	<b>Notes, Messages and Forms</b>	<b>Sociolinguistic</b>	<b>Pragmatic</b>	<b>Linguistic</b>

*Form A23: Graphic Profile of the Relationship of the Examination to CEFR Levels (Example)*

On the chart, the Y-axis (vertical, on the left) represents the CEFR levels. On the X-axis overall language proficiency and communicative language activities and aspects of language competence should be represented. Each column should be labelled with relevant categories from the CEFR. The cells of the chart that are covered by the examination in question should be shaded. If the examination requires a higher level in some categories, this is to be shown with shading, as in the example Form A23 above.

The labelling of the columns on Form A23 will not necessarily be the same as the names given to the subtests of the examination. Some columns may coincide with subtests, but other columns may also be added. For example, the examination might not have a separate subtest for linguistic competence, but the examination provider may wish to indicate to users the level of linguistic competence required.

The emphasis in the procedures presented in this chapter lies on both *process* and *outcome*. Users are encouraged to go through a process of content analysing and linking. It is strongly advised to reconsider every interim claim that has been made during the process. It is quite possible that the initial estimation of the relationship to the CEFR that was given in Form A8 will need to be revised. The user should revisit the analysis and make a considered judgment. The estimation (Form A8) is confirmed or revised in Form A24.

The following chapters of this Manual provide instruments to provide further evidence for the claim. Further research and in-depth analysis at later stages may cause a change to the claims made here. So the accuracy of the claim is subject to an extended process of verification that builds an argument. Examination providers are urged to involve colleagues in a process of discussion and interaction when completing the process.

Confirmed Estimation of Overall CEFR Level		
<input type="checkbox"/> A1	<input type="checkbox"/> B1	<input type="checkbox"/> C1
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> A2	<input type="checkbox"/> B2	<input type="checkbox"/> C2
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<p><b>Short rationale, reference to documentation. If this form presents a different conclusion to the initial estimation in Form A8, please comment on the principal reasons for the revised view.</b></p>		

**Form A24: Confirmed Estimation of Overall Examination Level**

*Users of the Manual may wish to consider:*

- *whether information or data needs to be collected and/or analysed before embarking on the Specification stage*
- *whether to use the CEFR Content Analysis Grids*
- *whether all examinations/tests are appropriate for CEFR-linking*
- *whether completing the Specification stage suggests any changes in the initial planning in the use of the Manual*
- *whether the experience of completing the Specification phase suggests changes in the existing test that might be taken into account at the next planned reform*
- *how they will conclude that Specification has been completed successfully*

## **Chapter 5**

### **Standardisation Training and Benchmarking**

#### **5.1. Introduction**

#### **5.2. The Need for Training**

#### **5.3. Advance Planning**

#### **5.4. Running the Sessions**

##### **5.4.1. Achieving and Verifying Consensus**

#### **5.5. Training with Oral and Written Performances**

##### **5.5.1. Spoken Performance**

##### **5.5.2. Written Performance**

#### **5.6. Training with Tasks and Items for Reading, Listening and Linguistic Competences**

##### **5.6.1. Familiarisation Required**

##### **5.6.2. Training for Standard Setting**

#### **5.7. From Training to Benchmarking**

##### **5.7.1. Samples Required**

##### **5.7.2. Achieving and Verifying Consensus**

##### **5.7.3. Data Analysis**

##### **5.7.4. Documentation**

## 5.1. Introduction

The purpose of the linking process is to enable a categorisation of test takers in terms of the proficiency levels of the CEFR, in such a way that this categorisation reflects in a truthful way what is meant by the CEFR. If a student is categorised as B1, one has to be quite sure that this student is well characterised by the “Can Do” descriptors for this level. This is the basic question of validity and the procedures to follow are referred to as standard setting (see Section B in the Reference Supplement to this Manual).

The way that levels are assigned to test takers falls roughly in two broad classes. Either the categorisation is based on a single holistic judgment by the teacher or examiner, or the test performance results in a numerical score. The former appears mainly with productive skills, while the latter is the common situation for receptive skills. The distinction, however, is not that clear-cut. In a writing examination two or three tasks might be given, and each task can be scored on a number of analytical criteria. The sum of the obtained scores by a test taker can then in principle be treated in the same way as a score on a reading test with a number of separate items. To avoid misunderstandings about this, the two cases will be referred to as indirect test (tests-with-a-numerical-score) and direct tests (holistically-rated-tests), respectively.

- **Direct Tests:** In holistically rated tests, the judgment on the level (here the six CEFR levels) is direct, and therefore it is important to assist raters in giving valid judgments. The main tool used for this special type of standard setting is called **benchmarking**. Benchmarking involves providing one (or more) typical sample(s) to illustrate performance at a given level both for standardisation training and to serve as a point of reference in making future decisions about performances of candidates.
- **Indirect Tests:** For tests with a numerical score, performance standards have to be set. A performance standard is the boundary between two levels on the continuum scale reported by a test that is represented by a “cut-off score”. A cut-off score of 30, for example, says that a numerical score of 30 or more on the tests grants a level of a particular level (e.g. B1) or higher, while a lower score points to a level lower than the level of the cut-off score (here: B1). The process to arrive at a cut-off score is commonly referred to as **standard setting**. In the case of receptive skills (reading and listening) or underlying competences (grammar, vocabulary), cut-off scores need to be decided upon.

Both **benchmarking** and **standard setting** are procedures which require group decisions, which in turn have to be carefully prepared by appropriate training. The main purpose of the present chapter is to give guidance for this training.

As benchmarking is a natural end product of training, it is included in the present chapter.

Standard setting, on the other hand, is a complex, widely discussed and in many respects a controversial topic, with a lot of literature. The procedures for standard setting are therefore discussed separately in Chapter 6. The coordinator can decide on the standard setting method(s) that best suits their context or purpose from the range described in Chapter 6, the Reference Supplement and the extensive literature on the subject.

Nevertheless, although the exact procedures to follow will depend on the standard setting method(s) selected, in the majority of the cases they will be similar to those described in the following sections of this chapter.

## 5.2. The Need for Training

The literature on standard setting includes numerous references to the importance of the panel that recommends the cut-off score(s) or performance standard, and discusses at length issues related to: how such a panel should be formed; how many panellists (= judges) it should involve; what background, skills, subject knowledge and expertise panellists should have; how, when and for how long they should be trained.

Detailed and useful information on how to plan and schedule activities prior and related to standard setting procedures is provided by Kaftandjieva in Section B of the Reference Supplement to this Manual (2004), Hambleton and Pitoniak (2006), and Cizek and Bunch (2007).

The aim of this section is to describe a series of procedures:

- (a) to help panellists to implement a common understanding of the CEFR levels;
- (b) to verify that such a common understanding is achieved, and
- (c) to maintain that standard over time.

The guidelines which follow draw on the experience collected from the reports of applications of different approaches and procedures in the piloting of this Manual, and on the consultation of the literature available.

Standardisation training in relation to the CEFR levels involves four steps:

- carrying out the CEFR familiarisation activities described in Chapter 3;
- working with exemplar performances and test tasks to achieve an adequate understanding of the CEFR levels;
- developing an ability to relate local test tasks and performances to those levels;
- ensuring that this understanding is shared by all parties involved and is implemented in a consistent fashion.

Before starting the training, the appointed facilitator(s)/coordinator(s) (henceforth “coordinator”) should read carefully the present Manual and then follow up the recommended literature references that are considered relevant in the context.

In order to help visualise the work that training entails, a Summary Table (Table 5.5) is provided at the end of this chapter. Table 5.5 can be used by institutions in order to estimate the amount of resources that need to be set apart for the whole process. The table may also be useful for coordinators, who can use it as a working checklist to plan and monitor the process.

The order in which the stages of the process are presented is not random. Training with spoken and written samples of performance – which are rated directly – is easier for the participants than the training with listening and reading items. Listening is the most difficult skill to work on and so should be treated last. Several case studies piloting the Manual showed a considerably higher level of rater agreement and a lower spread of scores for production samples than for reception items. This order is recommended as the most effective, but it will of course be modified according to the needs and constraints of the context.

More detailed guidelines for planning, including exemplar tables, figures and documents, can be found in Chapter 13: *Scheduling Standard-setting Activities* by Cizek and Bunch (2007).

Once training is completed, and common agreement on the assessment of illustrative samples is considered adequate (maximum spread equal to one and a half levels, e.g. A2+ to B1+), work with

local learner performances can start in order to carry out benchmarking (samples of production) or standard setting (for indirect tests with a numerical score).

### 5.3. Advance Planning

The coordinator is responsible for:

- The rationale to be followed, based upon this Manual and related literature.
- Decisions on what types of expertise should be drawn on and who should be involved in which roles at which stage.
- Decisions on the size and composition of the panel of judges. Twelve to fifteen judges can be considered a minimum, and experience in piloting this Manual and other standard setting projects suggests that it is a good idea to include panellists external to the institution producing the test in question, and experts/stakeholders representing different viewpoints.
- Mobilising for the judging panel(s) local professionals with particular experience in:
  - working with the CEFR;
  - producing syllabus and test specifications;
  - assessing productive skills in relation to defined criteria;
  - language test development and item writing;
  - coordinating and training groups of teachers or examiners.
- Obtaining copies of CEFR illustrative samples, plus their related documentation.
- The brief for collecting, to a locally defined standard format, the materials which will be used:
  - the local scripts of students' writing and video recordings of students' spoken performances that will be used to benchmark local performances to the CEFR illustrative performance samples and the CEFR itself;
  - the local test tasks that will be worked on in the judgment sessions.
- The decision whether to use the CEFR "plus levels" or not. Calibrated descriptors are available for levels A2+, B1+ and B2+.
- The preparation, development and photocopying of the materials to be used in the different stages of the process (see Table 5.5 for detail):
  - CEFR descriptors;
  - CEFR tables and rating instruments (e.g. CEFR Table 3 – Manual Table C2<sup>7</sup>);
  - selection of illustrative CEFR performance samples and tasks;<sup>8</sup>
  - selection of local performance samples and / or local test items;
  - reporting forms and documents to record information on the sessions.

---

<sup>7</sup> CEFR Table 3: Common Reference Levels: Qualitative aspects of spoken language use. (English: pages 28–29; French: page 28.)

<sup>8</sup> Please consult the up-to-date list of available materials on [www.coe.int/portfolio](http://www.coe.int/portfolio) At the time of writing, spoken and written samples for adult learners are available for English, French, German and Italian, with Spanish being planned to follow. A second CD of test tasks and items is currently being prepared; this CD includes a wider range of materials originating from case studies in piloting the Manual. A 2008 project in cross-linguistic benchmarking of spoken samples from French 16–18 year olds will later produce DVDs showing performance in English, French, German, Spanish and Italian.

- Checking that enough rooms are available to allow for group work and that all facilities needed are available – including tables or desks if working with writing samples of booklets of reading and listening items.
- The collection and analysis of data from the training sessions, presentation and copying of relevant results (e.g. empirical difficulty values of items; ratings of other groups with samples) so as to feed these into the sessions if and when appropriate.
- The organisation of the training sessions themselves in a way best adapted to the local context. The coordinator will have to decide on the number of participants per session as well as on the best timing and organisation. This includes:
  - deciding on who is to be invited (teachers/examiners/item writers) to which sessions and whether preparation for the sessions needs to vary according to the audience concerned;
  - ensuring the right atmosphere and appropriate grouping;
  - planning enough time (see below) to provide opportunities for extensive and in-depth reflection and discussion, which will contribute to achieving consensus in judgments;
  - summarising conclusions.
- The organisation of the documentation and reporting of work done at the training sessions, in order to give accountability, and to provide support for dissemination sessions and follow up sessions.
- The planning of continuous verification and on-going monitoring, dissemination and follow up actions.

**Time Required:** The time required will depend on:

- the expertise of the participants from attending previous rater training sessions;
- whether they are already familiar with the use of rating scales;
- their experience in item writing and in estimating item/task difficulty;
- whether pre-session familiarisation and practice e.g. using the “Dutch” CEFR Grid has been arranged.

With experienced participants it may be possible to complete the training for productive skills in one day, devoting the morning session to speaking and the afternoon session to writing. Then one can proceed to working with local performance samples on the following day. Alternatively the first day might be devoted exclusively to training and standardising activities with spoken performances, and the following day to written scripts. On each day, work should start with standardised exemplar performances in the morning and then proceed to local samples in the afternoon.

The time required for training for receptive skills will depend not only on how familiar the participants are with the process of scoring, selecting and writing test items and test tasks, how much concrete feedback they have received on item/task difficulties but also on the number of skills to be assessed. A similar pattern to the one described above for speaking or writing can be followed for each skill. If the first receptive skill – as recommended in this Manual – is reading, training with illustrative test items may take place in the morning and be followed in the afternoon by the judgment of local test items.

## 5.4. Running the Sessions

The training should take place in working sessions in which participants are made familiar with the CEFR, analyse and assess performances or test items and reach a consensus in terms of assigning them to a CEFR level.

During the sessions, the appointed coordinator(s) are responsible for:

- Checking that participants achieve a good background understanding of what the CEFR means and the extent to which they are aware of how the CEFR can contribute to improve their work. The Familiarisation activities in Chapter 3 should be used for this purpose.
- Ensuring, when rating performance samples, that a logical progression is followed in order to reach and reinforce consensus:
  - lead in and illustration;
  - individual rating;
  - small group rating;
  - whole group discussion.
- Collecting information and giving feedback throughout, as clearly and graphically as possible.
- Checking that an adequate consensus in the interpretation of the CEFR levels, as defined in the instructions, has been reached in terms of both the CEFR descriptors themselves and also in terms of performances or test tasks that operationalise them.

After the training, the appointed coordinators are responsible for ensuring that all necessary materials are available to all the members of the panel before the Benchmarking/Standard setting process starts.

### 5.4.1. Achieving and Verifying Consensus

Throughout each session coordinators should invite comments and discussion and summarise judgments in the way considered most appropriate within the context, in order to reach a reliable consensus.

It should be remembered that, as in any assessor training session, asking trainees to estimate the level of an already standardised sample is an exercise with a right answer. The correct answer is released only at a later stage by the coordinator. Unlike in the benchmarking or standard setting activities that follow, at this stage the group is not being invited to form a consensus on the level of the sample irrespective of previous evidence – but rather to arrive at the pre-established correct answer by applying the criteria.

This requires a certain skill on the part of the coordinator (a) to steer the group towards the right answer in these important initial experiences, and (b) to avoid publicly exposing participants who are too strict or too lenient in their interpretation before they have had a chance to tune in with the training – since this may upset them and destabilise their later judgments. The amount of time that this process takes should not be underestimated. It is essential to invest the necessary time for training before moving on to working with local samples.

There are two schools of thought as to how to steer the group to the right consensus.

**Sensitive Approach:** The first school suggests a sensitive approach that avoids embarrassing participants by maintaining anonymity of rating. This approach also ensures that participants record their individual judgment before discussion, are not “bullied”, and that the consensus which gradually

emerges is a genuine one. With this approach the individual is influenced by the other ratings: if a participant is an “outlier”, he/she sees this and may shift towards the mean.

- Rating slips that are passed around to the coordinator without comment safeguard privacy. In order to trace the panellists in data collection for possible later analysis, nicknames (e.g. Mickey Mouse) or numerical IDs pre-printed on the paper can be used. Swift collation of the anonymous slips onto an overhead projector or flip chart exposes but does not identify and embarrass “outliers” – unless they choose to argue!
- Electronic voting can be used to the same effect. The benchmarking seminars that produced the French, German, Italian and Portuguese DVDs used this approach. There are two rounds of voting: individual voting before discussion; voting to confirm the consensus after the discussion.

**Robust Approach:** The second school of thought takes a more robust approach: differences of opinion need to be expressed and discussed if a proper consensus is to be reached. Here the consensus is more conscious, as a result of argument – which may be swayed by an articulate speaker. For this reason, it is a good idea if the coordinator ensures that the participants are familiar with the standardised samples, and the reasons why a particular sample is a certain level, and the way this relates to the descriptors.

Working in pairs or small groups is something participants usually find very enjoyable. The coordinator can circulate and listen in to the discussion, where necessary steer a group in the right direction, and ask for a report back from a member of each group. The main advantage of group or pair work is that it naturally forces the participants to use the defined criteria to justify their judgments. Tallying the results, with the coordinator completing a grid on a flip chart or overhead transparency, is a simple way of recording results.

Whichever of the two approaches is chosen, the coordinator will need to calculate the percentage of participants who agree on the different ratings, or inter-rater correlation coefficients. The coordinator will need to decide whether, on this particular occasion, to share the figure with participants, if he or she thinks this will contribute to training and an increased convergence in judgment.

It is also a good idea to give a graphic presentation of the spread of ratings. Bar charts are produced easily with electronic voting. An alternative way to do this is by entering ratings into the data source for previously designed histogram in Microsoft Excel. A third method is to use the box plots produced by the test analysis program SPSS.

## 5.5. Training with Oral and Written Performances

It may well be that illustrative performance samples and/or test tasks are not yet available for the language concerned. In that case we recommend working with samples for a language that the panel has in common – provided panels possess a level of proficiency of this language, minimum B2/C1. If this is the case, it will need to be reported as an indirect training in the documentation.

The process starts with the analysis and assessment of CEFR illustrative performances of spoken performance and continues (if appropriate) with illustrative scripts of written performance. The majority of the illustrative spoken samples follow a similar format which includes a spoken production phase for each candidate (a sustained monologue in which one candidate explains something to the other, who asks questions) followed by an Interaction Phase (in which the two candidates discuss an issue spontaneously)<sup>9</sup>.

---

<sup>9</sup> This format was adopted for the Swiss research project that developed the CEFR descriptor scale and is shown on the initial (Eurocentres/Migros) DVD for English, which includes performances from that project. This approach, which is not a test situation, avoids examiner effects. It has been adopted by the developers of the DVDs for adult learners of French, Italian and Portuguese and for the Council of Europe/CIEP DVDs for teenager learners of English, French, German, Italian and Spanish.

For the assessment of writing, it is also important to see samples of both written interaction (e.g. notes, letters) and written production (e.g. descriptions, stories, reviews) from a candidate. This is particularly important at lower levels.

It is important to note that in the illustrative samples it is the overall proficiency of the candidate deduced from the complete performance that is rated, not the separate performances (monologue/interaction) themselves. The documentation gives a reasoned argument as to why the candidate is one level and not another level, with explicit citation of the CEFR criteria (CEFR Table 3/Table C2 for spoken performance; Table B4 for written performance). That is to say, the assessment tasks are designed to generate representative, complementary samples of the candidates' ability to perform orally in the language. On the basis of *all* of the evidence available, the panellist uses the generic criterion descriptors (CEFR Table 3/Table C2) to make a judgment of the competence of the candidate in as much as this can be deduced from the inevitably limited and imperfect sampling. The result – the competence glimpsed through the performance – is conventionally referred to in English as “proficiency”.

### 5.5.1. Spoken Performance

For this session it is essential that participants use an assessment grid made up of CEFR descriptors, such as those provided in Appendix B. We strongly recommend use of CEFR Table 3<sup>10</sup> (given as Table C2). In addition, panellists may find useful:

- a simplified, holistic assessment scale derived from CEFR Table 3 (Table C1);
- if “plus levels” are being employed, copies of the supplementary Grid based on CEFR Table 3 (Table C3);
- CEFR descriptor scales for Overall Interaction and Overall Production;
- the CEFR scale for Phonological Control, in case they are considered relevant<sup>11</sup>;
- a standard rating form to note their comments and assigned level for each performance (see Forms C2 and C3 as examples).

The session is organised in three phases:

**Phase 1: Illustration:** The session starts with two or three CEFR illustrative performances that the coordinator uses to introduce the levels. The coordinator plays the sample and then invites participants to discuss the performance with neighbours. At an appropriate point the coordinator should bring the group together, and elicit from the group the way in which the performance illustrates the level described on the CEFR Table 3 (Table C2) Grid, and why it is not the level described above or below.

It is best to play the whole recording of the sample, even though this may take 15 minutes. A candidate's performance in the Interaction Phase may be significantly different (better or worse) than performance in the Production Phase and – as mentioned in the introduction, it is the candidate's overall proficiency in the skill concerned that is to be rated - not one of their performances.

**Selection of Samples:** The following advice is based on experience in piloting the Manual, running the benchmarking sessions that produced DVDs of illustrative samples, and related projects.

---

<sup>10</sup> CEFR Table 3: Common Reference Levels: Qualitative aspects of spoken language use. English: pages 28–29; French: page 28.

<sup>11</sup> Pronunciation is not included in CEFR Table 3 because it is designed for use in international contexts and raters accustomed to working in a monolingual, national context can tend to be over-influenced by their lack of familiarity with the accents of speakers of other mother tongues.

- It is a good idea to start with Levels B1 or B2 and to show samples of performance at adjacent levels in order to encourage discussion of boundaries between levels, by referring to the criteria (CEFR Table 3/Manual Table C2).
- The first of these illustrative examples should show a performance with a relatively “flat profile” across the categories of CEFR Table 3/Manual Table C2 – namely a speaker who is, for example, B1, on all the categories Range, Accuracy, Fluency, Coherence, Interaction.
- One of these introductory standardised samples should show a more uneven profile, e.g. when the speaker is B1 for some categories but B2 or at least B1+ for others. Unless the issue of “uneven profiles” is discussed early in the training, it may become a complication later on.

In order to highlight the fact that some candidates may have very uneven profiles and that the different qualitative aspects (Range, Accuracy, Fluency, Coherence, Interaction) should be considered separately, coordinators may wish to consider rating several performances for *just one aspect*. This counteracts the panellists’ natural tendency to allow their overall impression to influence their judgments on each category (“halo effect”).

**Use of Rating Instruments:** The following advice is again based on experience in piloting the Manual, running the benchmarking sessions that produced DVDs of illustrative samples, and related projects.

- Participants may be asked to first use only the holistic scale (Table C1) that simplifies the CEFR Table 3 Grid (Table C2) in order to become consciously aware of their global impression of the candidates’ level, before they consider the categories in the CEFR Table 3 (Table C2) Grid.
- Having formed an initial impression of the level of the performance, they should then consult the more detailed descriptors for that level on the CEFR Table 3 (Table C2) Grid, read the descriptors for the level above and below for each category, and use the Grid to profile the candidates’ performance.
- If “plus levels” are being used, they should consult the supplementary Grid (Table B3) at this point to decide if the candidate is a “strong” example of the level – a “plus level” performance.
- They should then use the descriptors on the CEFR Table 3 Grid (Table C2) and if appropriate supplementary plus levels Grid (Table C3) to guide their discussion with their neighbour.
- During this discussion, they may wish also to consult the supplementary descriptor scales mentioned above.

**Phase 2: Practice:** In a second phase the role of the coordinator is to help individuals see if they are still tending to be too strict or too lenient. If voting is on paper, the coordinator will use the collation form (e.g. Form B3) to record the ratings onto a transparency or on a flip chart. Throughout this phase, the coordinator should graphically show the participants their behaviour as a group and monitor the discussion as discussed above, without embarrassing individuals. If no form of anonymous voting is being used, an effective technique here is to listen in to the group discussions, and when bringing the whole group together, to elicit “the answer” from groups most likely to get it right

It is good practice for the coordinator to lead a discussion in the whole group as to *why* the candidate is one level rather than the level above or the level below, with explicit citation of the criterion descriptors. This helps to prevent participants slipping back to pre-conceived notions of CEFR levels (often merely translated from another system) and makes it clear that the criterion descriptors are the sole point of reference.

**Selection of Samples:** Again the use of two to three samples is recommended.

**Use of Rating Instruments:** Coordinators should decide in advance whether to continue to use the global scale (Table B1) after the Illustration Phase. It is useful in that (a) it gives the participant a place to start reading on the Grid (CEFR Table 3; Manual Table C2), and in that (b) it helps the participant separate their initial impression from a considered judgment – especially if the two are recorded separately as in the record form given as Form C2.

**Phase 3: Individual Assessment:** The participants rate the rest of the performances individually, hand in their rating slips, and then discuss the CEFR levels these performances have been assessed to represent. It is recommended to continue to analyse performances in chunks of three performances. In this way the discussion will then be more easily focused on standardisation – rather than detailed discussion of the merits of certain performances. The last chunk should show good agreement. That is to say, the vast majority of participants should agree on the level, with the spread not exceeding one and a half levels. For example, for a performance generally agreed to be B1+, the spread of results should not exceed the range B1 to B2; for a performance agreed to B1, the spread should not be more than A2+ to B1+.

The session can end when this degree of agreement within the group is reached and the coordinator (and the participants) are satisfied with the degree of consensus in assessing standardised samples of oral performance.

Again the use of two to three samples is recommended.

Coordinators should decide in advance whether to continue to use the global scale (Table C1) after the Illustration Phase. It is useful in that (a) it gives the participant a place to start reading on the Grid (CEFR Table 3; Table C2), and in that (b) it helps the participants to separate their initial impression from a considered judgment – especially if the two are recorded separately as in the record form given as Form B2. However, it may be simpler to eliminate one of the pieces of paper panellists are working with. Experience shows that once panellists are accustomed to using CEFR Table 3 (Table C2), they do not really need the scale (Table C1) to arrive at an initial global impression.

**Selection of Samples:** It is recommended that at least one performance per CEFR level is analysed, assessed and discussed in the whole session.

**Use of Rating Instruments:** During the discussions, in order to contribute to a better understanding of the level, the coordinator will decide whether it is relevant to use further CEFR speaking scales and justify in more detail the level assignment.

### 5.5.2. Written Performance

A process parallel to that recommended for spoken performances is recommended.

The Assessment Grid to refer to is Table C4 in Section C of the Appendix. This Grid is an extension of CEFR Table 3, adding two columns on Description and on Argument that should only be used for those particular text types.

**Phase 1: Illustration:** The session starts with two or three written performances that the coordinator uses to illustrate the levels. For each sample, at a certain point the coordinator should bring the group together, and elicit from the group the way in which the performance illustrates the level described on the Table C4 Grid, and why it is not the level described above or below.

**Table 5.1: Time Management for Assessing Oral Performance Samples**

<i>Group size recommended: maximum of 30 participants</i>	
<b>Stage 1: Familiarisation</b>	<i>60 minutes</i>
<b>Stage 2: Working with Standardised Samples:</b>	
<i>Phase 1: Illustration with circa three standardised performances.</i>	<i>60 minutes</i>
<i>Break</i>	
<i>Phase 2: Controlled Practice with circa three standardised performances.</i>	<i>60 minutes</i>
<i>Phase 3: Free Stage with circa three standardised performances.</i>	<i>60 minutes</i>
<b>Lunch</b>	
<b>Stage 3: Benchmarking Local Samples:</b>	
<i>Individual rating and group discussion of circa three performances.</i>	<i>60 minutes</i>
<i>Individual rating of circa five more performances.</i>	<i>60 minutes</i>
<i>Break</i>	
<i>Planning follow-up activities and networking.</i>	<i>60 minutes</i>
<i>Summing up, closure.</i>	<i>30 minutes</i>
<b>Documents and tools to be prepared</b>	
<i>Photocopies for all participants:</i>	
<ul style="list-style-type: none"> <li>• <i>Assessment Grid CEFR Table 3/Manual Table C2.</i></li> <li>• <i>Assessment scale: simplifying the above: Table C1 (if considered necessary).</i></li> <li>• <i>“Plus Level” Grid: supplementing the above: Table C3 (if considered necessary).</i></li> <li>• <i>Rating sheets for participants: examples as Forms C2–C3.</i></li> <li>• <i>Selection of and copies of the relevant complementary scales or of Tables A1–A3.</i></li> </ul>	
<i>Plus:</i>	
<ul style="list-style-type: none"> <li>• <i>Standardised videos of performances.</i></li> <li>• <i>Manual.</i></li> <li>• <i>Collation forms for coordinator and transparency (Form B4).</i></li> <li>• <i>Local videos (to be recorded and/or selected according to the brief for Case Studies).</i></li> </ul>	

**Selection of Samples:**

- The first of these illustrative examples should show a performance with a relatively “flat profile” across the categories of Table B4 (namely a writer who is, for example, B1, on all three categories Range, Coherence and Accuracy, and equally good at description and argument).
- As with samples of spoken performance, the coordinator may consider rating some scripts with just one category in order to make participants aware of the “halo effect”.
- It is recommended that one of these introductory samples shows a more uneven profile, e.g. when the writer is B1 for some categories but B2 or at least B1+ for others. Unless the issue of “uneven profiles” is discussed early in the training, it may become a complication later on.

**Use of Rating Instruments:**

- The coordinator instructs the participants to read the script, and consider the performance in relation to the criteria in Table C4.

**Phase 2: Practice:** In this second phase – again using about three samples – the role of the coordinator is to help individuals see if they are still tending to be too strict or too lenient. If voting is on paper, the coordinator will use the collation form (e.g. Form C3) to record the ratings onto a transparency.

Throughout this phase, the coordinator should graphically show the participants their behaviour as a group and monitor the discussion as discussed above, without embarrassing individuals. If no form of anonymous voting is being used, an effective technique here is to listen in to the group discussions, and when bringing the whole group together, to elicit “the answer” from groups most likely to get it right.

**Phase 3: Individual Assessment:** The participants rate the rest of the performances individually and discuss the CEFR levels these performances have been standardised to.

It is recommended to continue to analyse performances in chunks of three performances. In this way, the discussion will then be more easily focused on standardisation – rather than detailed discussion of the merits of certain performances. The last chunk should show good agreement. That is to say, the great majority of participants should agree on the level, with the spread not exceeding one and a half levels. For example, for a performance generally agreed to be B1+, the spread of results should not exceed the range B1 to B2; for a performance agreed to B1, the spread should not be more than A2+ to B1+.

The session can end when this degree of agreement within the group is reached.

***Selection of Samples:***

- As for spoken performance, it is recommended that at least one performance per CEFR level is analysed, assessed and discussed in the whole session.

***Use of Rating Instruments:***

- As during the discussion on oral production and interaction samples, the coordinator may decide to use specific scales (e.g. Overall written production, Creative writing, Reports and Essays) to help reach agreement and better justify level assignments. Alternatively, coordinators might distribute Tables A2 and A3, used in conjunction with Chapter 4 on Specification.

**Table 5.2: Time Management for Assessing Written Performance Samples**

<i>Group size recommended: maximum of 30 participants</i>	
<i>Introductory tasks (Familiarisation)</i>	<i>60 minutes</i>
<b><i>Working with Standardised Samples:</i></b>	
<i>Phase 1: Illustration with circa three illustrative performances</i>	<i>60 minutes</i>
<i>Break</i>	
<i>Phase 2: Controlled practice with circa three–five illustrative performances</i>	<i>60 minutes</i>
<i>Phase 3: Free Stage with circa three–five illustrative performances</i>	<i>60 minutes</i>
<b><i>Lunch</i></b>	
<b><i>Benchmarking Local Samples:</i></b>	
<i>Individual rating and group discussion of high, middle and low performances</i>	<i>60 minutes</i>
<i>Individual rating of circa five more performances</i>	<i>60 minutes</i>

**Table 5.3: Documents and Tools to be Prepared for Rating Writing**

***Documents and tools to be prepared***

*Photocopies for all participants:*

- *Assessment Grid (Table C4)*
- *Rating sheets for participants: (Forms C2–C3 give examples)*
- *Selection of and copies of the relevant complementary scales*

*Plus:*

- *Standardised scripts*
- *Collation forms for coordinator and transparency (Form C4)*
- *Local scripts to be selected according to the brief for Case Studies)*

### **5.6. Training with Tasks and Items for Reading, Listening and Linguistic Competences**

The objective of the activities described in this section is to ensure that panellists can relate their interpretation of the CEFR levels to exemplar test items and tasks so that they can later build from this common understanding in order to:

- relate locally relevant test items to the CEFR levels;
- as added value, gain insights into developing test items that can eventually claim to be related to CEFR levels.

The techniques described can be used for test items and test tasks used to evaluate receptive skills and – where appropriate – to evaluate other aspects of language use, such as grammar and vocabulary.

Tasks which involve integrated skills (e.g. listening to a text and answering questions, and then using the information gained to make a summary) will need to be considered from the point of view of the difficulty of both the receptive and productive aspects of the task. There may be a deliberate difference in the difficulty level of the two parts of the task, and this needs to be addressed in training. Item difficulty may vary (and be varied systematically, if one so wishes) depending on the read or heard text, on the comprehension ability tested and on the response that the test taker needs to make to indicate comprehension.

As with performance samples, training with illustrative tasks and items with known difficulty values should take place first and then be followed by the process of analysing locally produced items (Chapter 6).

Training with illustrative test tasks and items includes, in this order:

1. Becoming fully aware of the range of CEFR subscales of descriptors for specific areas that are available in the CEFR (see Chapter 4).
2. Identifying the content relevance of the tasks analysed in terms of construct coverage vis-à-vis CEFR levels and scales. As mentioned in Section 4.3.2, the findings in the Dutch CEFR construct project (Alderson et al 2006<sup>12</sup>), and the resulting CEFR Content Analysis Grid for Listening & Reading<sup>13</sup> may be very useful.
3. Estimating the level each task and item represents in terms of the relevant CEFR descriptors.

<sup>12</sup> Alderson, J.C. et al (2006) Analysing tests of reading and listening in relation to the CEFR. 3(1) 3–30. Language Assessment Quarterly.

<sup>13</sup> Paper version in Appendix B1. Electronic version freely available on-line, with a training module, at [www.lancs.ac.uk/fss/projects/grid](http://www.lancs.ac.uk/fss/projects/grid)

4. Discussing the possible reasons for discrepancies between estimated and empirically established levels.
5. Confirming the level of difficulty against empirical data.

It is essential to start with the skill of reading. In the same way that it is easier to work on spoken and written performance (which can be observed directly) than to work on receptive skills (which cannot be observed), it is far easier to work on reading and rereading texts and items in print (that can be seen) than it is to work on listening items and texts (which cannot be seen) in several rounds of listening.

Once the process of assessing items for reading has been completed, organising the session for the skill of listening and working with listening texts will be easier, as the participants will already be familiar with the task to be done. The coordinator needs to decide how to organise the sessions and to estimate the duration of the sessions, depending on the context and the background of the participants.

### **5.6.1. Familiarisation Required**

Even if participants have already attended a general Familiarisation session described in Chapter 3, a sorting exercise with descriptors for the skill concerned before starting difficulty estimation and standard setting is a necessary training exercise.

The CEFR provides overall, general scales (e.g. “Reception”, “Overall Reading Comprehension”, “Overall Listening Comprehension”), and also specific scales that describe different receptive language activities (e.g. “Listening as a Member of an Audience”) and strategies (“Identifying Cues and Inferring”).

Coordinators need to decide on the most relevant scales for the examination in the context in which it is administered. Work should always start with analysis and discussion of overall scales (e.g. “Overall Reading Comprehension”). Then the coordinators may pool the most context relevant subscales for the skill concerned (e.g. “Listening as a Member of an Audience), or use the self-assessment reformulations of the CEFR descriptors employed in the DIALANG project (CEFR Appendix C), and ask participants to sort the descriptors into the six CEFR levels (see Section 3.2.1 Activity f).

Standardisation of items testing linguistic competences will need to take a slightly different approach to the one followed with reading and listening because of the need for a specification of the type of exponents that can be expected to be relevant to different levels. The CEFR provides general descriptors for elements of communicative language competence (CEFR Section 5.2; Manual tables A1–A3), but such linguistic specifications are unique to each language. Section 4.3 outlines the tools currently available. The DIALANG project also developed a set of specifications, with advice to item writers, for 14 languages.

### **5.6.2. Training for Standard Setting**

The standardisation process follows three phases similar to those training procedures employed with standardised performance samples:

**Phase 1: Illustration:** A first assessment of the level of one text and its corresponding tasks and items. This preliminary activity will help the participants tune into the CEFR levels for the skill being assessed.

It is essential to consider both the question of the **level of the source text** and the **difficulty of the individual item(s)** associated with it. A text does not have a “level”. It is the competence of the test takers as demonstrated by their responses to the items that can be related to a CEFR level. The most that can be said about a text is that it is suitable for inclusion in a test aimed at a particular level.

**Table 5.4: Reference Sources in the CEFR**

Area	CEFR Reference
Situations, content categories, domains	Table 5 in CEFR 4.1
Communication themes	The lists in CEFR 4.2
Communicative tasks	The lists in CEFR 4.3
Communicative activities and strategies	The lists in CEFR 4.4.2.2
Texts and text-types	The lists in CEFR 4.6.2 and 4.6.3
Text characteristics: length of test tasks, coherence of test tasks, structure of test tasks	The information in CEFR 7.3.2.2
Tasks	The description in CEFR 7.1, 7.2 and 7.3

In this respect, the CEFR Content Analysis Grid for Listening & Reading, described in the preceding chapter, can be very useful as an awareness raising instrument to highlight the features affecting level of difficulty.

Users will find it useful to refer to the relevant completed forms from Chapter 4 Specification, and to consider text and task difficulty in relation to the appropriate sections of the CEFR. For Reading Comprehension, for example, the form to use is Form A10 and the sections of the CEFR referred to are given in Table 5.4.

This task is to be done first individually and the coordinator will, as in relation to the work with learner performances discussed earlier in the chapter, raise awareness of agreement or disagreement across judges. The following points have been found particularly important:

- It is very important that participants actually read or listen to the text and answer the item/s individually before they estimate the difficulty of the question concerned and the CEFR level it best illustrates.
- After responding to the item(s), they should be able to compare their own response to the correct answer (and to the quality categories in the scoring rubric of polytomous items) for the item(s) concerned. Discussion to ensure clear understanding of the answer key or the scoring rubric should precede participants’ estimation of the item difficulty.
- It is also vital that the coordinator gives clear instructions in the form of the precise instruction that participants receive. The item is conceived as an operationalisation of a CEFR “Can Do” descriptor. Therefore the question is what level the learner has to be in order to be able to answer this question correctly – or reasonably well.

The precise instruction judges receive will depend upon the standard setting method being applied. The following example refers to the Basket method (Section 6.7.2):

*For items scored 1–0 (dichotomous items):*

*“At what CEFR level can a test taker already answer the following item correctly?”*

*For polytomous items:*

*“At what CEFR level can a test taker already answer the following item at score levels xxx (e.g. 2, 1, 0)?”*

- Participants individually note their ratings for the items, and then in pairs or small groups justify their decisions.
- Finally, the coordinator then provides “the” level that the item(s) really are calibrated to.

**Phase 2: Controlled Practice:** Once the illustration phase and the initial discussion have taken place and a common feel of the process to be followed has been achieved, different texts with their corresponding tasks and items will be assessed by participants, individually, relating them to CEFR levels and identifying the CEFR descriptors operationalised by each item/task.

As with the rating of spoken and written samples, it is a good idea to proceed with 4–6 items, or two or three testlets (a text with more than one item). Participants should be asked to:

- read the texts and answer their corresponding items;

and then complete a Grid (see below), providing their assessment of each item, in order to:

- identify the CEFR descriptors it operationalises;
- classify each item at one of the six CEFR levels.

Group discussion should take into account the following aspects:

- the type of item (selected response, constructed response) and how this may affect the difficulty of the item;
- the operationalisation of different CEFR descriptors in the text and task;
- the available evidence justifying the calibration of each item to its corresponding CEFR level;
- other relevant aspects for the text/item/response characteristics that participants have included under the “comments” column.

In this respect, it should be noted that panellists may tend to overestimate the difficulty of selected-answer items (e.g. multiple-choice), which tend to be easier than panellists often think. By the same token they tend to underestimate the difficulty of constructed-answer items (e.g. answer a question, complete the sentence), which tend to be more difficult than panellists think. Asking participants to actually respond to the items before embarking on difficulty discussions can go some way to reducing this problem. However, focusing on the interaction between text and item-type in determining difficulty – with regard to operationalisation of a CEFR descriptor – is necessary sensitisation training at this stage.

It may be useful to draw the panellists’ attention to the role of the complexity of the language, the length of the passage one needs to scan to find the correct answer, the plausibility of multiple-choice options etc. as factors contributing to item difficulty. Again, coordinators should invite comments and discussion, and summarise clearly and graphically the judgments, not only for the participants to see but also for future documentation.

**Phase 3: Individual Assessment:** The participants continue to work with the rest of the items individually and discuss the CEFR levels the items have been calibrated to. As with the spoken and written performances, it is recommended to proceed with chunks of 4–6 items. In this way the discussion will be more easily focused on standardisation rather than on the properties of the items or the different texts. The last chunk should show good agreement.

As with the performance samples, it is recommended that participants should continue working in the same fashion (using the Grid to write down their assessments) until a spread of results of not more than one and a half levels is achieved (e.g. A2+ to B1+).

The coordinator may use a global rating form like Form C4 in order to collate the participants' ratings of the items and to graphically show on a transparency or on a flip chart the variation in their agreement. The contents of this form will be necessary in the documentation.

Once training (Sections 5.4. and 5.5.) is complete and common agreement on the assessment of standardised samples and tasks is considered adequate, work with local samples that have previously been collected can start. The following section (5.6.) provides a step-by-step account of how to proceed for benchmarking local samples of speaking and writing. The procedures to follow are very similar to those followed in the training (5.4.).

As for establishing cut-off scores on locally developed tests for reading, listening or underlying language abilities, the choice of standard setting procedure(s) from those described in Chapter 6 in this Manual (or from other literature on standard setting) will influence the procedures to follow. Users of this Manual should read Chapter 6, decide on one or more methods, and, following the structure of the training described in this section, develop their own context-relevant step-by-step procedures. The extensive literature available will be of great help in drawing up the procedures, but the points described in the following section for benchmarking in relation to sampling/choice of items, data analysis and documentation need to be considered.

## **5.7. From Training to Benchmarking**

The application of the understanding of the CEFR levels to the benchmarking of local samples (of spoken or written performance) or local tasks/items (for scored tests for listening, reading and linguistic competence) should take place as soon as possible after the standardisation training. It is highly recommended that it should take place in the same session, in the afternoon or on a second day. The coordinator will be the best judge of whether this is feasible, or whether it would be better done at a later stage. If the sessions with local samples are delayed, then a “tuning-in” phase is recommended, showing participants extracts from a couple of the standardised performances rated in the previous session, and reminding them of the discussion.

The procedures to follow for benchmarking are similar to those applied in the training.

### **5.7.1. Samples Required**

It is worth investing time and energy in collecting a representative set of local samples of high quality, even if this imposes a delay in the timing of the project. Once they have been benchmarked to the CEFR, such samples are likely to acquire a significant status as points of reference. Therefore it is advisable to make a conscious selection of the items to ensure quality, representativeness (in terms of test takers) and content coverage.

The collection process could be undertaken much in the same way as an item production process:

- definition of the selection criteria;
- identification of candidate samples;
- workshop to study and screen the samples for quality;
- selection;
- verification of sufficient coverage in the set;
- supplementation with other samples, if feasible, to “complete” the set.
- documentation of the features of the samples for benchmarking using a tool like the CEFR Grids for Writing and Speaking Tasks (Appendix B2).

It is essential that the local performance samples to be used for benchmarking include different discourse types for the same candidates, covering a range of the activities described in the CEFR.

For speaking, this suggests an activity with phases eliciting different discourse – spoken production as well as spoken interaction. The technique used for filming the illustrative samples was designed to avoid examiner effects, and to provide a balanced sample of both spoken production and spoken interaction.

For writing, it suggests different text types. It is best if written samples encompass both freer tasks (e.g. a letter to a friend, a description) and more formulaic tasks in which the candidate follows a learnt model (e.g. a letter confirming a hotel booking). This is particularly important at lower levels.

It is vital that during the process of task production, task administration and task recording or documenting, special care is paid to obtaining good, usable samples. In the case of videos this means good sound and image<sup>14</sup>; in the case of scripts it means performances uncontaminated by external influences such as extra time, use of dictionary, poor handwriting, etc.

Completing the CEFR Grids for Writing and Speaking Tasks as suggested in the previous section helps to ensure that the selection of samples is balanced and that the basis for documentation is available.

### 5.7.2. Achieving and Verifying Consensus

In general, the procedures to be followed are those outlined in Sections 5.3 and 5.4 for standardisation training with illustrative samples. This will include:

- using the same rating instruments that were used in training (Tables C1, C2 and possibly C3 (plus levels); Table C4 for written performances; CEFR scales and/or Tables A1, A2 and A3 for receptive/linguistic texts and items);
- individual rating followed by small group discussion leading to group consensus;
- discussion of spread across individual ratings and iteration until suitable agreement (maximum spread equal to one and a half levels), is reached.

Here an important point to emphasise is that the individual ratings must be recorded *before* any discussion. Actually, experience in the benchmarking seminars that produced the illustrative DVDs suggest that it is the spread of ratings that is affected by discussion (as outliers conform to the norm), not the mean and hence result. Nevertheless, it is the mark of a successful benchmarking seminar that aggregated individual judgments and the final consensus should give the same CEFR levels for a sample or item. Demonstration of this with uncontaminated data is part of providing evidence<sup>15</sup>.

If agreement is NOT reached, the coordinator should discuss with participants why they are having such a problem in contrast to their success with the illustrative samples. The coordinator will need to

---

<sup>14</sup> If a video is later to be copied onto a “master”, and that master copied for distribution, then users will have a third generation copy that magnifies any sound defects. For this reason, even with digital DVD technology, it is *always* advisable to use an external microphone and *not* the microphone built into the camera. With an external, omnidirectional limited range (1–2m) microphone it is perfectly possible to get acceptable sound quality without a recording studio.

<sup>15</sup> This is not necessarily the case with standard setting for indirect, scored tests. Because standard setting is an indirect process, in many methods it is conducted in rounds. Later rounds generally introduce information to guide panellists towards less inaccurate judgments – hence aggregated initial individual judgments will not coincide with the final results of a successful standard setting seminar. The information conventionally provided to help standard setting panellists includes empirical item difficulty; projected consequences that cut scores set with the judgments made would have on the percentages of people reaching the level concerned, etc. and other information; please see Chapter 6.

make a judgment on the reason for the problem, and take appropriate action. Some possible reasons, and possible courses of action, might be:

#### **Problem**

- Local samples only have one task and this task is too different from the CEFR samples
- The rating Grid (e.g. Table C2) seems inappropriate for samples (e.g. vocational context, narrowly defined task)
- Some participants start to apply other standards when now rating “their own” learners.

#### **Possible Action**

- ➔ Check that a sufficient range of discourse/text is provided  
Get other samples closer to the CEFR format
- ➔ Revise Grid, consulting CEFR scales to do so
- ➔ Juxtapose CEFR sample and local sample directly to try and “force” people to apply the same CEFR standard.

### **5.7.3. Data Analysis**

Ratings of the local samples that are the subject of the benchmarking should be analysed statistically (a) in order to confirm the relationship to the levels and (b) in order to calculate intra-rater reliability (consistency) and inter-rater reliability (consistency).

The degree of agreement amongst the participants should be assessed, and the mean level of the samples confirmed, by analysing the ratings during the benchmarking process. The main advantage is that panellists who are inconsistent in their behaviour can be identified and they can be excluded from the analysis, if this seems appropriate.

There are several methods that are suitable for this purpose, described in the Reference Supplement to this Manual. In addition to inter-rater reliability correlations, there is, for example, the multi-faceted Rasch model operationalised in programs such as FACETS.

### **5.7.4. Documentation**

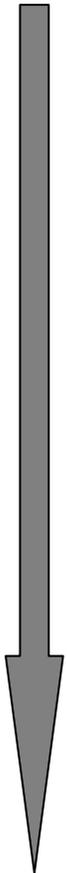
It is essential that at the end of the session the set of benchmarked samples are filed together with the records kept during the session(s). It is very helpful in future training if there is detailed documentation for each extract why a particular sample represents a certain level. In this respect the documentation provided with the DVDs of illustrative samples can serve as a model.

An audio recording of the discussion in the session can be a useful source for the preparation of such notes on each benchmarked sample. The coordinator may also decide to ask one or more of the participants to assist in taking notes explaining the reason why samples were identified as particular levels. These notes could then be standardised into a set of coherent documentation and circulated to participants after the session.

*Users of the Manual may wish to consider:*

- *how they can ensure a balanced and representative panel for the project*
- *how large a panel it is feasible and sensible to have*
- *what overall strategy is likely to be best in the context (in terms of resources, planning, implementation, analysis)*
- *whether the project will aim to benchmark “local” samples to use as context-specific illustrative samples in future*
- *how to ensure that such “local” material for benchmarking (and future training) purposes is of good quality*
- *what form documentation for the local material should take, and how it will be provided*
- *how much training is likely to be needed*
- *whether all participants will need to start from the same point – or whether some could be given a more elaborate “pre-task” than the others*
- *whether to use “plus levels” (there are arguments on both sides; what is important is not to change approach once the process has started)*
- *whether to use the CEFR-based rating Grids in Appendix C or develop other more sector specific CEFR-based instruments*
- *how to publish and/or disseminate the results of the standardisation process to the field*
- *how to ensure good “local” dissemination and follow up*

**Table 5.5: Standardisation Training and Benchmarking: Summary**

	Activity	Materials needed	Time	People	Suggestions
	<b>FAMILIARISATION</b>	<ul style="list-style-type: none"> <li>• Question checklists based on framework reminders (boxes)</li> <li>• Photocopies of Question checklists</li> <li>• Photocopies of CEFR Tables 1 and 2</li> <li>• Cut out versions of CEFR Table 2, other scales</li> </ul>	2 hours	Coordinator Big groups possible	Using self-training on-line package if available
	<b>TRAINING (Productive skills)</b>	<ul style="list-style-type: none"> <li>• Standardised performance videos (8 minimum)</li> <li>• Standardised scripts ( idem )</li> </ul> Photocopies of skill specific scales: <ul style="list-style-type: none"> <li>• CEFR Table 3/Tables B1–B3 (spoken performance)</li> <li>• Table B4 (written performance)</li> </ul> Photocopies of: <ul style="list-style-type: none"> <li>• Participant rating sheets (Forms B2–B3)</li> <li>• Coordinator rating forms (Form B4 )</li> </ul> Photocopies of other complementary scales, as relevant	3–4 hours/skill: 30min Introduction 90min Illustrative samples 90min Local samples	Coordinator 30 people max.	Doing two skills per day, or doing a half-day on training and half a day on benchmarking in relation to just one skill.
	<b>TRAINING (Receptive skills)</b>	Photocopies of skill specific scales: <ul style="list-style-type: none"> <li>• Overall Reading</li> <li>• Overall Listening</li> </ul> Photocopies of: <ul style="list-style-type: none"> <li>• Participant rating sheets (Appendix 2)</li> <li>• Coordinator rating forms (Appendix 3)</li> <li>• Photocopies of other complementary scales, as relevant</li> <li>• Calibrated model items</li> </ul>	3–4 hours/skill: 30min Introduction 90min Illustrative samples 90min Local samples	Coordinator 30 people max.	Doing two skills per day is possible as participants will at this stage be very familiar with the CEFR levels and with the standardisation activities.
	<b>BENCHMARKING PERFORMANCE SAMPLES (Productive)</b>	<ul style="list-style-type: none"> <li>• Local videos (8 minimum)</li> <li>• Local scripts ( idem )</li> </ul> Photocopies of skill specific scales: <ul style="list-style-type: none"> <li>• CEFR Table 3 /Tables B1-B3 (spoken performance)</li> <li>• Table B4 (written performance)</li> </ul> Photocopies of: <ul style="list-style-type: none"> <li>• Participant rating sheets (Forms B2–B3)</li> <li>• Coordinator rating forms (Form B4)</li> <li>• Photocopies of other complementary scales, as relevant</li> </ul>	3–4 hours/skill: 30min Introd. 90min Calibr. 90min Local.	Coordinator 30 people max.	Doing two skills per day, or doing a half-day on training and half a day on benchmarking in relation to just one skill.

## **Chapter 6**

### **Standard Setting Procedures**

#### **6.1. Introduction**

#### **6.2. General Considerations**

##### **6.2.1. Organisation**

##### **6.2.2. Concepts**

#### **6.3. The Tucker-Angoff Method**

##### **6.3.1. Procedure**

##### **6.3.2. The Minimally Acceptable Person**

##### **6.3.3. Probability Statements**

##### **6.3.4. Aggregating Individual Standards and Rounding**

#### **6.4. Two Variations of the Tucker-Angoff method**

##### **6.4.1. The Yes-No Method**

##### **6.4.2. The Extended Tucker-Angoff Method**

#### **6.5. The Contrasting Groups Method and the Borderline Group Method**

##### **6.5.1. The Contrasting Groups Method**

##### **6.5.2. The Borderline Group Method**

#### **6.6. The Body of Work Method**

##### **6.6.1. Training, Rangefinding and Pinpointing**

##### **6.6.2. Calculating the Standards: Logistic Regression**

#### **6.7. The Item-descriptor Matching Method and the Basket Method**

##### **6.7.1. The Item-descriptor Matching Method**

##### **6.7.2. The Basket Method**

#### **6.8. The Bookmark Method**

##### **6.8.1. The Task for the Panel Members**

##### **6.8.2. Content of OIB Pages**

##### **6.8.3. Technical details**

#### **6.9. A Cito Variation of the Bookmark Method**

#### **6.10. Special Topics**

##### **6.10.1. Standard Setting Across Skills**

##### **6.10.2. Standard Setting and Test Equating**

##### **6.10.3. Cross Language Standard Setting**

#### **6.11. Conclusion**

## 6.1. Introduction

The basic output from taking a test is a numerical score. In the case of highly itemised tests used for Reading and Listening for example, this score usually is the number of correct responses. In the case of productive skills the task performance is mostly judged on a number of aspects, and for each aspect the test taker receives a number of “points” (ranging for example from zero to four or five). The test score in such a case is the total number of points collected by the test taker across all aspects and all tasks he or she has made.

Based on this score a decision on the examinee’s ability is taken, the main one being a pass/fail decision: has the candidate performed satisfactorily on the test? If an examination is to be linked to the CEFR another decision has to be made as well, the decision whether the candidate has reached a particular CEFR level (e.g. B2) or not. Both decisions (pass/fail; attainment of a CEFR level) involve the determination of a *cut score* defining a *performance standard*. In a pass/fail decision, the cut score is the minimum score on the test that will lead to the decision “pass”; scores lower than the cut score lead to the decision “fail”. Similarly, a cut score for B2 is the minimum score that will lead to the decision/classification that the ability of the candidate is at Level B2 or higher; lower scores are interpreted as “lower than B2” (= B1 or lower).

It is possible that multiple standards have to be set for the same test. In linking to the CEFR, one might wish, for example, to set a cut score for A2, B1 and B2. It is important to understand what is precisely meant by the preceding sentence. A cut score is to be conceived as a border between two adjacent categories on some scale. So the example should be understood in the sense that every test taker will be classified either as A2, B1 or B2, and hence we need two cut scores: one that marks the border between A2 and B1 and one for the border between B1 and B2. In general the number of cut scores is one less than the number of classification categories.

To avoid confusion between categories (the levels) and cut scores (the boundaries between them), one often denotes the cut scores by naming the two adjacent categories. In the example in the last paragraph with three categories, the cut scores could be indicated as A2/B1 and B1/B2. One should be careful with the labelling of the two extreme categories: labelling the lowest category in the example as A2 could imply that any test taker having a score lower than the A2/B1 cut score is at Level A2, including the ones having a score of zero. Therefore it is better to make the label all inclusive and to call it, for example “A2 or lower”. Similarly, using “B2 or higher” is more appropriate for the highest category in the example.

Determining the cut scores or *setting the (performance) standards* is usually a group decision. The group that makes such decisions is normally called a *panel*. Panel-based approaches typically take many days. Most of the time is spent with activities which are described in the previous chapters. For linking examinations to the CEFR, panellists have to be familiar with the CEFR itself (Chapter 3), they will have to ensure that the coverage of the examination itself is related to the CEFR (Chapter 4), and they will have to be trained in how to apply the CEFR descriptors to the examination (Chapter 5). In the present chapter, the attention is focused on the more formal aspects of the group decision making: the kind of judgments made by the panellists, the kind of information they have available and the way their judgments are treated and aggregated to arrive at single or multiple cut scores. Such procedures have been formalised and are known as *standard setting procedures*.

Standard setting can have important consequences for individuals and for policy makers. It requires careful judgment and this means that “standard setting is perhaps the branch of psychometrics that blends more artistic, political and cultural ingredients into the mix of its products than any other” Cizek (2001, p. 5).

## 6.2. General Considerations

An essential part of any standard setting procedure is the efficient organisation of the meetings. Usually, part or all of the Familiarisation, Specification and Standardisation phases described in earlier chapters of this Manual form an organic whole together with the standard setting procedures (in the strict sense) that are discussed in this chapter. Therefore, the whole procedure is rather demanding and requires efficient organisation. An excellent introduction can be found in the first chapters of Cizek & Bunch (2007). In this section, attention is therefore restricted to the standard setting proper, and essential elements will be outlined only briefly.

### 6.2.1. Organisation

Panel-based standard setting procedures usually take two or three days, starting with one or more sessions on familiarisation, discussion of the test specification, training with illustrative material and a vital step in which all the panel members complete the test paper made up by the items under consideration. After suitable instruction the panel members give their judgments, usually in two or three rounds separated by discussion phases and the provision of feedback and additional data.

In the sessions between rounds, essentially two kinds of information are given. After the first round, information is given about the behaviour of the panel members themselves, showing that some members give very outlying judgments. This kind of information is called *normative* information, and is intended primarily to detect and eliminate misunderstanding of the instructions. It is good practice to let panel members discuss this information in small groups. The danger of public discussion is group pressure towards the viewpoint of one or more dominating personalities in the group (see suggestions in Section 5.4.1). It is the task of the group leader to lead the discussions in such a way that panel members do not feel under pressure from such behaviour.

After the second round, a different kind of information called *impact* information is usually given. This shows the consequences of the panel's judgments by computing the proportion of students who would have reached or failed to reach each standard based on the provisional cut-offs determined by the result of the previous round. Of course, to be able to do so, one has to have collected the scores of a representative sample of students.

The preceding paragraph may be confusing in some sense. Standard setting as described in this chapter treats test performance clearly from a criterion referenced perspective: qualified judges are asked to formulate the minimal requirements (in terms of test performance) to pass an exam or to earn the qualification "B2", and they are supposed to be led by an application of a general system (in our case the CEFR) to a concrete test or examination. One could think therefore that it does not matter whether 10% or 90% of the test takers in some population will pass the exam. But one should not forget that high stakes standard setting is usually embedded in a social and often political context, and that it is wholesome therefore to confront panel members with the societal consequences of their decisions. It may happen that after providing impact information, a number of panel members change their mind and become more strict or more lenient, for opportunistic reasons, than they were before. If this happens, it does not imply necessarily that their changed opinion is the final decision; on the contrary: large shifts in the standards after providing impact information should be used for an in depth discussion with the aim to find a rational and reasonable compromise between two highly different group decisions, and this may be sufficient to organise a fourth round of judgments.

It should be borne in mind that the systematic presentation of normative and impact information needs a lot of preparatory work so that the resulting computations (which depend on the judgments of the panel) can be undertaken efficiently, (e.g., during a lunch break) so that the information is available for the next round.

For almost all standard setting procedures described in the literature, many variations have been tried out, shaped to particular needs or inspired by shortcomings in earlier experiences. Some applications exemplify what is essentially the same procedure, but may differ in the number of judgment rounds, in the organisation of the discussions (plenary versus small groups), etc. There is no need in any application to follow all details of a described procedure, and variations deemed to serve better a particular setting can be introduced. In the remainder of this chapter, procedural details and possible variations are not discussed; the features described for all methods are to be considered as essential to the methods.

To make a judgment on the validity and efficiency of any procedure that is applied in any project, however, it is essential that adequate documentation on all steps and procedural details is available. Without such procedural detail, professional judgment on the results is difficult and one cannot claim to have built an argument.

## 6.2.2. Concepts

Recognising that standard setting cannot be carried out properly by just following mechanically any particular method, this chapter will provide a discussion of some fundamental concepts that come up in various standard setting methods. Such concepts include:

- probability statements;
- mastery probability or response probability;
- partial credit scoring;
- concepts related to IRT (difficulty parameter, difficulty level, discrimination);
- decision tables;
- ordered item booklet (OIB), and
- threshold region.

It is difficult to introduce such concepts in the abstract. Therefore they are introduced in the chapter as they first become necessary in order to describe a particular method. The order in which the concepts are presented is purely to assist the user in following the development of the concepts. No specific implication that methods presented earlier are in some way “worse” is intended. The chapter presents a range of standard setting methods to choose from but, as the standard setting contexts vary, it does not advocate the use of any single one of them.

Sometimes standard setting methods are divided into test-centred and examinee-centred methods. Three methods of the latter category are discussed. The Contrasting Groups method and the Borderline Group methods use direct judgment of test takers by a rater who knows them well. The Body of Work method asks holistic judgments on all the “work” from a sample of students that is used to determine their score on the test or examination; this may be answers to multiple-choice questions, to constructed response items, but it may also be as broad as an essay or even a portfolio. The important characteristic of these examinee centred methods is that *specific* examinees are classified (as passed or failed, or as B1, B2, or as a borderline case) by a *holistic* judgment.

In the older methods such as the Tucker-Angoff method or the Nedelsky method<sup>16</sup>, panel members are asked to make a judgment on each item. These judgments are based on the perceived characteristics of the items by the panel members and the whole procedure can be applied without any empirical data from test takers taking the test. For these methods, the term test-centred is indeed appropriate. With the growing popularity of item response theory (IRT), however, methods have been developed where the distinction between examinee centred and test centred methods is less clear. In these methods information is available for the panel members, which derives directly from the performance of a *group* of test takers. Usually this information takes the form of item difficulty estimates. Availability of such information is meant to help the panel members and to exempt them from the difficult task to provide difficulty estimates based exclusively on the perceived features of an item.

The methods discussed in this chapter may therefore be categorised in three groups. The first group will be labelled examinee-centred (E-C), the second group test-centred (T-C) in the sense that it can be applied without any empirical test taking data and the third group will be labelled IRT, meaning that panel members use a summary of empirical data (usually provided via an IRT analysis).

Table 6.1 provides an overview of the various methods discussed, their classification as given above and the section number where the method is discussed. In Section 6.10 some special topics are discussed.

The quality of standard setting can vary extensively. Whichever method or combination of methods is adopted, it cannot be assumed that standard setting has been done properly just because certain procedures have been followed. There is a need to collect *evidence of the quality of the outcomes* of the procedures and to report these in a sufficiently detailed and transparent manner. This validity-related issue will be discussed in greater length in the final chapter of this Manual.

---

<sup>16</sup> This method is probably the oldest method of standard setting. It is not discussed in this Manual. A good description can be found in Cizek and Bunch (2007, Chapter 4).

**Table 6.1: Overview of the methods discussed**

<b>Method</b>	<b>Section</b>	<b>Class</b>
Tucker-Angoff	6.3.	T-C
The Yes-No Method	6.4.1.	T-C
The Extended Tucker-Angoff Method	6.4.2.	T-C
The Contrasting Groups Method	6.5.1.	E-C
The Borderline Group Method	6.5.2.	E-C
The Body of Work Method	6.6.	E-C
The Item-descriptor Matching Method	6.7.1.	T-C
The Basket Method	6.7.2.	T-C
The Bookmark Method	6.8.	IRT
A Cito Variation on the Bookmark Method	6.9.	IRT

### 6.3. The Tucker-Angoff Method<sup>17</sup>

Although the method was introduced in 1971 as a kind of side remark in a chapter on scaling, norming and test equation that Angoff wrote for the second edition of the reference book *Educational Measurement* (Thorndike 1971), it is still, after more than 35 years, one of the most widely used standard setting methods. Many variations of it have been proposed, and in this chapter two of them will be discussed. We start with what is nowadays known as “The Angoff method”, although actually Angoff presented it only in a footnote as a variation of the procedure proposed in the main text.

#### 6.3.1. Procedure

A basic concept, which also appears in many other standard setting procedures, is the concept of the “minimally acceptable person”, also referred to sometimes as the “borderline person” or person “just barely passing” or “minimally competent candidate”. Where a standard has to be set, for example, for CEFR Level B1, a minimally acceptable person has the competencies, skills and abilities to be labelled as “B1”, but only to such an extent that the slightest decrease in those competencies, skills and abilities would suffice in order not to grant this qualification. The task for the panellists is to keep in mind such a person or collection of persons during all the judgmental work they have to do.

For each item in the test, the panel members have to give the probability that such a minimally acceptable person would give a correct answer. So the basic data collected in a judgment round can be presented in a table like Table 6.2, where 15 raters have formed a standard setting panel for a test of 50 items.

As a next step in the procedure, the probabilities are **summed** across items for every rater. For rater one in the example, this sum amounts to 17.48. As the probability of a correct answer with a binary item equals its expected score (see Section C in the Reference Supplement), the sum of the probabilities across items equals the expected test score of the minimally competent person, according to rater one. In the example we see that these sums differ across raters, and this is always the case in real settings. So there remains the problem of aggregating the sums of the individual raters in some reasonable way to come to a final standard. One method, often applied in practice, is just to take the **average** of the sums, and to consider this average as the standard.

To summarise: three components are essential in the procedure: the concept of the minimally acceptable person, the assignment of a probability for a correct response for such a person (to be given for each item by each of the panel members) and the aggregation of the sums of these probabilities across panel members. Each one of these aspects will be commented upon in the following sections.

<sup>17</sup> In the literature this method is usually called the Angoff method, but Angoff himself attributed the method to his colleague at ETS, Ledyard Tucker.

**Table 6.2: Basic Data in the Tucker-Angoff method**

	Rater 1	Rater 2	...	Rater 15
<b>Item 1</b>	0.25	0.32	...	0.35
<b>Item 2</b>	0.48	0.55	...	0.45
<b>Item 3</b>	0.33	0.38	...	0.28
...	...	...	...	...
<b>Item 49</b>	0.21	0.30	...	0.35
<b>Item 50</b>	0.72	0.80	...	0.90
<b>Sum</b>	<b>17.48</b>	<b>19.52</b>	...	<b>18.98</b>

### 6.3.2. The Minimally Acceptable Person

The concept of a minimally acceptable person or borderline person is central in this approach. In the training of the panellists great care must be given to provide a reasonable definition of it, and to make sure that the internal representation panel members have of such an (abstract) person is (a) highly consistent among panel members and (b) is in accordance with the purpose and interpretations of the test results.

Suppose a standard is to be set for the Level B1, i.e., a cut-off for A2/B1. To be sure that the cut-off reflects this boundary and not something else, one has to ascertain that the panel members have an accurate grasp of what is meant by A2 and B1, or more generally, that they are intimately familiar with the CEFR. Moreover, they should have a clear and consistent idea on how the CEFR applies to each item, meaning that they have to know which “Can Do” descriptors are relevant in answering each item, and in particular they should have a clear idea of which descriptors are the critical ones: the ones that distinguish best between A2 and B1. The process of arriving at a good understanding of the critical difference between A2 and B1 with respect to each item in the examination is a time consuming and onerous activity. Guidance to organise this activity can be found in the preceding chapters.

In some variations of the Tucker-Angoff method, it is suggested that panel members have a **concrete** person in mind, whom they typically would consider as a borderline person, for example a student they know very well. The argument put forward for this procedure is that it helps the panel members to have a **stable** idea of the borderline person when they go through the list of items. Although this is admittedly true, working with concrete persons has two disadvantages. First such a person is usually known by only one of the panel members, and it may be fairly difficult to use characteristics of such a person in group discussions, because nobody – except one panellist – knows that person. The second and more important disadvantage of using concrete persons is that if everybody is thinking of their own separate concrete person, it will be harder to correct misconceptions one might have about the correct meaning of the (abstract notion of the) borderline person. This problem can occur when starting the standard setting, and it might well appear during training and group discussions. In any case, it should be clear that working with “private” concrete borderline persons cannot be a substitute for thorough training.

### 6.3.3. Probability Statements

For each item, the panel members have to state the probability that a correct answer would be given by the borderline person. As people not acquainted with probabilities might be scared by such a task, it may be helpful to concretise the task a bit. One might say for example, “suppose that 100 borderline persons answer the item, how many of them do you expect to give a correct response?” The number given by the panel member is then divided by 100 and considered as his or her probability estimate. This probability estimate is nowadays commonly referred to as *Angoff rating*.

The use of the number “100” in the above example has two advantages: firstly, the answer given by the panel member can be directly interpreted as a percentage, and secondly the number of possible answers (0, 1, 2, ..., 100) is large enough to warrant accurate expression of probabilities. Suppose a panel member has in mind a probability of  $\frac{2}{3}$  or 0.6666... In answer to the question allowing only 100 persons, he will probably say 67<sup>18</sup>.

There are two aspects to be kept in mind when panellists are asked to make probability statements. The first is that with multiple-choice questions, the probability of a correct answer can be substantial, even if the level of the candidate’s ability is far less than that of the borderline person. The reason is correct guessing. It is useful to remind panel members of this and for example to urge them not to state probabilities that are below chance level (one divided by the number of response alternatives). This is an important issue for the discussions between rounds and during the training.

The other aspect has to do with a tendency to avoid extreme statements. This means that when fed with enough information to make extreme probability statements, there exists a tendency in human judgment to avoid these by giving values larger than the “real” values when these are very low, or lower than the real values when they are very high. If such a tendency is present when using the procedure, the effect will differ depending on the general level of difficulty of the test or examination. Suppose the test is quite easy for the borderline person, leading to quite high probabilities for many items. If these probabilities are systematically biased downwards because of the tendency to avoid extreme (high) estimates, the net effect on the cut-off score will be that it will be lower (more lenient) than in the case without such a tendency. If on the other hand the test is quite difficult for the borderline person, the opposite will occur: the prevailing low probabilities will be overestimated, and the standard will be biased upwards.

Of course it is very difficult to measure the extent to which such a conservative tendency occurs in a particular standard setting project, but one can try to avoid these phenomena in two ways. The first way applies to all judgmental standard setting methods: be modest in your ambitions. It is an illusion to think it is possible to build a test and to set standards for the six basic levels of the CEFR (A1 to C2) within the same test or examination by using test-centred standard setting methods. For the Tucker-Angoff method this would imply that for the A1/A2 borderline person there would be many very difficult items (needed for the C1/C2 standard), and conversely that for the C1/C2 borderline person there would be many very easy items (needed for the A1/A2 standard). Even a weak tendency to give conservative probability estimates may have a quite substantial effect on the cut-off scores, being too harsh for the lower levels and too lenient for the higher levels.

The second way to avoid systematic distortions in the probability estimates is to provide panellists with what Cizek and Bunch call *reality feedback*. This can be done in the following way and on the condition that real test data is available. After the first round of standard setting, provisional standards can be computed. Suppose that in the 50 item test used in the example in Table 6.2 the average sum of probabilities is 18.52, so the standard will be a score of 18 or 19. If this standard is not too far from the final standard, it is reasonable to consider students with a score in the vicinity of this provisional standard as borderline students. For these students one can compute the proportion of correct answers to each item, and give the results of these computations as feedback to the panel when they are preparing for the next round. These proportions are empirically based estimates of the proportion of correct answers for borderline persons. Panel members may

---

<sup>18</sup> This is not the same as 100 times  $\frac{2}{3}$ , but the error is small enough not to cause any systematic effect (a bias) in the final result. If one uses “10” instead of “100” (or asks to give probabilities rounded to one decimal, i.e., the possible answers are 0, 0.1, 0.2, ..., 1), then systematic distortions in the final result will occur, especially if the standard is set near either end of the score range. (Reckase 2006a; 2006b.)

compare their own estimates with it and be led to make reasonable adjustments. From the probability statements in the next round, it can then be seen whether and to what extent possible conservative statements have been adjusted in the desired direction.

To define a reasonable vicinity of the provisional standard, one will usually have to compromise between the width of the range allowed and the number of students having a score in this range. Suppose one sets the provisional standard at a score of 19 points, and suppose only 15 students have obtained this particular score. The proportion of correct answers for each item in this small group will have a large standard error because there are few students. Widening the definition of the vicinity from 17 to 21 for example, may raise this number considerably, but on the other hand, if the standard is really 19, it may be disputed whether persons with a score of 17 or 21 can still legitimately be considered as borderline. A defensible strategy is to define the vicinity as the provisional standard plus and minus the standard error of measurement. To avoid biases, it is important to take the range of the vicinity symmetric around the provisional standard.

### 6.3.4. Aggregating Individual Standards and Rounding

Summing the probabilities over items for an individual panel member yields the individual standard for that panel member. Taking the average of all these individual standards may be considered as the standard set by the whole panel. This may seem to be a bit trivial as the only reasonable way to aggregate the different individual standards. But this is not so. In some respects averages are vulnerable measures as representatives of a whole group. They are especially vulnerable to outliers, which can come about if one or two panel members are very stubborn in keeping to very extreme standards, or have not understood the procedure. To avoid such extremes influencing the group decision too much one might take a more robust measure. The most popular one is the median, but another useful one is the trimmed average. A trimmed average is the average of a set of data where a certain percentage of the data is excluded from the computations. The excluded data are the most extreme ones (high as well as low). If there are 20 panel members, and the percentage of trimming is set at 10%, then the highest and the lowest value are excluded and the average is computed on the 18 remaining values.

Usually the individual panel member standards, as well as the group standard, be it an average, a trimmed average or the median, will be fractional numbers. But fractional scores cannot occur in practice as the outcome of individual test taking. Therefore the fractional outcome will have to be rounded to the integer score just below or just above it. This may look like a trivial problem – round the integer score closest to the fractional standard; in the example this would mean round 18.55 to 19 – but the issue is more complex than this.

To understand this, one should realise that any standard setting, no matter how carefully it is set up, will inevitably lead to classification errors because the test scores themselves are not perfectly reliable. But these classification errors can go in two ways: a student with a true score at or above the cut score can be classified as not having reached the standard (a false negative), and conversely, a student with a true score below the standard can, due to the measurement error, be classified as having reached the standard (false positive). Classification errors have consequences at the individual level and possibly at a societal level, and more importantly, the consequences of false negatives may be different from the consequences of false positives. If the latter are deemed more serious, then there is reason to make the standard harsher and thus to round the fractional standard **upwards**. More detailed discussion on the consequences of classification errors will be addressed in the next chapter.

One final warning about rounding is in order here. Rounding numbers, and doing further calculations with rounded numbers, can have unwanted and unforeseen consequences. Therefore, rounding should be postponed as long as possible. It is bad practice, for example, to round the individual standards (the bottom row in Table 6.2) for each panel member to the nearest integer, then compute the average of the rounded numbers and to round the result again. A simple example can show this: suppose there are three raters with individual standards 17.01, 17.51 and 17.53 respectively. The average is 17.35 which yields 17 when rounded. Rounded individual standards gives 17, 18 and 18 with an average of 17.67 and 18 as a rounded average.

## 6.4. Two Variations of the Tucker-Angoff Method

In applications of the Tucker-Angoff method, the task of estimating the probabilities of a correct response is often felt to be difficult and hard to understand. A variation of the method, called the Yes-No method<sup>19</sup> avoids this problem.

The original proposal by Angoff was exclusively directed to tests consisting of **binary** items. In many tests, especially with productive skills, some items also have polytomous scores, where one can earn, for example, zero, one, two or three points. The Tucker-Angoff method can (in principle) be extended to such cases as well. In this section both variations are discussed briefly.

### 6.4.1. The Yes-No Method

The clearest description one could wish for is the original text by Angoff himself:

“A systematic procedure for deciding on the minimum raw scores for passing and honours might be developed as follows: keeping the hypothetical ‘minimally acceptable person’ in mind, one could go through the test item by item and decide whether such a person could answer correctly each item under consideration. If a score of one is given for each item answered correctly by the hypothetical person and a score of zero is given for each item answered incorrectly by that person, the sum of the item scores will equal the raw score by the ‘minimally acceptable person’ (Angoff 1971 pp. 514–515).

So instead of giving probability statements (numbers from zero to one), the panel members assign only one (saying Yes) or zero (saying No). Although good results have been reported with this method (see Cizek and Bunch 2007, pp. 88–92 for some results), the method can lead to severely biased results.

To see this, one could consider the answers given (0 or 1) as probabilities rounded to zero decimals. Now consider a rather homogeneous test that is relatively easy for the borderline person. This could mean that for all items, the borderline person has a probability of over 50% to give the correct response, so that a rational panel member should answer Yes for every item. But if he does do so, his individual standard will be the maximum test score, while real borderline persons might obtain on average a score which is only slightly larger than half of the maximum score.

From this we can deduce a more general principle about the meaningfulness of standard setting. In the previous example it will be clear that a meaningful result can only be obtained if there are items in the test which the borderline person can get right (with a probability substantially higher than 0.5) and cannot get right (with a probability substantially lower than 0.5). This will prevent that the cut-off score is very extreme (close to zero or close to the maximum). In more abstract terms this means that the test should convey sufficient information about the ability of the borderline person, and this leads to the same conclusion as was reached in the previous section: if multiple standards have to be set for abilities that are quite far apart (for example for A1/A2 and B2/C1) using the same test, one has to collect sufficient information on several quite disparate ability ranges, which is usually not feasible, unless the test is very long. Ignoring this principle can lead to absurd results as is shown by the following example. Suppose a test is constructed to make a distinction between B2 and C1 as its primary purpose. Using this test to set the standard for A1/A2 will probably yield a cut-off score of zero in the Yes-No method (a borderline A1/A2 person does not answer correctly to any of the items), and lead to the absurd conclusion that one is at Level A2 if one obtains a score of zero on this test.

---

<sup>19</sup> In fact, this was what Angoff originally proposed as his method of standard setting. The method discussed in the previous section was proposed in a footnote.

### 6.4.2. The Extended Tucker-Angoff Method

A generalisation of the method to tests that consist of any mixture of binary and polytomous items is easy to understand, if one sees that the probability of a correct answer on a binary item is the same as the **expected score** for that item (see Section C of the Reference Supplement). For **polytomous** items, it is far more difficult to specify response probabilities, because then we have to specify the probability of obtaining a score of 1, 2, etc. until the maximum score for that item. One can, however, circumvent this problem by specifying the **expected score** for a polytomous item. The instruction for the panel members in such a case could go like this:

*“Suppose that 100 borderline persons answer the item, where one can earn up to [4] points, what would be in your view the average score obtained by these 100 persons?”*

Instead of filling out a probability in a table like Table 6.2, one fills out the expected average score as specified by the panel member. The remaining parts of the procedure (summing and aggregating) remain the same as in the Tucker-Angoff method for binary items.

The only extra problem with this method is that one should ascertain that panel members understand well what an average score is. In particular, they should understand well that the average can be a fractional number although individual scores can take only integer values. A good method is to teach them to set up for themselves a frequency table of the possible scores for the 100 borderline persons and then to compute the average. An example of such a table is given in Table 6.3 for an item with a maximum score of 3. The basic task of the panel member then consists of filling out the frequency column in the table (and checking that the sum is 100). The third column (score times frequency) then follows mechanically, and from the example in Table 6.3, one derives immediately the result that the expected score is  $75/100 = 0.75$ . If one suspects that constructing the third column and doing the multiplications and sums is too hard for some panel members, one can just prepare a simple table, leaving out the third column, and let the panel members only specify the frequencies. The necessary computations to arrive at the average can then be done off-line.

**Table 6.3: Computing the Expected Score of 100 Borderline Persons**

Score	Frequency	Score * Frequency
0	45	0
1	35	35
2	20	40
3	0	0
<b>sum</b>	100	75

**Concluding remark:** The Tucker-Angoff method and its many variations are *a typical test-centred method*, because the primary task for the panel members is to concentrate on the characteristics of the items and to classify these items with respect to the ability of an abstractly defined borderline person. This classification is absolute (in the Yes-No method) or probabilistic. Considered from a purely formal viewpoint, one could say that to apply this method, panel members need not to have any teaching or other experience with real students in the subject matter of the test, but in practice using such a panel might result in totally unacceptable standards. Even with experienced teachers, the task setting is quite abstract, and teachers usually find it quite difficult to give the required judgments. Therefore, all variations on the method nowadays use several rounds and provide information about real students' performances to moderate the standard setting. Providing impact data gives evidence on the consequences for groups of students and can lead to important adjustments. Providing reality data, the proportion correct calculated on a borderline group, which is defined in terms of the provisional standards, can give clues which help to adjust probability estimates to more realistic values. Yet, even with these provisions, the main focus of the method is on the characteristics of the test, the qualification of the method as test centred remains justified. In the next section, two examinee centred methods will be described.

## 6.5. The Contrasting Groups Method and the Borderline Group Method

These two methods form a strong contrast with the Tucker-Angoff method in the sense that the judgments of the panel members are based primarily (and almost exclusively) on the performances of **real students** on the test. Therefore they are a prototype of **examinee-centred** methods.

Common to both methods is the requirement that test scores from a sample of students are available. As is common to all standard setting methods, one should take care that the sample is representative of the target population. Moreover, the students must be well known by (at least) one of the panel members. In practice this usually means that the panel will consist of the teachers of the sampled students, and consequently that each student in the sample is well known by exactly one of the panel members.

### 6.5.1. The Contrasting Groups Method

The task for the panel members is to assign each student to one of two categories (in the case of a single cut-off score) or to  $k+1$  categories when there are  $k$  cut-off scores. If the purpose of the standard setting is to establish, for example, the standard for B1/B2, every student is categorised by the panel members as either B1 (or lower) or B2 (or higher).

Once this information is available, a frequency table with two columns can be constructed. The rows represent the score on the test, and the two columns display the frequencies of the scores for the groups of students categorised as B1 or B2 respectively. An example, based on artificial data for a test of 50 items is given in Figure 6.1, where the two frequency distributions are displayed graphically. The total sample consists of 400 students, 88 were categorised as “B1” and 312 as “B2”. The distributions displayed have a number of features that occur quite often in practice: they are very irregular (a consequence of the moderate sample sizes) and they have considerable overlap. Therefore it is not immediately clear where to place the cut-off score. Moreover, the two groups differ markedly in size, as is often the case in pass–fail decisions.

The average score of the B1 students is 16.78 and for the B2 students it is 34.24. An acceptable cut-off score, at least provisionally, is the value midway between these two averages, yielding  $(16.78 + 34.24)/2 = 25.51$ . One should be careful, however, in taking this value (or a rounded value near it) as the definite standard.

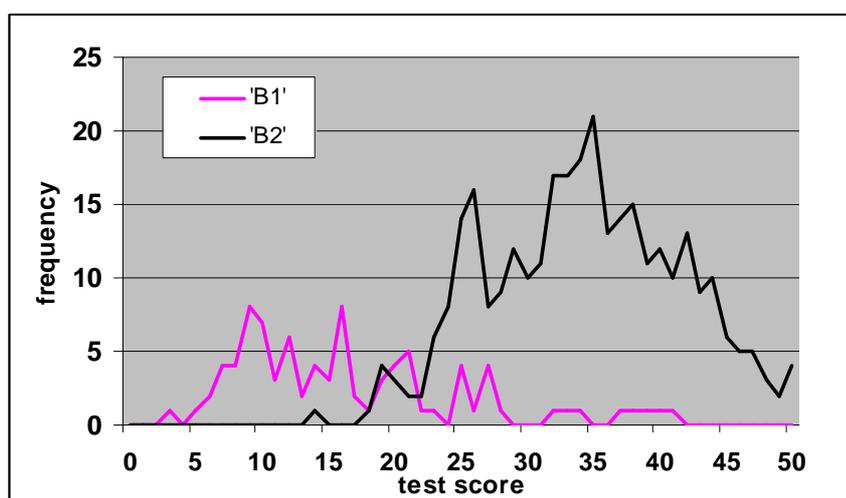


Figure 6.1. Frequency Distributions of Test Scores in Two Contrasting Groups

In the B1 distribution, seven students (out of 88) obtained a quite high test score (of over 30 points), and could be considered as outliers. It is worthwhile to check whether these seven students were categorised by the same teacher or not. If they are, it might be a point of discussion in the panel to see whether this teacher has not been too strict in his/her judgments, and if necessary obtain revised judgments. But even without outliers, overlap in the distributions will in general be observed.

A good technique to make a rational choice is to construct **decision tables for several cut-off scores**. This technique is illustrated next. In Table 6.4, the frequency table corresponding to Figure 6.1 is displayed in a compressed form: low scores (up to 20) and high scores (from 28 on) are taken together; the other scores are displayed separately.

**Table 6.4: Frequency Distribution Corresponding to Figure 6.1**

Score	B1	B2
0–20	63	9
21	5	2
22	1	2
23	1	6
24	0	8
25	4	14
26	1	16
27	4	8
28–50	9	247

The five subtables in Table 6.5 are directly derived from Table 6.4. Take the cut-off score of **24** as an example: from Table 6.4 it is seen that 18 students, categorised as B1 by their teacher “pass” the test, and are thus considered as B2 on the basis of their test score. These 18 are false positives. Similarly, 19 students categorised as B2 by their teacher do not “pass” the test, and are false negatives. Taken together, this means 37 misclassified students on a total of 400 students which is 9.3%.

**Table 6.5: Decision Tables for Five Cut-off Scores**

Classified as:	Cut-off = 21		Cut-off = 22		Cut-off = 23		Cut-off = 24		Cut-off = 25	
	B1	B2								
<b>Below cut-off</b>	63	9	68	11	69	13	70	19	70	27
<b>Cut-off or higher</b>	25	303	20	301	19	299	18	293	18	285
<b>Total</b>	88	312	88	312	88	312	88	312	88	312
<b>% misclassifications</b>	8.5		7.8		8.0		9.3		11.3	

From Table 6.5 it is seen that the percentage of misclassifications changes as the cut-off score varies. It reaches its minimal value at a cut-off score of 22, and changes very little at 23. Therefore 22 or 23 might be preferred to the provisional values of 25 or 26 determined by the midpoint of two averages.

There is another aspect of this procedure that one should not lose sight of. The *numbers* of misclassifications, i.e., the number of false positives and the number of false negatives, at every cut-off point in Table 6.5 are not equal, but they are reasonably similar. But this comparison is not to the point because the number of students classified as B1 and B2 by their teachers are very dissimilar. Take a cut-off of 24 as an example (where the numbers of false positives and false negatives are almost equal). The false positives represent 18 out of 88 or 20.4% of the B1 students, while the 19 false negatives represent only 6.1% of the B2 students. At the cut-off of 22, these percentages are 22.7% and 3.5% respectively, representing a kind of dilemma that can occur in practice: the optimal cut-off score in terms of the total percentage of misclassifications is in general not optimal with respect to the balance of false positives and false negatives. Careful considerations about the relative costs of false positives and false negatives and overall costs of misclassifications may be needed to arrive at a final decision.

There are two considerations to be careful about when one applies this method in high stakes situations. The first is a statistical one, the second is of a more methodological nature. As to the first, the sample sizes used in the example above are moderate, especially in the B1 group. This makes the numbers in Table 6.5 statistically unstable, meaning that upon replication with another sample of the same size, the corresponding table might change substantially, and lead to another choice of the optimal cut score.

The other consideration is still more serious. The whole reasoning in constructing the tables and interpreting them is based on the assumption that the judgment of the teachers is completely trustworthy and corresponds

to reality ('if your teacher says you are a B1, then you are a B1'). Of course this is not the case and teacher's judgments, however well trained they may be in the CEFR, will not be completely valid. It is true that an overestimation of some student by one teacher may be compensated by an underestimation of another student by another teacher, but the problem is that one has almost no control over this, because students are *nested* within teachers. If one or two teachers are too lenient, say, leading to too many B2 categorisations in the example, it is almost impossible to detect such a leniency. Even if they have substantially more B2 judgments than their colleagues in the panel, this is not a proof of their leniency, because it is possible that they have more able students. One could try to check this by using the test scores, showing for example that the average score of their students is about the same as the average of the other students, and that therefore they have to adjust their judgments. But this is dangerous practice. The whole method of using contrasting groups for standard setting rests on a comparison of two variables: the test scores and the judgments of the teachers. To be a sound method, the data for the two variables should be collected *independently*, meaning for example that the teachers have to give their judgments on the students without knowing their test scores. Now if one uses information from one of these variables to adjust (change) the other, one destroys this independence. In fact, by doing so one manipulates the data (towards a certain decision) and this jeopardises the whole procedure.

### 6.5.2. The Borderline Group Method

This method is very similar to the Contrasting Groups method: it also rests on a judgment of the level of concrete students. The judgments themselves, however, are meant to identify those students who can be conceived as being borderline cases at the intended standard. Continuing the example of the preceding section, one would try to identify the students who are somewhere near the border of the B1-level and the B2-level.

Once this group is identified, the cut-off score is defined as some central value of the test scores of this group, for example the average or the median, and then rounded appropriately.

The principle of this method is very simple, but the implementation may encounter several difficulties. Some of them are discussed next.

The first, and perhaps most delicate one, is a clear definition of what is meant by a borderline student. In the CEFR, levels are operationalised by "Can Do" descriptors, but borderline cases are not explicitly described. Defining them as "something in between two 'Can Do' statements" may be too fuzzy to ensure a common understanding of the CEFR, from which unwanted and uncontrollable variation among panel members may crop up. A good method to guide panel members in their understanding of borderline cases would be to use benchmarks: annotated examples of borderline performance.

The second difficulty is of a statistical nature. It is not uncommon that the size of the borderline group is moderate, not to say small, so that the average or median test score of this group will have a rather large standard error. Moreover, in applying this as a standalone method, useful information on the other students' performance on the test is not used. A way out of this is to combine the Borderline Group method and the Contrasting Groups method. This is discussed next.

Consider again the example of setting the cut-off score B1/B2. Instead of asking the panel members to classify students as borderline or not borderline, one might ask them to classify their students into *three* categories: "B1", "B1/B2", or "B2". The two groups "B1" and "B2" can then be used in a Contrasting Groups method, and the borderline group "B1/B2" can be used for the Borderline Group method, giving two provisional standards. This is useful information for the validation of the procedure, and more on this will be said in the next chapter. To set up the decision tables (see Table 6.5), the results can easily be combined, or even better, the tables can be set up separately, giving information on the rate of misclassifications for students who were definitely not borderline (according to the panel judgments), and for those who were judged borderline, the former being more serious than the latter.

This method functions satisfactorily when one can be sure that all students in the sample are either B1 or B2 (or somewhere on the border between them). If there is a suspicion that weaker or stronger students have

participated in the examination, it is safer to add one or two extra judgment categories, which might be labelled for example 'A2/B1 or lower' and 'B2/C1 or higher'. Even if one does not have the intention of setting the cut scores for A2/B1 or B2/C1, these extra categories may help in purifying the contrasting groups B1 and B2.

A further advantage of this combined method is that it avoids forced choices from teachers in case they have doubts themselves about the definitive category to place their students in.

## 6.6. The Body of Work Method

The Body of Work method (Kingston et al 2001) is perhaps the most suitable one for handling holistic judgments, although it can be used with any mixture of item types and tasks. It is examinee centred and it does not use IRT. Here is a brief list of what is needed to apply the methods:

- A collection of the work of a sample of examinees. The total work can consist of only answers to multiple-choice questions, or a mixture of multiple-choice questions, constructed response questions and essays or even a complete portfolio. A necessary condition, however, is that the work (test performance, portfolio) has received a *numerical score*.
- The sample does not need to be representative for the target population of the test. It must, however, cover most of the range of the possible scores, independent of the relative frequency of these scores which are available before the standard setting.
- The task for the panel members is to give a *holistic judgment* on each of the work samples presented to them. In the framework of the CEFR such a judgment will be the allocation of the examinees to one of the predefined levels one wishes to set the standard for. Suppose one wants to set standards A1/A2 and A2/B1, then the judgment asked from the panel members is to categorise each student's work either as A1, A2 or B1 (or higher).
- The kind of judgment asked from the panel members is the same as in the Contrasting Groups method or the Borderline Group method. The essential difference with the two latter methods is that here all panel members judge the same collection of work samples, in such a way that group discussion between rounds makes sense. Typically the Body of Work method (BoW) needs two rounds, although the need may be felt to add a third round.
- The scores of the sampled works are not known by the panel members.
- To convert panel judgments into cut-off scores, one has to take recourse to a special technique, called logistic regression. The reason for this is that the sample of works used is highly selective, such that applying the usual methods (e.g. taking the midpoint between averages as in the Contrasting Groups method) may lead to serious biases.

In the remainder of this section some details are given on the organisation of the method (Section 6.6.1), and on the statistical analysis technique required (Section 6.6.2). More detail can be found in Kingston et al (2001) and in Cizek and Bunch (2007, Chapter 9).

### 6.6.1. Training, Rangefinding and Pinpointing

These three terms refer to different phases in the procedure but at the same time to different samples of work to be used. To be concrete, it will be assumed that standards have to be set for A1/A2, A2/B1 and B1/B2, and that the panel consists of 15 members.

The training materials consist of a fairly small sample of work samples, carefully selected so as to cover a broad range of scores and levels. In the example it would be worthwhile to select two or three cases at each of the Levels A1, A2, B1 and B2, and to try to select the work samples in such a way that they represent the

substantial variation in the scores obtained within the level. For this selection, one can rely on expert judgments. For the training phase itself, the reader is referred to Chapter 5. Kingston et al insist that work samples with unusual or conflicting score patterns be avoided, e.g., work with some very high scores on some constructed response items and very low scores on other but similar items.

After the initial training a first round of judgments is organised, called rangefinding. The material presented to the panel members is a sample of students' work representing the whole range of obtained scores. The sampled work is presented in a number of folders, and each folder contains a small number of work samples. The work samples within a folder all have similar scores with few variations. The work samples within a folder are presented in an increasing order of score. The folders are presented also in increasing order of the scores of the work samples they contain. For a test with a maximum score of 55, one could prepare 10 folders with three works per folder, such that works with 30 different scores are represented to all panel members.

**Table 6.6: Summary of the Rangefinding Round**

Folder	Score	A1	A2	B1	B2	Total
1	13	15	0			15
	15	15	0			15
	16	14	1			15
2	18	13	2			15
	19	11	4			15
	21	9	6			15
3	23	10	5			15
	24	7	8			15
	26	5	10			15
4	27	3	10	2		15
	28	0	12	3		15
	30	1	11	3		15
5	32		9	6		15
	33		11	4		15
	34		8	7		15
6	35		7	8		15
	36		8	7		15
	37		6	8	1	15
7	39		3	12	0	15
	41		1	14	0	15
	42		1	12	2	15
8	43			10	5	15
	45			11	4	15
	46			8	7	15
9	48			4	11	15
	49			1	14	15
	51				15	15
10	52				15	15
	53				15	15
	54				15	15

The task for each panel member is to assign each work sample to one of the categories of the CEFR; in the example to A1, A2, B1 or B2. After this, the judgments are collected and staff members prepare a frequency table of the judgments given as exemplified in Table 6.6. From this table one can deduce useful information to reduce the amount of work in the second round of judgments.

- For the work samples in folder 10, the judgments are unanimous (B2), so that it can safely be assumed that the standard B1/B2 will be lower than a score of 52, the lowest score in folder 10. Similarly for

folder 1, where there is almost unanimity for category A1, it may be deduced that the standard A1/A2 will be higher than 16.

- Standards where panel members disagree most are likely to be found between adjacent categories. For the standard A1/A2 this is at score 24 (folder 3), for A2/B1 it is for the scores 34 and 35 (folders 5 and 6) and for B1/B2 the most disagreement is found at score 46 (folder 8).

These scores indicate the approximate value of the standards, and to avoid unnecessary work for the panel members in the second round, new folders are composed consisting of works with scores in the neighbourhood of these provisional standards. For the example in Table 6.6, work samples with scores in the range 21–27 for A1/A2, in the range 32–38 for A2/B1 and in the range 42–48 for B1/B2 may represent a suitable choice. These work samples may be collected in six folders, say, of three or four work samples, and these are the samples to be judged in the same way as in the first round. This second selection pinpoints the samples under study to narrower ranges than in the first round; hence the name pinpointing for the second round.

The work sample to be judged in the second round may consist in principle of entire new work, or only of the same work that has been used in the first round, or of a mixture of old and new work. The decision about the precise mixture will depend mainly on the time needed to go through completely new work, but as a general principle it is advisable to try to compose an even mixture of old and new work. New work creates the opportunity to judge on the generalisability of the procedure and inclusion of old work allows one to evaluate the consistency of the panellists' judgments.

### 6.6.2. Calculating the Standards: logistic regression<sup>20</sup>

The technique used to calculate the standards is called logistic regression. Like in all regression applications there is a dependent variable and one or more independent variables. In this application there is only one independent variable: the score on the test. The dependent variable is the judgment of the panel members, which can take only two different values for a particular standard, say A2/B1: the work has reached the standard (symbolised by a value of one) or not (value of zero). The regression model applied, however, is not the usual linear model between independent and dependent variables, but a linear model between independent variable and *the logit of the probability of getting a '1' on the dependent variable*. With a formula, this is given as

$$\ln \frac{p}{1-p} = a + bs$$

Where 'ln' symbolises the natural logarithm,  $s$  is the score on the test, and  $a$  and  $b$  are the two regression coefficients to be estimated. The symbol  $p$  stands for the probability of reaching the standard. Of course, this probability is not known, but we can approximate it by the proportion of panel members having judged that the standard is reached.

In Table 6.7 the results for the second round are displayed for the seven works around the provisional standard A2/B1. Notice that to compute the proportions, one has to take into account *all* cells indicating that the standard has been reached. In particular, for the score of 38, 10 panel members have indicated Level B1 and one panel member has indicated B2, making a total of 11 out of 15, leading to a proportion of  $11/15 = 0.733$ .

The regression analysis to be carried out is a simple linear regression analysis where the independent variable is the score and the dependent variable is given by the rightmost column in Table 6.9. If this table is contained in an Excel spreadsheet, the regression analysis can be performed directly in Excel.

---

<sup>20</sup> The technique discussed in this section uses the general approach of logistic regression, but the way the coefficients are estimated is not what is usually done in logistic regression techniques. However, the technique as presented here is easier to understand and its results are useful.

**Table 6.7: Results of the Pinpointing Round (partially)**

Score	A2	B1	B2	$p$	$\ln[p/(1-p)]$
32	10	5		0.333	-0.6931
33	11	4		0.267	-1.0116
34	9	6		0.400	-0.4055
35	7	8		0.533	0.1335
36	8	7		0.467	-0.1335
37	6	9		0.600	0.4055
38	4	10	1	0.733	1.0116

The estimates of the regression coefficients are

$$a = -10.3744 \text{ and } b = 0.29358$$

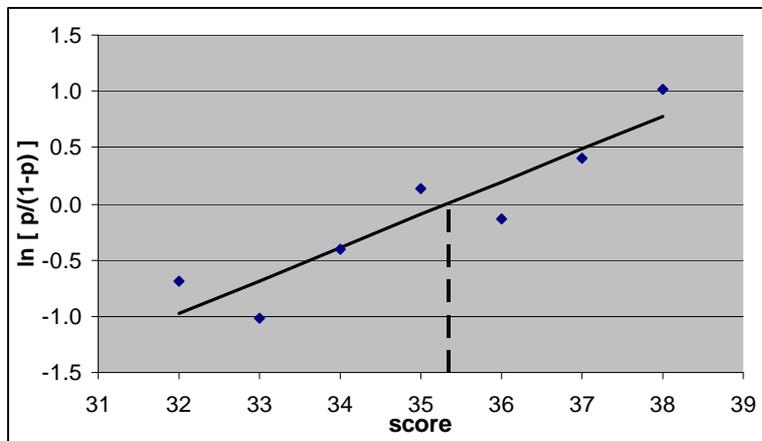
The final step is to compute the standard itself from these two coefficients. The cut-off score is conceptualised as the score where the probability of reaching the standard is exactly 0.5 and the logit of  $p = 0.5$  is  $\ln[0.5/(1-0.5)] = \ln(1) = 0$ . So, we look for the score for which it holds that

$$\ln \frac{0.5}{1-0.5} = 0 = a + bs,$$

from which it follows immediately that

$$\text{cut-off score} = \frac{-a}{b} = \frac{10.3744}{0.29358} = 35.34,$$

which will be rounded to 35 or 36. In Figure 6.2, the seven data points (from Table 6.7) are displayed graphically together with the regression line. The cut-off score is to be read on the horizontal axis at the point where the regression line crosses the zero grid line, as indicated by the vertical dashed line.



**Figure 6.2. Logistic Regression**

### 6.7. The Item-descriptor Matching Method and the Basket Method

In their book on standard setting, Cizek and Bunch (2007) give the following comment in the introduction to the Item-descriptor Matching method (p. 193):

“Performance level descriptors (PLDs) form the foundation of many modern standard setting methods, and are one of the key referents that participants rely on when making whatever judgments a particular method requires.”

And further:

“In a sense, it may not be an exaggeration to claim that standards are set more by the panels who craft the PLDs than by those who rate items or performances. This claim is most defensible under two very common conditions:

1. when PLDs are highly detailed and include very specific statements about examinee abilities at the given performance levels; and
2. when a standard setting panellist in the course of making a judgment about an item or task in a test relies – as he or she should – on the PLDs for a dispositive<sup>21</sup> indication of how performance on the item or task relates to the performance levels.”

In the CEFR, the performance levels are A1 to C2 (or more refined ones), and their descriptions are the “Can Do” descriptors, placed in a proper context and possibly further elaborated by benchmark examples. The preceding chapters, describing the necessary activities for the panellists to undertake in preparing for their rating task, as well as the detailed test specification can both be considered as an example par excellence of the fulfilment of the above mentioned conditions.

The two methods to be discussed in this section directly use these PLDs to arrive at one (or more usually) several cut-off scores. They are discussed in turn.

### 6.7.1. The Item-descriptor Matching Method

The method is relatively young; it has been proposed by Ferrara, Perie and Johnson in 2002<sup>22</sup>. The judgmental task asked from the panel members is to put every item in the level category (A1, A2, etc.) where it belongs according to the following requirement: “To which performance level description (i.e. CEFR level or category) are the knowledge, skills and cognitive processes required to respond successfully to this item most closely matched?” (Ferrara, Perie and Johnson 2002, p. 10).

From this quotation it is immediately seen that this method is test-centred. The task for the panellists is to assign a level for each item. The authors present an ordered list of the items (together with a short description). The order is an increasing order of difficulty, and an index of difficulty is given. Such a list is called an *ordered item booklet* (OIB) in standard setting literature. The method has been developed for cases where an IRT analysis has been used to estimate the difficulty parameters of the items.

The procedure to convert these judgments to a cut score (for each panel member) uses the important concept of a *threshold region*, which will be explained with the help of an example. In Table 6.8 a fictitious example of a judgment form is given for a test which is deemed suitable to set the standards A2/B1 and B1/B2. The form is a bit abridged because the item descriptions are left out. The rightmost column contains the judgments of a panel member. The column labelled “difficulty” contains a difficulty parameter estimate from an IRT-model used. The higher the numbers, the more difficult the item is. The column labelled “item-ID” identifies the item in the test, so that it can be looked up during the judgment procedure.

We will assume that all the judgments of this panel member for the items 1 to 10 were either A1 or A2 and that after item 21 no such judgments appear. One can see from the table that according to the judgment of the panel member there is no sharp cut between A2 and B1 items that is consistent with the ordering in difficulty: items 15 and 18 are judged to match with Level A2, although there are easier items, which are judged B1. The range of items, which are preceded by a clear (i.e. steady and unambiguous) sequence of judgments at the lower level and followed by a clear sequence of judgments at a higher level, is called the *threshold range*. In the example, this range contains the items 14 through to 18.

The basic idea of the method is that the threshold range, and the corresponding range of the underlying variable (the latent variable), indicate a region where the cut-off point has to be located. For the underlying variable, the difficulty parameters may be used, so that the cut-off point is to be located somewhere between -1.63 and -1.20. The midpoint might be a reasonable choice. Of course, every standard setting must deliver a

---

<sup>21</sup> “Orderly assigned”.

<sup>22</sup> In fact the method was presented as a conference paper at the 2002 meeting of the American Educational Research Association in New Orleans with the title *Setting performance standards: the item descriptor (ID) matching method*.

cut-off score in the score domain. So the cut-off point on the underlying variable must be converted to a cut-off *score*. This conversion is technically quite involved, and is discussed in Section 6.8.3.

As to the definition of the threshold range, the authors of the method propose that the starting point is the item that is preceded by at least three consecutive judgments at the lower level. In the example this is true because items 11, 12 and 13 are judged to match A2 and the end point is the item number which is immediately followed by at least three consecutive judgments at the higher level (the items 19, 20 and 21 are judged to match B1).

**Table 6.8: Example of an ID Matching Response Form (abridged)**

Rank number	Item-id	Difficulty	Judgment
...	...	...	...
11	22	-2.13	A2
12	13	-2.11	A2
13	7	-1.84	A2
14	1	-1.63	B1
15	4	-1.48	A2
16	8	-1.47	B1
17	3	-1.32	B1
18	17	-1.20	A2
19	15	-1.06	B1
20	9	-.97	B1
21	19	-0.94	B1
...	...	...	...

For applications in relation to the CEFR, the success of this method seems to depend quite critically on the tight relation between difficulty of the items and the level of the items. Ideally, one would say that an item that only requires abilities and skills described at Level A2 is easier than an item developed for Level B1. This, however, might be too simplistic a view for a sound theory on item difficulty. If there is great variability in difficulty within the levels attributed to the items in such a way that many of the hardest items from a lower level are more difficult than the easy items from a higher level, this will cause very broad threshold ranges, and make the intuitive appeal of the method disappear.

### 6.7.2. The Basket Method

A method that has many similarities to the Item-descriptor Matching method was used for the standard setting in the Dialang project (Alderson 2005) and is presented in Section 5.6 in Training for Standard Setting. The similarity is that it also requires a comparison of the demands of an item in terms of the PLDs, i.e. in terms of the “Can Do” descriptors of the CEFR. The basic question asked from the panel members, however, is not a judgment on the items but focuses on an abstract examinee, having capacities at a certain level. The basic question to be asked can be phrased as follows:

*“At what CEFR level can a test taker already answer the following item correctly?”*

If the scope of the test is broad, for example covering all levels from A1 to C2 as was the ambition in DIALANG, the same question has to be asked for each item at each level. Although such a procedure certainly has advantages in order to investigate the validity of the method and its outcomes (see next chapter), it is quite time consuming and this might have adverse effects on the motivation of the panel members.

Therefore a shortcut method was devised. The panel members are asked to put each item in a basket corresponding to one of the CEFR levels. If an item is put in basket B1, this means that a person at that level

should be able to give a correct response to this item. Here it is assumed that, if this is the case, persons at higher levels should also be able to give the correct response. Notice that this judgment does not imply that persons at a lower level should *not* give the correct response; it only means that (in the eyes of the panel member) a correct response should not reasonably be required at lower levels.

Notice that the task for the panel members in this abridged method is logically the same as in the Item-descriptor Matching method. In both methods a matching has to be found between a PLD (a CEFR-level) and the requirements implied by the items. In the Basket method, however, no information on the difficulty of the items is given to the panel members.

The method to convert judgments to cut-off scores was based on the reasoning that through the outcome of the Basket method, the panel member sets minimum requirements for each level. Suppose that for a 50 item test, two items are placed in Basket A1, seven in Basket A2 and 12 in Basket B1, then it follows that according to this panel member,  $2+7+12 = 21$  items should be responded to correctly by any one who is at Level B1 or higher. This number, the minimum requirement, is interpreted as the cut-off score.

A small technical note is in order here: it may be the case that a panel member judges that an item is so difficult that it cannot reasonably be expected to obtain a correct response at the highest level. For the procedure this means that the item does not fit in any of the baskets provided. One can anticipate such a situation by adding an extra basket with the label “higher than [C2]”. Of course, if a test aims at Level B1, it is not necessary to provide baskets explicitly for all levels. The three highest ones could be labelled as “B1”, “B2” and “higher than B2”.

It may be that the equating of the minimum requirement and the standard leads to standards that are too lenient. It might be reasonable to expect that a person at some level will also be able to answer correctly some items which are required at a higher level. This is not taken into account in the method, but some comparative studies (not published yet) show that the Basket method tends to produce lower (more lenient) standards than other methods.

This section is concluded with some remarks.

- Both methods discussed in this section are rather recent and reflect the importance of the Performance Level Descriptors (PLDs), which in the case of the CEFR are operationalised as “Can Do” descriptors. It is difficult to imagine that either of these methods can be meaningfully applied in case of pass/fail standard setting. The reason is that for each performance level (A1, A2, ...) the performance is described in a positive sense (what one can do), while it is not easy to describe in a positive way what a person deserving to fail an examination is able to do.
- In principle both methods can be used for binary items (such as MC items, yielding either a right or a wrong answer), and for constructed response items or tasks, (yielding a partial credit in the range 0–2 or 0–3, for example), which are more likely to occur with productive skills. One should, however, not underestimate the burden of work implied for the training phase in this case. Suppose that for a speaking task, a student can earn up to three points. This task will then appear in the list of items/tasks three times, the first time as a task-response combination leading to a score of 1 point (rather than zero), the second time as a task-response combination leading to a score of two points (rather than zero or one, but not enough for three points) and a third time, leading to the full credit of three points. In the three cases the task description will of course be the same, but the quality of the responses will differ. To ascertain a good understanding of these differences, one should refer to the rubric of the task (which is part of the test specification), and probably add sample answers that illustrate the intended use of the rubric. This illustrates the necessity of good rubrics: one cannot expect good standard setting using a rubric that says: “zero points for a bad answer, one point for an answer that is not too bad, two points for an answer that is a bit better and full credit for a perfect answer”. To select good sample answers (local benchmarks), one has to make sure that the judges giving the marks also have a good understanding of the rubric and have followed them strictly. This illustrates the fact that the whole process of constructing a test or an examination, from the first step (defining the purpose of the test) until the last step (setting the standards) is a long chain of interrelated decisions. As the standard setting is logically the last step, carelessness in

one or more of the earlier steps is likely to show up in the fact that the standard setting procedure “does not seem to work”.

- In their discussion of the Item-descriptor Matching method, Cizek and Bunch state that the items should be presented to the panellists in increasing order of difficulty, and, moreover, that an index of difficulty should be provided (as in Table 6.8). It is important to notice that for the task given to the panellists, these indices are not used. They only become important when the judgments of the panellists are to be converted to a cut score, but this conversion is usually not done by the panellists themselves, but off-line by staff members conducting the standard setting procedure. This conversion will be discussed in Section 6.8.3. It may even be advisable not to present such numerical values, because they can easily be misinterpreted, and may divert the attention of the panellists from their main task: the match between the requirement of the items and the descriptor(s) of a CEFR level.
- Although the formal characteristics of the method are easy to implement (the judgment form is easy to develop, and one for the ID matching method can be downloaded from the website [www.sagepub.com/cizek/IDMform](http://www.sagepub.com/cizek/IDMform), it would be illusory to think that a “quick and dirty” application of the method will guarantee useful results. The success (in terms of validity, to be discussed in more detail in the next chapter) depends critically on three factors:
  - Firstly, the clarity and discriminative power of the descriptors.
  - Secondly, complementarily with the first feature, the degree to which panel members understand well the meaning of the descriptors. This implies thorough familiarisation with the CEFR itself and a good standardisation in the sense used in the preceding chapter.
  - The third prerequisite is that the items or tasks of the test or examination itself can be clearly and unambiguously described and understood in terms of specific level descriptors. Panellists have to understand clearly which “Can Do” statements do apply and which ones do not apply in each and every item or task.
- The latter requirement also makes clear why more than one round of judgments is strongly advised. A second round with normative data (prepared quickly between the first and second round) showing particular cases of disagreement and discussing them in small groups, is not meant to enforce unanimity, but to stimulate discussions which lead to a clearer understanding of the CEFR and the relation between the descriptors and the requirements of the items or tasks.

## 6.8. The Bookmark Method

The Bookmark method (Mitzel et al 2001) is gaining very rapidly in popularity in the US. Most of its ingredients have already been discussed in previous methods, except for one, which will be explained in more detail in this section. We start with an overview of the important features.

- The method is test centred and it is applicable for binary as well as for polytomous responses (constructed responses, CR).
- Panel members use the concept of a minimal competent person or borderline person. For multiple standards (as e.g. A1/A2, A2/B1 and B1/B2 for the same test) the procedure has to be repeated for each standard. The burden of work, however, is less than in the Tucker-Angoff method because of the next feature.
- Items or tasks are presented to the panel members in increasing order of difficulty. For CR responses the task appears several times in the list. For example, if 0, 1 or 2 points can be earned on a task, this task appears twice, once as an instance where one can earn 1 point and once where one can earn 2 points. The ordering of the items in difficulty is not trivial, and will be discussed in Section 6.9. Notice that this ordered presentation is also used in the Descriptor Matching method, discussed in Section 6.7.1. Items and tasks are physically prepared in the form of a booklet. Each page refers to an item (in case of binary items) or to a task-partial credit combination in case of constructed responses. The content of each page

will be described in more detail. In the standard setting literature, this booklet is referred to as the *Ordered Item Booklet (OIB)*.

- The concept of mastery of an item or a task. Mastery here is defined in probabilistic terms. If a person masters an item, one can expect that he/she will give the correct response with a rather high probability. The exact definition of “rather high probability” is in principle arbitrary, but in many cases it is set at  $\frac{2}{3}$ , although some authors prefer to set it at 50% and others at 80%. In standard setting literature this mastery criterion is referred to as the *Response Probability (RP)*. Panel members have to decide for an item if a borderline person (at the given standard) masters the item or not. For  $RP = \frac{2}{3}$  this means that they have to decide whether the borderline person will give the correct answer in at least two of the three cases. (If  $RP = 80\%$ , it is a correct answer in at least four of the five cases.) It is important to make sure that panel members understand the notion of RP very well, and special attention to this understanding should be given in the training phase. Although there is no strict rationale to choose a particular value for RP, the choice one has made has definite consequences on the standards that one will find. In general it holds that the higher the RP, the higher the standards will be.
- For task-partial credit combinations, the RP has a special meaning. Suppose the maximum score on a task is 3. If the partial credit equals one, the RP refers to the probability of obtaining a partial credit of *one or higher*. If the partial credit is two, it refers to the probability of obtaining *two points or more*. If the credit equals the maximum score the RP refers to the probability of obtaining it.

### 6.8.1. The Task for the Panel Members

The panel members are instructed to start with the lowest standard (e.g., A1/A2), and go through the booklet from easy to hard, and to decide for each item whether the probability of a correct answer is RP or higher. If the answer is affirmative, this means that the borderline person masters the item, from the viewpoint of the panel member. As the judgments start with the easiest items, it is to be expected that the answer will be affirmative for some items in a row, but that at a given item it will be judged that the borderline person does not master the item any more. Suppose this happens at item 11, then a bookmark (real or symbolic) is placed at that page. As soon as this happens, the panel member switches to the next higher standard (A2/B1 in the example), and continues the judgmental work from the item where he/she was.

If there are three standards, then in principle the work ends as soon as the third bookmark is placed, and this may be well before the last item. It is good practice, however, to urge the panel members to look at all items, and even to consider the possibility of replacing earlier placed bookmarks as they continue to proceed through the OIB.

In each round each panel member indicates his/her provisional standard in a table like the one displayed in Figure 6.3, for the case three standards have to be set. The cells with the page numbers have to be filled out by the panel members. It is preferable to let the participants indicate two page numbers as in Figure 6.3. The page numbers 11/12 for the standard A1/A2 mean that (in the view of the participant) a borderline person at the Level A1/A2 has a probability of RP or more to answer item 11 correctly, but not for Item 12.

The information collected after a round is collected by staff members to make overviews to be used in the next round or in a concluding session.

Round 1			
Standards:	A1/A2	A2/B1	B1/B2
Page numbers:	11/12	24/25	38/39

**Figure 6.3: Panel Member Recording Form for Bookmark Method**

## 6.8.2. Content of OIB Pages

Each page of the ordered item booklet contains the following information:

- The page number within the booklet. This number is placed in evidence (bold face) at the right upper corner of the page, since this is the position panel members have to refer to in their judgments.
- The position of the item in the test or the examination (upper left corner). If the easiest item in the test is item number 5, the left upper corner must state “item 5”, while the right upper corner will state “1”, as it is the easiest item and takes position 1 in the OIB. In the case of items with partial credits, a double reference is needed. For example “item 13-2”. This refers to item 13 earning a credit of two points. If three points can be earned on this item, there will be three pages referring to this item, by references “13-1”, “13-2” and “13-3” respectively.
- In the top centre of each page, the RP and the scale value at RP is stated in texts like the following ones:
  - For binary items: “*Achievement level required for a 2/3 chance to answer correctly = -1.84.*” The RP is set at 2/3, and the value of the latent ability to have a probability correct of 2/3 is -1.84. Section 6.8.3. explains how to compute this value.
  - For partial credit items (as with constructed responses) the text is: “*Achievement level required for a 2/3 chance to obtain a partial credit of 2 points or more = 1.38.*” This will appear on the page with the item reference “nn-2”. For the highest score on a partial credit item the addition “or more” is omitted.
- The text of the item (the question) and in addition to this:
  - For MC items, the response alternatives.
  - For partial credit items, the precise scoring rule (rubric) for obtaining the specified partial credit. It is advisable to also add in such a case the scoring rule to obtain one point less and one point more, in such a way that the panel member can see the differences in scoring at the same page of the OIB.
- The correct response(s):
  - For MC items, this will be the key.
  - For partial credit items, one or more sample responses earning the specified score may help the panel members to focus on the precise meaning of the scoring rule.
- Reference to a source book:
  - With a reading test, where several questions (items) are asked about a single text (=testlet), it is advisable to collect all the texts in a source book, e.g. with numbered passages, and to refer to the relevant passage in the lower right corner of the pages in the OIB.
  - With listening tests things are a bit more complicated so a computer for every panel member may appear indispensable, in order to allow panel members to listen to the spoken passages as they feel the need to do so.

## 6.8.3. Technical Details

*On the value of RP for the Bookmark method*

The Item-descriptor Matching method and the Bookmark method are developed in the context of IRT calibrated tests, and typically make use of the calibration results. We illustrate this for the simplest case of binary items, which are calibrated with the Rasch model. Details for the case of partial credit items can be found in Cizek & Bunch (2007, Chapter 10).

In the Rasch model, the item response function is given by

$$P(X_i = 1 | \theta) = \frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} \quad (1)$$

where  $\beta_i$  is the difficulty parameter of item  $i$ . (Its value is known from the calibration.) First consider the case where the ability equals the difficulty of the item, i.e.  $\theta = \beta_i$ , then we can write equation (1) as

$$P(X_i = 1 | \theta = \beta_i) = \frac{\exp(\beta_i - \beta_i)}{1 + \exp(\beta_i - \beta_i)} = \frac{\exp(0)}{1 + \exp(0)} = \frac{1}{1+1} = \frac{1}{2}$$

meaning that for a latent ability equal to the difficulty parameter of the item, the probability of a correct response is exactly 0.5, and conversely, if RP is set at one half, the required ability for mastery is equal to the difficulty parameter of the item.

If one sets the RP at another value,  $p$  say, then one needs to find a value of  $\theta$  such that

$$\frac{\exp(\theta - \beta_i)}{1 + \exp(\theta - \beta_i)} = p$$

The solution is given by

$$\theta = \beta_i + \ln \left[ \frac{p}{1-p} \right]$$

where ‘ln’ indicates the natural logarithm. If  $p = 2/3$ , we have  $(2/3)/(1/3) = 2$  and  $\ln(2) = 0.693$ , whence we find  $\theta = \beta_i + 0.693$ , and this is the scale value that is printed on the OIB pages as the value of the achievement level (see Section 5). Notice that to shift from an RP of one half to  $2/3$  (as a requirement for mastery), the scale value increases with 0.693 logits. If one sets the mastery criterion RP at  $3/4$ , the increase is  $\ln(3) = 1.098$ , and for an RP of  $4/5$ , the increase is  $\ln(4) = 1.386$ .

#### *The provisional standard for the Bookmark method*

As an example, some page numbers are displayed in Table 6.9, together with the achievement level for RP = 0.5 (second column) and for RP =  $2/3$  (rightmost column). The difference between the two latter columns is  $\ln(2) = 0.69$ . Assume that RP has been set at 0.5, and some panel member has put the bookmark for A1/A2 at 13/14. This implies that according to this panel member the borderline person masters (with an RP of 0.5) the items 1 through 13, but not item 14, which implies that the achievement level (latent ability) of the borderline person must lie somewhere between -1.84 and -1.63. Usually one takes the lesser value of these two as the provisional standard (of that panel member). Notice that this provisional standard is a value on the latent scale. To arrive at the group standard, the provisional standards are aggregated (by taking the (trimmed) average or the median), so that the group standard is also expressed as a value on the latent scale.

#### *Converting latent scale standards to cut-off scores in the Bookmark method*

The simplest way to convert standards on the latent scale to cut-off scores in the score domain is to use a table that gives good estimates of the latent value for all possible scores in the test. An example is given in Table 6.10. Suppose the standard on the latent scale is -1.35. From the table one sees that a score of 9 (items correct) leads to an estimated latent value of -1.409, smaller than the standard, while a score of 10 has an estimated value of -1.257, higher than the standard. This will lead to a cut-off score somewhere between 9 and 10, and this value has to be rounded, taking all considerations of false positives and false negatives into account, as explained in Section 6.3.4.

**Table 6.9: Bookmarks and Achievement Levels**

Page number	Achievement level for RP = 0.5	Achievement level for RP = 2/3
...	...	...
11	-2.13	-1.44
12	-2.11	-1.42
13	-1.84	-1.15
14	-1.63	-0.94
15	-1.48	-0.79
-----	-----	-----
19	-1.32	-0.63
20	-1.20	-0.51
21	-1.03	-0.34
...	...	...

**Table 6.10: Estimated Theta**

Score	Estimated theta
...	...
5	-2.153
6	-1.938
7	-1.746
8	-1.571
9	-1.409
10	-1.257
11	-1.114
12	-0.977
13	-0.845
14	-0.717
15	-0.592
16	-0.471
17	-0.351
...	...

An important question, however, is which estimate of the latent variable should be used. In Section G.7 of the Reference Supplement, several estimates are discussed, and it was shown that the maximum likelihood estimate can be seriously biased. Therefore it is advisable to use the Warm estimator, contrary to what Cizek and Bunch advise<sup>23</sup>. This is especially important if the standard in the domain score happens to be rather extreme, relatively low or relatively high.

*An extra problem with the Item-descriptor Matching method*

In the Bookmark method, the RP-value has to be introduced to the panel members and clearly explained. This is important, because the higher the RP, the stricter the standard will be, and panel members must be clearly aware of the meaning of the RP.

In the ID matching method, to the contrary, the concept of RP does not enter the game, because panel members only have to indicate at which level (A1, A2, etc...) each item fits best. From the difficulty level displayed in Table 6.9 (third column) one cannot deduce if these are the difficulty parameters or the achievement level for some other value of RP than 0.5. As was argued above, these numbers are of no use in the judgment task of the panel members beyond giving a clue that the items are ordered in difficulty. But once the *threshold region* has been determined, these numbers play a central role, because they are used to determine the provisional threshold (for each panel member), and ultimately to calculate the group standard.

We can see the problem in a thought experiment using two groups of well trained panellists. In one group the difficulty levels displayed for them equal the difficulty parameters as they have been found in the Rasch calibration; in the other group the difficulty levels are the difficulty parameters plus  $\ln(2)$ , corresponding to an RP of 2/3. Since the basic task of the panellists is to concentrate on a match between the requirement of the items relative to the CEFR levels, it can be expected that the threshold regions will not show systematic differences between the two groups of panel members, and will not be influenced by the magnitude of the numbers displayed for each item. But the standards computed from the difficulty values will differ by a value of approximately 0.693 ( $= \ln(2)$ ) in the two groups. More generally, this means that the standards arrived at are arbitrary to a large degree, depending on which values happen to be displayed as difficulty levels.

---

<sup>23</sup> In the literature it is advised to use the test characteristic function to convert latent values to scores. In the Rasch model and the two parameter model, however, this conversion is the same as using maximum likelihood estimates. Warm estimates are provided by default in the software package OPLM, which is available on simple request from [norman.verhelst@cito.nl](mailto:norman.verhelst@cito.nl)

## 6.9. A Cito Variation on the Bookmark Method

The Bookmark method may get more complicated if the items do not discriminate equally well (which is more often the case than not). A simple example with two items is displayed in Figure 6.4., where the dashed curve represents the best discriminating item. The two curves represent item response functions: they relate the latent ability (horizontal axis) to the probability of obtaining a correct response (vertical axis).

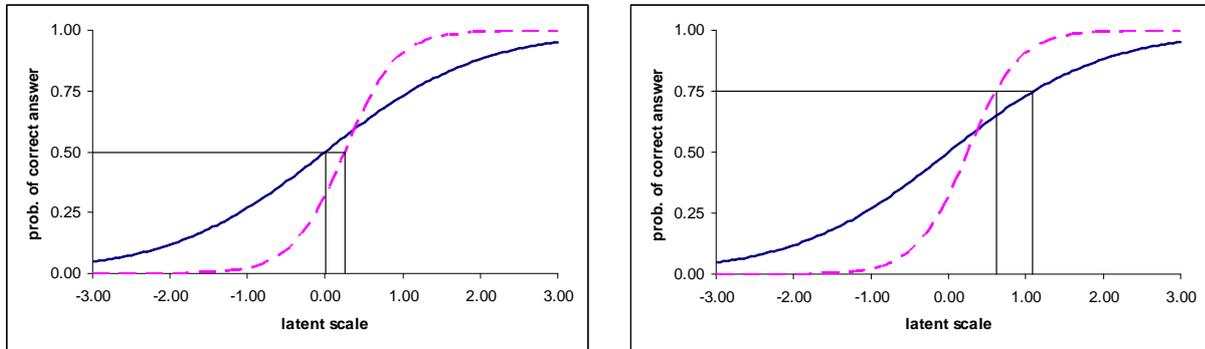


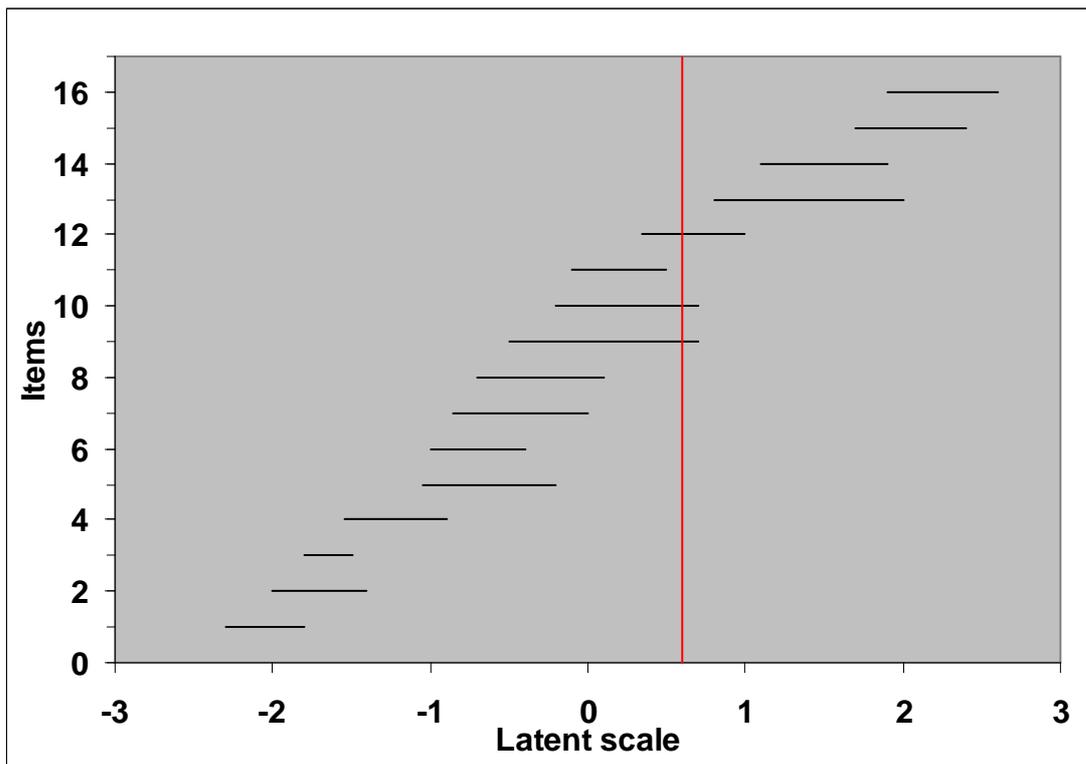
Figure 6.4: Items with Unequal Discrimination

If one uses the Bookmark method with  $RP = 0.5$  (left-hand panel), the dashed item will have a higher page number (being presented as the most difficult of the two) in the OIB than the other item, while with an  $RP$  of  $0.75$  (right-hand panel), the reverse holds: the dashed item will now appear as the easiest of the two. This illustrates the fact that “difficulty of an item” is not a simple concept, and presenting the ordering of the difficulties by a simple number may confuse panel members.

The method developed at Cito (Van der Schoot 2001) aims at presenting in a graphical way difficulty and discrimination values of all items in a single display. Consider the least discriminating item in Figure 6.4: for  $RP = 0.5$  the required ability is  $0$ ; for  $RP = 0.75$ , the required ability is about  $1.1$ . One could designate a chance of  $50\%$  to get an item correct as “borderline mastery”, while a chance of  $75\%$  correct could be called “full mastery”. To go from borderline to full mastery the ability must increase from  $0$  to  $1.1$ . One can display this graphically in a figure like Figure 6.5, which is an item map for  $16$  items that contains information about the difficulty and discrimination of each item. Each item is represented as a piece of line, stretched horizontally. The left end corresponds to the difficulty parameter of the item ( $RP = 0.5$ ), and the length is indicative for the discrimination value: the longer the line, the less the item discriminates. The right end corresponds to a higher  $RP$ ,  $0.75$  or  $0.80$ , say. The display is constructed in such a way that the left ends of the item lines increase as one goes from bottom to top. One should take care that the lines are properly identified, such that panellists can associate each line clearly with an item in the test.

The vertical line symbolises the provisional standard of a panel member, and by drawing this line (or holding a ruler) the panel member can quickly have an overview of the consequences of his/her decision. In the example the proposed standard implies full mastery of the items  $1$  to  $8$  and of item  $11$ . For the items  $9$  and  $10$  there is almost full mastery. For item  $12$ , borderline mastery has been reached, and for the items  $13$  to  $16$ , borderline mastery is not reached at all.

To apply the method, the panel members can be asked to draw a vertical line, or to give a numerical value that corresponds to the location where the vertical line touches the horizontal axis in the figure (which is  $0.6$  in the example).



**Figure 6.5: Item Map, Indicating Difficulty and Discrimination**

Notice that from Figure 6.5 one cannot deduce in any way what the distribution of latent abilities in the target population looks like. To avoid all associations with, e.g. a normal distribution, it may be advisable to change the scale values that are displayed along the horizontal axis in the figure to a convenient scale having no negative values, and an easily understood unit. For example, adding 8 to all numbers displayed along the axis in the figure, and the multiplying by 10 will make the numbers range from 50 to 110, avoiding interpretations in terms of percentages, and being fine grained enough to require provisional standards expressed as whole numbers<sup>24</sup>. After the standard setting is completed, the resultant standards can easily be transformed back to the original scale, and standards in the score domain are determined in the same way as in the Bookmark method (see Section 6.8.3.)

## 6.10. Special Topics

In this concluding section a number of special topics will be touched upon briefly. These topics are:

- standard setting with heterogeneous tests (across skills);
- standard setting and test equating (across administrations);
- cross language standard setting.

### 6.10.1. Standard Setting Across Skills

In some settings the requirement might be to report one single, global result as to an examination candidate's CEFR level, while the examination itself may consist of three or more parts, with each of these subtests addressing a different skill.

<sup>24</sup> An alternative or supplementary approach might be to include CEFR descriptors relevant to the test tasks in the piloting or pre-testing as items for teacher-assessment or self-assessment and then, at an appropriate point in the series of standard setting rounds, to show panellists where they appear calibrated to the latent scale shown on Figure 6.5. See Section 7.5.4.2).

One can take different viewpoints regarding such a situation. Two viewpoints are discussed, a compensatory and a conjunctive approach. It is argued that both approaches, if applied to the extreme, can lead to unacceptable results; a reasonable solution in the form of a compromise is discussed as well.

**Compensatory approach:** On the one hand, as an extreme position, one could consider all tasks and items in the mixture of skills and apply any of the methods discussed above on the whole collection of items and tasks simultaneously. In proceeding this way one must realise that test scores are per definition compensatory in nature, since they are sums of item and task scores. Failing on some tasks may be compensated by good performance on other tasks. As long as the test is homogeneous with respect to the nature of the tasks, such a compensatory mechanism is quite natural, and one does not have to be concerned with the precise items and tasks that are solved or failed.

However, with a more heterogeneous test, this compensatory viewpoint may not be adequate. For example, suppose that a certain national examination for English consists of a reading test, a listening test, a speaking test and a writing test, with a maximum score of 100 points on the four parts taken together. Suppose further that the Body of Work method is applied to set standards and that care is taken to collect work samples from different regions in the country. If regions differ markedly in their teaching investment and/or expertise for one or more of the skills, typical profiles on skills may show different patterns across regions. If in some region little attention is paid to speaking, even the best students may be characterised by poor speaking and perform at the same level as the average student in regions where sufficient attention has been given to this skill. Taking all skills together would hide possibly important differences in profiles.

Therefore it is important that a thorough study is undertaken to investigate the extent to which a unidimensional approach is appropriate. In addition to studying the structure of the different skills, possible differences in structure between schools, regions or instruction methods used, surfacing as differential item functioning (DIF), would have to be examined before a unidimensional approach could be justified. If there are in fact marked differences or only moderate correlations between the skills, one has to face several problems, two of which we mention here:

1. A rational decision has to be made on the weighting with which each skill is represented in the total score. If there is some legal provision that says for example that each skill is equally important, then this problem is solved.
2. But even with an imposed weighting, one has no guarantee that in examinee centred methods such as the Body of Work method, the panel members will indeed use this imposed weighting when they have to come to a holistic judgment of the student's level.

**Conjunctive approach:** The alternative is an approach that takes each skill separately, which implies that standard setting is carried out for each skill separately. The conjunctive decision rule states that one has globally reached a certain level only if one has reached that level for each skill. Applying this rule in all its rigidity may lead to unacceptable results, as a student may not be granted Level B1, even if he has reached B2 in three of the four skills but not the A2/B1 standard for the fourth.

A compromise between compensatory and conjunctive rules may seem reasonable in this context: a general conjunctive rule may be set with some compensatory exceptions, as in the example just mentioned, where it may be reasonable to grant Level B1. The exact nature of the compensatory exceptions must be considered with care, and a good approach is to discuss them with the panel members after they have set the standards for each skill separately.

### 6.10.2. Standard Setting and Test Equating

As standard setting is a rather expensive undertaking, it may be worthwhile to investigate possibilities to avoid a lot of the work, certainly in cyclical examinations where the test specification tends to be repeated from year to year without major modifications.

If careful standard setting has been carried out for one year's form of the examination, the results of the standard setting may be transferred as it were to a new examination form (e.g. for the following year) by

applying a technique called test equating<sup>25</sup>. Loosely speaking, test equating designates a collection of techniques in which for each score in one test an equivalent score in the other test is determined. Suppose the standard A2/B1 has been set for the first year's examination at a score of 35. If the equivalent score of 35 on the second year's examination is 37, this automatically entails that the cut-off score for this year is 37.

Applying test equating has two aspects which must be carefully taken into account. The first is of an almost purely technical nature, the second is conceptual.

To apply equating techniques, it is essential that the two samples of students taking each examination are comparable in some way. Such comparability may be ensured either by using common items in both examinations or by taking measures such that the two samples are statistically equivalent. Neither approach can be implemented easily in an examination context: usually it is not possible to repeat last year's examination in the current year for reasons of secrecy, and equivalence of samples is difficult to obtain, since students usually cannot be assigned randomly to either examination. A slightly more able population this year may cause this year's examination to look easier than it is. If this is not recognised, and the two populations are considered as equally able, this will lead to strict standards.

Using IRT techniques similarly requires that the two examinations are anchored in some way, meaning that parts of both examinations have been administered to a sample of students. (See Section G of the Reference Supplement for more details; see also Section 7.2.3.)

The conceptual issue has to do with the construct validity of both examinations. Although using the same specification is a reasonable measure to obtain equivalent constructs, it may not be sufficient, as nobody has a complete understanding of the composition of the constructs measured by a language examination. Techniques to investigate the dimensionality of a complex test such as Factor Analysis (Section F of the Reference Supplement) may offer a solution here.

But the safest way to guarantee the validity of transferring standards by equating is to carry out standard setting on the new examination anyway, to check whether the standards obtained by transferring them through test equation do indeed correspond to the standards set by an independent panel of expert judges.

### **6.10.3. Cross Language Standard Setting**

Perhaps the most challenging aspect of linking examinations to the CEFR is to find methods that show that examinations in different languages are linked in a comparable way to the common standard.

Although it might be theoretically possible to administer two examinations in different languages to the same sample of students, this would presuppose that each student in the sample has the same level of competence in both languages, which clearly would be nonsense. Therefore, methods must be looked for which assume that any student has taken only one of both examinations, treating each student's performances in different languages as those of unrelated candidates.

To link both examinations to the CEFR, use can be made of plurilingual panel members, who can give trustworthy judgments either on the items (in test centred methods) or on students' work in both languages. The Body of Work method may be a good candidate in the latter case. For test-centred methods, any method that does not presuppose IRT calibration can be used in principle. IRT based methods will not work because of the impossibility to scale both exams on the same scale, because the design will not be linked through common persons (see above) or common items.

As such cross-linguistic standard setting is relatively new<sup>26</sup>, care must be given to a number of threats to the validity of the procedure. Here are some topics to pay attention to:

---

<sup>25</sup> A good introduction can be found in Kolen & Brennan (2004).

<sup>26</sup> A cross-language CEFR benchmarking seminar was hosted by the CIEP in Sèvres 23–25 June 2008. At this event, samples of French teenagers speaking English, German, French, Italian and Spanish were rated onto the CEFR levels in multilingual teams. A report is available on the Council of Europe website ([www.coe.int/lang](http://www.coe.int/lang)).

- As it is impossible to hide the language of the test to the panel members, thus excluding “blind” judgments, it is important that there are not too many systematic differences in construct between the two language tests, to avoid the panel members using different constructs for the two languages. Therefore, care must be taken that the two examinations or tests have the same or very similar specifications.
- Care must be taken in the composition of the panel to have a well balanced design with respect to expertise in both languages. If the two tests are in English and French, care must be given to the language and training background of the panel members. For example half of them may be native in English, the other half in French, or a balance must be sought for the main rating task: half of the panel members being teachers of French with sufficient proficiency in English, and vice versa for the other half.
- This balance must be maintained in sub-groups of the panel formed for discussion.
- In a similar way, the material to be judged (be it performance samples or items) should be presented in a well balanced way with respect to sequence of presentation as well as with respect to content.
- Steps must be taken during standardisation training to ensure that that all members of the panel apply the same standard to each language. There is a danger of systematic distortions caused by the traditions, reference publications and terminological differences associated with different pedagogic cultures. It is vital that panel members use and refer to the official criteria and benchmarks – not preconceived internal standards.
- Detailed records of the procedure must be kept, and if possible, results from the bilingual standard setting procedure should be compared with the results of a monolingual procedure in either language, collected with an independent set of panel members.

## 6.11. Conclusion

This chapter has given an overview of a number of standard setting procedures, but it pretends in no way to be exhaustive. A comprehensive overview can be found in Section B of the Reference Supplement and additional procedures exploiting teacher judgments and IRT to incorporate an external criterion into the linking study are presented in Extra Material provided by Brian North and Neil Jones. The accent in this chapter has been put on the feasibility and appropriateness of the selected methods for language testing and for linking to the CEFR by stressing a good understanding of the basic notions.

Of course, during and after the application of the procedures, one needs quality monitoring focussed on several questions:

- Has the procedure of standard setting had the effects as intended: was the training effective, did the panel members feel free to follow their own insights? and similar questions are relevant here. These are questions of procedural validity.
- Are the judgments of the panel members to be trusted: is each panel member consistent with himself/herself across the various tasks he has carried out; are panel members consistent with each other in their judgments and to what extent is the aggregated standard to be considered as the definite standard, or do they have some measurement error just like test scores? These questions and their answer constitute the internal validity of the standard setting.
- The most important question, however, is whether the results of the standard setting – allocating students to a CEFR level on the basis of their test score – is trustworthy, and the basic answer to this question comes from independent evidence which corroborates the results of a particular standard setting

procedure. It is the task of everyone applying such a procedure to provide an answer to that question, and this is precisely what is meant by validation. Such evidence may come from different sources, such as:

- **Cross Validation:** repeating the standard setting procedures with an independent group of panellists;
- **Complementary Standard Setting:** carrying out independent standard setting using a different procedure that is appropriate to the context;
- **External Validation:** conducting an independent study to verify the results of the standard setting by referencing them to an external criterion. This external criterion might be a test for the same skill(s), known to be reliably calibrated to the CEFR. However, it might be judgments of teachers or learners trained with CEFR descriptors.

All these issues are considered in Section 7.5.

*Users of the Manual may wish to consider:*

- *whether specialist support or further reading on standard setting is needed*
- *what method(s) is/are the most appropriate in the context*
- *whether to adopt a method judging the difficulty of individual items (e.g. Descriptor matching or Basket methods) or a method judging the cut score on the reporting scale for the trial test (e.g. Bookmark, Body of Work methods)*
- *whether two methods might be used to validate each others' results*
- *how panellists will be given "normative feedback" on their behaviour after the first round; is electronic voting<sup>27</sup> feasible?*
- *whether IRT difficulty estimates will be available to inform the standard setting process or whether p-values will have to be used*
- *what sort of "impact data" on the effects of provisional standard setting might be made available to inform later rounds*
- *what support may be needed in applying the chosen methods*

---

<sup>27</sup> For information on the application of electronic voting, see Lepage and North (2005).

# **Chapter 7**

## **Validation**

### **7.1. Introduction**

### **7.2. Pre-requisites: The Quality of the Examination**

- 7.2.1. Content Validity**
- 7.2.2. Operational Aspects: The Pilot**
- 7.2.3. Operational Aspects: The Pretest**
- 7.2.4. Psychometric Aspects**
- 7.2.5. The Timing of the Standard Setting**

### **7.3. Procedural Validity of the Standardisation Training and Standard Setting**

### **7.4. Internal Validity of the Standard Setting**

- 7.4.1. Intra-judge Consistency**
- 7.4.2. Inter-judge Consistency**
  - 7.4.2.1. Agreement and Consistency**
  - 7.4.2.2. Three Measures of Agreement**
  - 7.4.2.3. Evaluating Indices of Agreement**
  - 7.4.2.4. Finding Problematic Items**
  - 7.4.2.5. Indices of Consistency**
- 7.4.3. Accuracy and Consistency of the Standard Setting Method**
  - 7.4.3.1. Standard Error of the Cut Score**
  - 7.4.3.2. A Paradoxical Situation**
  - 7.4.3.3. Accuracy and Consistency of Decisions**

### **7.5. External Validation**

- 7.5.1. Cross Validation**
- 7.5.2. Comparison of Marginal Distributions**
- 7.5.3. Bivariate Decision Tables**
- 7.5.4. Some Scenarios**
  - 7.5.4.1. Taking Advantage of IRT Calibration**
  - 7.5.4.2. Using “Can Do” Statements**
  - 7.5.4.3. Cross Language Standard Setting**

### **7.6. Conclusion**

## 7.1. Introduction

Linking an examination to the CEFR is a complex process involving many steps, which all require a professional approach. Validation concerns the body of evidence put forward to convince the test users that the whole process and its outcomes are trustworthy. Test users are to be understood in a very broad sense; they range from students (or their legal representatives, like parents) taking the test, educational and political authorities using test results for policy decisions, textbook developers and teachers, testing agencies, employers and trade unions, the scientific community involved in language testing, and if the stakes are really high, also legal authorities. Although the present Manual focuses on the linking process in a rather strict sense, culminating in the application of one or more standard setting procedures, it would be mistaken to assume that the validation process can be restricted completely to the activities and outcomes described in Chapters 3 to 6. In the present chapter, most of the procedures and techniques to be discussed will also be focused on the linking process proper. However, a separate section (7.2.) will be devoted to general prerequisites, pertaining to the quality of the examination: content validity of the examination, the pilot, the pretest, some psychometric aspects, and the timing of standard setting.

The discussion about the validity will be organised thereafter in three separate sections, two of them dealing with the validity or trustworthiness of the procedure itself and its basic outcomes. In Section 7.3, **procedural validity** will be discussed and in Section 7.4, the **internal validity**, to be understood as internal consistency, will be given attention to. In Section 7.5, finally, the most important and most difficult part of the validation process, **the external validity**, is focused upon. In general, external validity refers to all *independent* evidence that other methods come essentially to the same conclusions as the methods and procedures in the current study.

Validity is not a question of all or nothing, but a matter of degree. A report on validity will require attention to the many facets involved, putting forward well considered arguments and empirical evidence to underpin any statements and claims to generalisability. For this reason, it is indispensable for a good validation study to have all activities carefully documented.

The end of the chapter will conclude the Manual with some reflections on the state-of-the-art in standard setting and a brief outlook on the future.

## 7.2. Pre-requisites: The Quality of the Examination

Linking a qualitatively poor examination to the CEFR is a wasted enterprise that cannot be saved or repaired by careful standard setting. In this section a number of important aspects of the examination itself will be reviewed briefly from the perspective of a good linking process. They refer to the content of the examination, its operational and its psychometric aspects.

### 7.2.1. Content Validity

Usually the content of an examination is dictated by curricular prescriptions that leave limited room for manoeuvre. Although the CEFR “Can Do” statements are formulated in quite an abstract manner, it may happen that curricular requirements and the way the CEFR is articulated conflict. It may happen that some items in the examination are so complex that an unambiguous allocation to one of the CEFR levels is impossible, while on the other hand, taking away the ambiguity may conflict with curricular requirements.

To solve this problem, different viewpoints can be taken:

- The most extreme position is to abstain completely from the linking to the CEFR. Although it might not solve problems in the short term, publishing arguments may be helpful for a revision or extension of the CEFR, or for a revision of the curricular requirements to make them more compatible with the CEFR.

- A more nuanced approach might be to seek for a compromise and to base the linking on only part of the examination, leaving out for example 25% of the tasks and items used in the examination, because they are difficult to relate to CEFR categories or levels.
- An alternative would be to select a standard setting method which is less analytical, for which no reference to specific CEFR descriptors is necessary. Some standard setting methods rely on broad, holistic judgments (e.g. the Body of Work method: Section 6.6.), whilst others involve global judgments about where to place the cut-off between levels on a test, informed by a lot of psychometric information (e.g. the Bookmark method or its Cito variant: Sections 6.8.–6.9.).

Another aspect of the same problem is the extent to which the relevant activities and competences described in the CEFR are covered by the examination. The specification of the examination (Chapter 4) details what is included in the examination, but not what has been left out. Omission of important parts and aspects from the CEFR construct can lead to one-sidedness and make claims of generalisability in the linking unjustified. There exist methods to quantify the content validity of an examination and a practical example has been given by Kaftandjieva (2007). In order to avoid any danger of overgeneralisation, it is a good idea to state explicitly what the content coverage (content representativeness) of the examination is.

### 7.2.2. Operational Aspects: The Pilot

Before an examination is administered in a real examination context, data may be collected at several stages. Usually one distinguishes between piloting and pretesting.

A pilot is usually meant to try out the test material in order to eliminate ambiguities, to check on the clarity and comprehensibility of the questions and their rubrics, to have a first impression on the difficulty of the tasks and items and to estimate the time load involved. Such a pilot can be conducted on a small scale (one or two classes usually suffice), but it is useful not to present the material exclusively as a test, but to try to elicit as much feedback as possible about the quality of the test material. Qualitative methods such as interviews and cognitive labs<sup>28</sup> can reveal a lot of interesting information about the planned examination, and participants in such a pilot can be students and teachers. By good piloting unpleasant surprises at the time of the pretesting and the real examination can be avoided.

One aspect that is easily overlooked in the construction of itemised tests is the dependency between items. A test yields its maximum information about the construct to be measured if each item is a new and fresh opportunity for the test taker to show his or her ability or proficiency. An item  $i$  that can be answered correctly only if another item  $j$  has been answered correctly, or a construction where a wrong answer on item  $i$  entails a wrong answer on item  $j$  are extreme examples of dependency, usually called *functional dependency*. But more subtle forms of dependency can occur as in the case where working on an item  $i$  releases information about the correct answer on item  $j$  without being fully informative. Moreover, this information may be selective so as to be helpful only if the correct answer to item  $i$  has been found. This kind of dependency is called *statistical dependency*. Ignoring dependency may have awkward consequences for the psychometric characteristics of a test (such as leading to the inflation of the reliability coefficient) and also for the standard setting. Particularly in ambitious projects where a calibrated item bank is built, and an examination is constructed by selecting a set of items from the bank, dependency can have serious consequences. If items  $i$  and  $j$  have been administered jointly to collect the data for the bank calibration, and if there is statistical dependency between them, then the psychometric characteristics of either of them in isolation, only one of them being part of the examination, is unpredictable.

As the demonstration of statistical independence is not easy, it is certainly worthwhile to try and detect the subtle strategies test takers may use to relate one item to another during piloting. Well-constructed feedback from candidates during piloting is a good way of identifying any such problems<sup>29</sup>.

<sup>28</sup> A cognitive lab is a procedure where participants are invited to take the test whilst thinking aloud and making explicit the way in which they understood the questions, their strategy to answer and the different steps they take.

<sup>29</sup> For a statistical and psychometric treatment of dependencies using IRT, see Verhelst & Verstralen (2008).

### 7.2.3. Operational Aspects: The Pretest

A pretest is usually designed to get information on the main characteristics of a planned examination. Apart from psychometric features (to be discussed subsequently), operational characteristics should also be observed. A major source of information to be collected in this respect is the time allotted and needed for the pretest. Although the number of items towards the end of the test that candidates do not manage to complete may give useful information in this respect, at least two aspects go usually undetected:

- Students in need of time may pick the easy looking items to give a response. Especially if the examination is a mixture of multiple-choice and constructed response items, students may tend to pick the MC items as a strategy to collect the highest possible score. In such a case, non-response is difficult to interpret: it may be caused by the intrinsic difficulty of the items or by a time pressure strategy. A short questionnaire administered to (a subset of) the students or to the teachers after the pretest may be helpful in finding a reasonable explanation of non-response behaviour.
- It may happen that the total time allocated for the test is overestimated, causing a loss of information. A simple means to detect this is to ask the teacher to note for each candidate the exact time that he or she hands over the finished examination.

Apart from being a kind of rehearsal for the examination to come, pretesting also has a central function in linking examinations to each other. As examinations tend to be unique in composition from year to year and because the target populations have no students in common<sup>30</sup>, data from two examinations cannot be meaningfully compared: differences in the average score may have been caused by systematic differences between the two student populations or by a difference in difficulty between the two examinations or by any mixture of these two causes, and there is no way to find out to what extent both reasons apply unless the data are linked in some way.

Because presenting item material to the same students in a pretest and in the examination itself has unpredictable consequences due to memory effects, good practice will require that pretesting and linking is done two years (or periods) in advance of the examination proper. Supposing that the examinations for Year 1 and Year 2 are to be linked, then the pretest that links them will have to be organised two years in advance of Examination 2, i.e., in Year zero.

It is advisable to plan the pretesting in what is called a “balanced incomplete block design”. The item material for the two examinations (together with some reserves) is partitioned into a number of item blocks. Each student participating in the pretest takes the same number of blocks, but no student takes all of them. A balanced incomplete block design has the following characteristics:

- each block is presented to an equal number of students;
- each pair of blocks is presented to an equal number of students;
- each block of items occurs in each serial position.

To accomplish these requirements, restrictions have to be put on the number of blocks. Balanced incomplete designs are possible for 2, 3, 7 and 13 blocks, but not for other numbers lower than 13. For any of the given numbers, the number of different test forms to be prepared equals the number of blocks. Table 7.1 shows the design for three blocks and Table 7.2 the one for seven blocks. In Table 7.1, each student gets one of the three test forms. The numbers in the row for the test form indicate the content of the test and also the sequencing of the blocks. It is easy to check that the three requirements for a balanced incomplete block design listed above are fulfilled here, as well as in the design with seven blocks.

---

<sup>30</sup> Even if a student takes two forms of an examination (because of grade repetition, for example), one cannot assume that his or her ability is the same at the two examination moments, and in all psychometric analyses such a student is treated as representing two (statistical) students.

**Table 7.1: Balanced Incomplete Block Design with Three Blocks**

Test Form	Item blocks	
1	1	2
2	2	3
3	3	1

**Table 7.2: Balanced Incomplete Block Design with Seven Blocks<sup>31</sup>**

Test Form	Item blocks		
1	1	2	4
2	2	3	5
3	3	4	6
4	4	5	7
5	5	6	1
6	6	7	2
7	7	1	3

Care must be given not to administer the same test form to all students from one class or school, because systematic differences between classes or schools may possibly bias the estimates of the  $p$ -values of the items. In principle all test forms should be administered an equal number of times in each class. A practical way of implementing this principle is *spiralling*. The test forms are distributed in the class room in a fixed sequence: if the first student gets Form 4, the next gets Form 5, then 6, 7, 1, 2, 3 and the sequence is repeated. Do not start the sequence in each class with Form 1. The test form starting the sequence should be picked at random, or should be one higher than where the preceding class ended. All this requires good planning but it is a good safeguard against unforeseen biases, which are difficult to repair.

Use of a balanced incomplete block design has useful advantages for constructing the examination. Whatever subset of items is selected to be included in the examination of Year One, each item has been observed in conjunction with every other item. For the items of the examination in Year Two the same applies, as well as for the items not used. Every item from examination one is linked to every item of examination two. To obtain balanced contents in each of the test forms used, it is important to make each block as heterogeneous as possible with respect to content and to difficulty.

To continue this process, we look what happens in Year One. The examination for Year One is administered, but in the same year pretesting is necessary for the following two years. Applying the same principle as before, in the Year One pretest, item material for Year Two and Year Three is to be pretested, to guarantee the link between the examinations of Year Two and Year Three. So the material for Year Two has to be pretested again. This illustrates a basic principle: in order to have good year-to-year linking of examinations, the item material has to be pretested twice.

Having completed the above, it is important that a sufficient number of test takers provide responses for each item. Classical Test Theory is poorly equipped to handle data collected in incomplete designs, such that one

---

<sup>31</sup> If one considers the three columns of numbers one notices that they start at a certain value in the top row, then climb to 7 and restart from 1. The starting values for the columns are 1, 2 and 4. For 13 blocks, one can apply the same principle: the starting values are 1, 2, 4 and 10 respectively. Of course, the table has 13 rows in this case, and each test form has four blocks of items. For test forms containing five blocks of items, one needs 21 different test forms, but in practice this is seldom feasible.

will probably have to have recourse to IRT. Good use of IRT, however, requires substantial sample sizes. 200 responses<sup>32</sup> per item can be considered as a minimum in order to provide sufficiently stable estimates.

#### 7.2.4. Psychometric Aspects

It is important that the pretest gives sufficient data for approximate psychometric characteristics of the examination to be indicated. The first aspects concern characteristics at the item level such as the difficulty ( $p$ -value) and the discriminatory power of the items. If one sticks to indices from Classical Test Theory, one should realise that these indices are population dependent, and that their values are only indicative of the values in the target population if the pretest sample is representative of this target population. Relying exclusively on a number of schools for convenience (e.g. schools of teachers who are members of the construction team of the test) may lead to serious biases.

Secondly, the reliability of the examination is important to a good CEFR linking project, as it has an impact on the accuracy and consistency of the classification to the levels of the CEFR, as will be demonstrated below. In estimating the reliability two aspects are to be kept in mind:

- It often happens that the KR20 (or Cronbach's alpha) are reported as reliability coefficients, but these indices are not the reliability, they are lower bounds to the reliability and with heterogeneous tests they may substantially underestimate the reliability. A far better index is the greatest lower bound (GLB) to the reliability as explained in Section C of the Reference Supplement.
- If an incomplete block design has been used for the pretest, the GLB is only available per test form. To have a reasonable estimate of the reliability of the examination as a whole, this index can be computed per test form only for the items that will be selected for the exam. On each such estimate, the Spearman-Brown formula can be applied to estimate the reliability of the full-length examination. Taking the average of all these estimates will give a reasonable approximation to the examination's reliability if care has been taken to make the item blocks heterogeneous and as representative for the final product as possible.

#### 7.2.5. The Timing of the Standard Setting

If linking to the CEFR is high stakes, there will usually not be enough time between collecting the data from the examination administration and the release of the results to organise a complete standard setting procedure and to assess its validity.

As it is advisable to use real data from students even in test centred methods of standard setting (impact information, reality feedback; see Chapter 6), the time between pretesting and final administration of the examination will probably be the time best suited for organising the standard setting. Using the two-year planning period, as described above, even offers the possibility to cross-validate two standard settings. This is explained in more detail in Section 7.4.

In the present section, the discussion will be restricted to the consequences of what is sometimes called the *pretest effect*. This term refers to all systematic differences between pretesting and real examination, which may influence performances. The main influence probably comes from a difference in motivation and all factors directly linked to motivation like seriousness of preparation and test anxiety. If the examination is high stakes and the pretest low stakes, all these factors may work in the same direction, that of lowering the performance in the pretest as compared to the examination. If this is the case, the impact data presented to the panel during standard setting will be biased and may have a systematic effect on the proposed standards: if panel members consider themselves as being too strict as a consequence of this biased information, this may lead to lower standards.

Here are some suggestions about what one possibly could do to avoid or control the pretest effect:

---

<sup>32</sup> It is highly advisable not to take this number as an established rule: it just gives an indication of the order of magnitude of the sample size. In high stakes applications one needs the professional advice of a trained psychometrician who can judge, probably with the aid of computer simulations, the appropriateness of the sample size.

- Try to organise the pretest under conditions as similar as possible to the real examination. Presenting a pretest as a kind of general rehearsal for the real examination, close in time and with as high stakes as possible might tend to make motivation and preparedness more similar for the two sessions.
- Adding a short questionnaire after the pretest may be helpful. For example students showing low interest in the pretest or asserting having had “no time or opportunity” for a serious preparation might be excluded from the analysis.
- If one succeeds in doing pretesting in the way described for several years, pretest data and real examination data may be compared to make an estimate of the pretest effect. If one obtains a fairly stable estimate over time, the pretest effect could be explained to the panel members, and a kind of “corrected” impact data could be presented during the discussion sessions. For example, if the pretest effect is estimated at two score points (the average being two points higher in a real examination than at pretest time), one could add this effect to each score obtained in the pretest to compute the proportion of students in each level using the provisional standards. Of course, one has to tell the panel members about this correction (and its justification); nobody stands to win by withholding information and lying can have serious consequences.

### 7.3. Procedural Validity of the Standardisation Training and Standard Setting

In the preceding chapters, a number of procedures have been described to familiarise panel members with the CEFR, to understand the specification of the examination, to determine useful benchmarks and to set the standards. The standard setting sessions themselves then need to start with explanations and instructions so that panel members feel confident in completing their tasks. All these procedures can be considered as steps following good practice; ignoring them puts the outcomes at risk. Following such procedures can be considered as a *necessary* condition for a good result, or to put it in a more direct way: they exemplify the saying “garbage in, garbage out”.

The validity problem is concerned with the *sufficiency* of the procedures. Taking the examples of Familiarisation (Chapter 3) and Standardisation training (Chapter 5): if there is no training at all in the understanding of the CEFR, one cannot count on achieving a valid result. If, on the other hand, the suggested training procedure has been followed, there is no guarantee that the result will be successful; training is necessary, but was it sufficient? Validation of this aspect involves showing that the training has been effective: if one trains people to understand something, one has to show that they really do understand it after the training.

A number of aspects for demonstrating such procedural validity will be discussed in turn. They are explicitness, practicability, implementation, feedback and documentation.

**Explicitness:** This term refers to the degree to which the standard setting purposes and processes were *clearly* and *explicitly* articulated *a priori*. It means that the whole process is defined before it starts, that the steps are clearly described, and that the conditions and expected outputs for every step are described as a fixed scenario.

A good criterion to judge on the explicitness is to check whether the planning is such that it could serve as a guide for a genuine replication of the whole procedure. One way of checking whether the explicitness criterion is fulfilled is to ask the participants if they got a clear understanding of the purpose of the standard setting meeting and how clearly the standard setting task was explained.

**Practicability:** Although some procedures are quite complicated, the preparation must be practical (see Berk 1986), such that:

- The standard setting method can be implemented without great difficulty.
- Data analysis can be addressed without laborious computations. This does not mean that the computations cannot be complicated; but the preparatory work – like producing Excel spreadsheets with the appropriate formulae – must be completed well before the sessions.
- The procedures are seen as credible and interpretable by non-technicians.

One way of checking whether the practicality criterion is fulfilled is to ask the participants if the training was really helpful for them to understand how to perform the task.

**Implementation:** This criterion refers to how systematically and rigorously the panel was selected and trained, how well the CEFR levels were internalised and how effectively the judgment data were dealt with. Information on these points should be provided.

**Feedback:** This criterion has to do with how confident the panel feels in the standard setting process and in the resulting cut scores. Are the panellists happy that they achieved the right result? Information needs to be collected and reported.

**Documentation:** This has to do with how well the standard setting procedure is documented for evaluation and communication purposes.

#### 7.4. Internal Validity of the Standard Setting

Questions of internal validity try to answer questions about the *accuracy* and the *consistency* of the standard setting results. Lack of consistency may be due to a general weakness of the method applied or it may be localised within one or two judges or a few items. If the weakness is a local one, one might consider removing certain panel members from the whole process (or the analysis following it) or basing the linking process on a subset of the items and tasks in the test, excluding those that have caused problems.

- In removing judges, one should be careful not to influence the outcome of the standard setting in a direction desired by the organiser. If evidence can be found that a panel member did not understand the instructions or intentionally ignored them, this may be a valid reason to remove this panellist's data from the analysis. Post-session interviews and a well-conceived questionnaire may provide such evidence. Such a removal should be well documented and in the final report it should be mentioned how many panel members are removed and why.
- Removing items or tasks is an even more delicate problem. If linking to the CEFR is the main purpose of the examination, e.g., by applying the rule that a fail in the examination is the same as not having reached the standard B1/B2, then removing certain items could seriously bias the content validity of the test. This in turn could create ethical problems by having students do preparatory work for an examination, which turns out to be partially a wasted effort. If, on the other hand, linking to the CEFR is considered a side product of the examination, removing items from the linking study, whilst keeping them in the analysis for reporting candidates' results, can be justifiable.

The rest of this section discusses a number of topics that are all related to consistency and accuracy:

- the intra-judge consistency: where indications are sought to show that a single judge has been consistent in his/her judgments with other sources of information one has about the test;
- the inter-judge consistency: where one investigates to what extent panel members agree with each other in their judgments;
- the stability of the results, expressed as the standard error of the cut-offs;
- the accuracy and consistency of the classification based on the standard setting.

Not all methods proposed to check consistency are applicable to all standard setting methods discussed in the preceding chapter. Therefore we will use the modified Tucker-Angoff method as a working example, and add comments for other methods when appropriate.

##### 7.4.1. Intra-judge Consistency

In this context, one can ask two sensible questions: is the judge (panel member) consistent with himself/herself and second, are his/her answers consistent with other information one has about the test?

To answer the first question, it is necessary that the panel member gives an answer twice to the same question (or to two very similar questions). This could be accomplished during standard setting in a special, repeated measurement set up in which the final round is a partial repetition of items from earlier rounds. When working with multiple standards (for several levels) and the Tucker-Angoff method, one could ask each judge to give their probability estimates a second time for one of the standards. Since the probability estimates are fractional numbers, a scatter diagram and a correlation coefficient can give insight into the internal consistency of the judgments. The correlation can directly be interpreted as the *reliability* of the judgments. Comparisons of these reliabilities across judges may give useful information about outlying panel members, and this may be used to possibly exclude one or two panel members' data from further analysis.

In giving probability statements for a borderline person, panel members give implicitly an indication of the difficulty of the items. Judging that the borderline person has a probability of 0.6 of giving a correct answer to item  $i$  and of 0.4 for item  $j$  means that the panel member judges that item  $i$  is easier (higher values indicating easier items). These estimated probabilities may also be correlated with empirical indices of difficulty, such as  $p$ -values (where one expects positive correlations) or estimated difficulty parameters in an IRT application (where the correlation is expected to be negative). This kind of indices can be considered as a *validity* coefficient as they express the relation between the judgment on a set of items with an external criterion, the empirically determined difficulties from students' responses.

Setting rules of thumb for adequacy of correlation is a difficult problem and one should be careful with such rules. The value of the correlation will depend to a high degree on the standard deviation of the item difficulties, lower values of the SD leading to lower values of the correlations (restriction of range effect). But as with the reliabilities, comparing the correlations across panel members may give valuable information with respect to outliers.

Computing these indices after the whole procedure is finished is certainly worthwhile for reporting and publishing purposes, but they can be very useful during the sessions as well. After each judgmental round these correlations and associated scatter diagrams can easily be produced and be used in the discussions to point to misunderstandings or disagreements that one might wish to solve.

Similar techniques can be used with other standard setting methods as well. We discuss two cases: the Body of Work method and the Basket method.

- In the Body of Work method students are allocated to a CEFR level on the basis of a holistic judgment of a dossier of their work. One can consider these CEFR levels as ordinal variables, A2 ranking higher than A1, B1 ranking higher than A2, etc. For all of the students under consideration, the test score is known and the correlation between test score and allocated level can be computed; the bivariate data (scores and allocated level) can also be graphically displayed in a scatter diagram. For the computation of the correlation, it is advisable to use a rank correlation coefficient, Kendall's tau-b<sup>33</sup>, which allows for a correction of ties.
- In the Basket method, the same rationale can be used to relate the allocated levels for the items and their empirical difficulty.

We end this discussion with two warnings:

- In the Body of Work method as described in the preceding chapter, the dossiers of the students are presented in the order of increasing scores. Either this rank ordered information is conveyed to the panel members, or not, in which case they soon will find out that there is such an ordering. By presenting dossiers in rank order in this way, internal consistency is induced to some extent by the method itself: a panel member will realise very soon that the higher the rank order of a dossier in the files of dossiers he has to judge, the higher the level that should be allocated. This may induce a kind of social desirability (the panel member not "daring" to give a high level to a work early in the row or a low level to a work later in the sequence). This tendency may obscure to some extent what the panel member is really

---

<sup>33</sup> See, for example, Siegel & Castellan (1988).

thinking (and this may have odd consequences for the outcomes of the procedure), while it will at the same time lead to an increase of the correlations as discussed above.

- Some methods give so much information to the panel members that it is virtually impossible to exhibit inconsistent behaviour. Typical examples are the Bookmark method and its Cito variation, where for each standard a single holistic judgment has to be given. In the Bookmark method as discussed, it is even impossible by the way the procedure is defined, to generate a lower standard for A2/B1 than for A1/A2. This does not mean, however, that intra-judge consistency is not important in these procedures. In the Cito variation of the Bookmark method, for example, the operational task for the panel members is so simple (drawing a line or writing down a number; see Section 6.9.) that an arbitrary individual standard set by an uninterested panel member may go unnoticed. Therefore it is advisable to check the intra-judge consistency in this procedure by an extra task. This could go as follows. Once the individual standard has been set, intra judge consistency can be derived for each item if at the value of the standard “No mastery”, “Borderline mastery” or “Full mastery” is required. Referring to Figure 6.5, “No mastery” of an item means that the vertical line, representing the individual standard, passes to the left of the horizontal line representing the item itself; “Borderline mastery” means that the vertical line crosses the item line, and “Full mastery” means that the vertical line passes to the right of the item line. So with respect to the individual standard, all items can be classified as belonging to one of these three categories. In an independent task, the panel members could be asked to classify all items into one of these three categories without the psychometric information (See Figure 6.5) being available. These two classifications, one derived from the provisional standard and one collected by the blind allocation, can be displayed (per panel member) in a 3 x 3 frequency table, and indices of agreement can be computed for them.

#### 7.4.2. Inter-judge Consistency

In judging inter-rater consistency, one tries to determine the extent to which panel members agree with each other or – in a weaker sense – give similar judgments. The latter is usually called consistency. It is important to make a clear distinction between these two concepts. We consider a small example to explain the difference.

##### 7.4.2.1. Agreement and Consistency

Suppose that 30 items are to be allocated to one of five CEFR levels, as in the Basket method, and the judgments of two judges are summarised in a two dimensional frequency table (see Table 7.3). One can see that Panel Member 1 allocated seven items to Level A1, and that Panel Member 2 allocated the *same* seven items to Level A2. So for these seven items, the two judges disagree completely on the level these items are to be placed at. But the same holds for the other items as well, as we can easily see in the table, because all the frequencies on the main diagonal (the underlined numbers) are zero. But in spite of this total disagreement, one cannot say that there are no systematic similarities between the decisions of the two judges: Panel Member 2 places all items one level higher than Panel Member 1, meaning that Panel Member 2 is more lenient in his judgments than Panel Member 1.

**Table 7.3: Example of High Consistency and Total Disagreement**

		Judge 2				Total
		A1	A2	B1	B2	
Judge 1	A1	<u>0</u>	7	0	0	7
	A2	0	<u>0</u>	11	0	11
	B1	0	0	<u>0</u>	12	12
	B2	0	0	0	<u>0</u>	0
	Total	0	7	11	12	30

Since the four CEFR levels are clearly ordered, one can compute a rank correlation coefficient between the judgments of both judges. Kendall’s tau-b in this case equals 1, expressing complete *consistency* in the

judgments of these two panel members. In general, then, we can say that consistency measures, generally expressed by a correlation coefficient, are not sensitive to systematic shifts in the judgments which can be associated with relative harshness or leniency in the judgments. Therefore, it is useful to pay attention both to agreement and to consistency when judging the work of the panel members<sup>34</sup>.

#### 7.4.2.2. Three Measures of Agreement

To illustrate these measures, we use a more realistic outcome than the highly artificial data in Table 7.3. Suppose 50 items are to be allocated to four levels and for two judges one finds the bivariate frequencies displayed in Table 7.4.

**Table 7.4: Bivariate Frequency Table using Four Levels**

		Judge 2				Total
		A1	A2	B1	B2	
Judge 1	A1	7	2	1	1	11
	A2	1	10	2	1	14
	B1	1	2	12	2	17
	B2	0	1	0	7	8
	Total	9	15	15	11	50

The index of *exact agreement* is the proportion of cases (items) where the two judges come to exactly the same judgment. The frequencies of exact agreement are given by the cells on the main diagonal (dark grey) of the table. So, in this example

$$p_{exact} = \frac{7+10+12+7}{50} = \frac{36}{50} = 0.72,$$

which is not impressively high in this type of context. Of course for items where the two judges disagree, the disagreement may vary in degree: an outcome where an item is placed three levels apart is more worrisome than a case in which the allocated levels are adjacent. These latter cases are displayed in the light grey cells of Table 7.4 along the main diagonal. In total there are 2+2+2+1+2 = 9 such items. The *index of adjacent agreement* is the proportion of items leading to exact agreement or to a difference of one level. In the example we find that

$$p_{adj} = \frac{36+9}{50} = \frac{45}{50} = 0.90.$$

Even if the two judges give their judgments at random, the indices of agreement will not equal zero, but will take a positive value whose magnitude will depend on the marginal frequencies (the bottom row and the rightmost column in Table 7.4). The *expected* number of cases in each cell – under the hypothesis of random responses but with fixed margins – is given as the product of the row total times the column total divided by the grand total. For the cell (A1, A1) in Table 7.4 this gives 11 x 9 / 50 = 1.98. For the other three cells on the main diagonal the expected frequencies are 4.20, 5.10 and 1.76, and the sum of the expected frequencies for all cells on the main diagonal is 13.04. Therefore, if the judges answer at random, one expects an index of (exact) agreement equal to

$$E(p_{exact}) = \frac{13.04}{50} = 0.26.$$

A widely used index of agreement, Cohen's kappa, takes this agreement by chance into account. It is defined (for the exact agreement) as

$$\kappa = \frac{p_{exact} - E(p_{exact})}{1 - E(p_{exact})}.$$

<sup>34</sup> A multifaceted IRT analysis of the judgment data using the program FACETS is one way of doing this.

In the numerator of this formula, the empirical proportion of agreement found is compared with what could be expected under random responding. The function of the denominator is to keep the maximum value of kappa equal to 1. Notice that kappa can be negative: if the agreement found is even lower than one could expect under random responding.

### 7.4.2.3. Evaluating Indices of Agreement

As is the case with many psychometric indices, it is hard to evaluate the results found in a study in an absolute manner. Formulating absolute benchmarks is barely feasible and can be risky.

- Take the index of absolute agreement as an example. If the items to be judged form a fairly homogeneous set, for example being constructed for the Levels A2+ and B1, an average index of agreement of 0.8 may be exceptionally high. On the other hand in a case of a very heterogeneous collection of items across a wide range of levels, the same value may be unsatisfactory, even pointing to a non-serious attitude of one or more panel members.
- It is useful to consider very carefully the set-up of a standard setting study, and to keep in mind that the method used may induce high or low values for the agreement between panel members. The Body of Work method offers a nice example. In this method, students are allocated to a level, but the material selected for the range finding round has to be very heterogeneous, covering the whole score range, and this heterogeneity will facilitate high agreement. If one works with an absolute criterion (of say 0.8) for the average index of agreement, reaching this value may create a feeling of satisfaction. However, at the same time it may happen that this apparently high index actually obscures the fact that one or two panel members did not understand the instructions, and have substantially influenced the final standards in an undesirable way.

A more fruitful approach is offered by taking a relative viewpoint. The indices discussed above are defined for pairs of judges. With 12 panel members, this means that there are  $12 \times 11/2 = 66$  pairs and for each pair one or more indices can be computed. Of course, these indices will show variability among them, and the important question is whether one can study this variability to improve the results (in a subsequent round with discussion focused on the problematic areas) or to identify and remove some badly performing judges or items in order to improve the overall quality of the standard setting.

Although there are methods to generalise indices like Cohen's kappa to more than two judges, such summary indices may tend to obscure outlying patterns and are seldom useful in pinpointing the weak points in a multi-rater study. Here we will sketch an easy way to evaluate the strengths and weaknesses of the inter-judge agreement. As an example we will use Cohen's kappa index, but the same procedure can be applied with the index of exact or adjacent agreement.

- Arrange the indices in a square table. The entry in cell  $(i,j)$  is the kappa coefficient computed for panel members  $i$  and  $j$ . The table is symmetric, and the entries on the main diagonal are left undefined. They do not enter any of the calculations to follow.
- Useful information can now be extracted by computing two indices for each column of the table:
  - Compute the average for each column, yielding an index for each judge which expresses the general level of agreement with all other judges. A graphical display of these column averages will immediately point to the panel members disagreeing most with the others, as they will get the lowest values.
  - Compute the standard deviation in each column. The joint evaluation of the average and the standard deviation may give additional information. If the average is low and the standard deviation is small, this means that the panel member disagrees with the others and does so in a systematic way. This may occur in a situation where the panel member has systematic deviating ideas on the meaning of the CEFR or the meaning of the items. If, on the other hand,

the SD is high, this may point to erratic behaviour. A scatter plot of averages and standard deviations may be helpful in diagnosing problems with one or more panel members.

The technique explained in the preceding paragraphs is useful in cases where only a few panel members show behaviour which deviates from that of the majority of the other panel members. In cases where for example the group of panel members falls apart in two subgroups, who agree to a high level with members of their own subgroup, but disagree substantially with the members of the other subgroup, this technique may fail. In such a case, it is advisable to use techniques which can reveal a complicated structure in the matrix of agreements. Cluster analysis and multidimensional scaling may be appropriate here.

#### 7.4.2.4. Finding Problematic Items

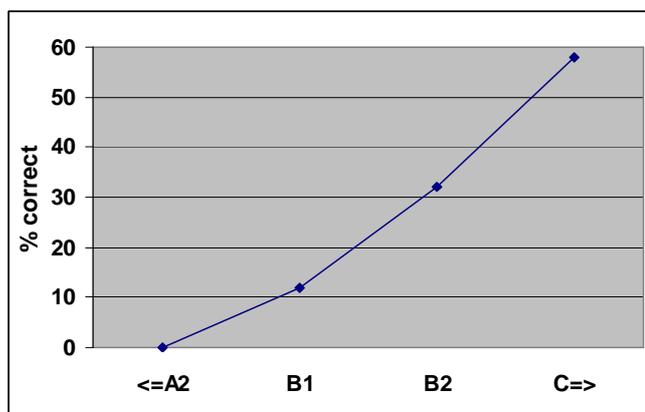
In standard setting procedures where panel members allocate items or tasks to a level (like the Basket method or the Item-descriptor Matching method), there are two easy ways to find out whether the relative lack of agreement can be attributed to a few items or not.

The first one is to construct a table or a graphical display per item (e.g. a column diagram) giving the frequencies (absolute or relative) of allocation to each level. An example of a problematic item is given in Table 7.5<sup>35</sup>. In Figure 7.1 the empirical item characteristic curve for this item is displayed. Students have been grouped in levels (as indicated along the horizontal axes) using the standards as set by the panel, and for each group the percentage of correct responses on this item is displayed.

**Table 7.5: Frequencies of Allocation of a Single Item to Different CEFR Levels**

Level	A1	A2	B1	B2	C1	C2
Frequency	0	17	11	5	0	1

From the figure, two important characteristics of this item can be derived: (a) it is a fairly difficult item, which cannot be solved by students at the A-level, and (b) it has a proportion correct of less than 0.6 for students at the C-level. Furthermore, the curve is increasing rather steeply, indicating a good discriminating power for the item. Combining these empirical facts with the judgments of the panel leads to the question: How can one explain the fact that the majority of the judges allocated this item to Level A2? Furthermore, one sees that only one panel member locates the item at a C-level, while a simple analysis of Figure 7.1 seems to show that he/she is actually right! This again teaches us that applying a simple majority rule and suppressing disagreement with a consensus is not always a good decision. It is clear that the presentation of Table 7.5 and Figure 7.1 would be valuable input for a discussion round.



**Figure 7.1: Empirical Item Characteristic Curve for a Problematic Item**

A second method to generate an overview of problematic items is to use the information from the bivariate frequency tables as shown in Table 7.4. In that table, there are five items which are allocated two or more

<sup>35</sup> This is a real example from a recent standard setting seminar.

levels apart from each other by two panel members. If one identifies these items, and does so for each pair of panel members, one can construct a frequency table as exemplified in Table 7.6.

The rows of the table correspond to items, and the cell entries are the number of times the item has been allocated to different levels. The entry ‘3’ in the row for the first item and in the column labelled ‘two’ says that there were three pairs of panel members that have placed this item in places exactly two levels apart. Items with the highest frequencies in the rightmost column are probably the most problematic ones and deserve the most attention in discussion rounds. From the table, it is clear that item 3 deserves most attention.

**Table 7.6: Summary of Disagreement per Item**

Item ID	Two levels apart	Three levels apart
1	3	1
2	2	0
3	3	7
4	0	0
5	2	0
⋮	⋮	⋮

#### 7.4.2.5. Indices of Consistency

Three different methods to judge the consistency or lack of consistency in the rater judgments will be discussed: the intra-class correlation, a method which is a direct application of Classical Test Theory, and, very briefly, a measure of consistency appropriate for judgments on an ordinal scale.

**The Intra-class Correlation:** Consider the modified Tucker-Angoff method. The basic outcomes of the procedure are so-called Angoff ratings, i.e. statements about the probability of a correct response by a borderline person. These data can be arranged in a rectangular array, the rows indicating the items and the columns the panel members or judges.

In the ideal case where all judges would agree completely with each other, all columns of this table would be identical. This means that all variability between the numbers in the table can be attributed to the items. If there is variability due to the raters as well, there is departure from the ideal situation, which is precisely what is called inconsistency. A way to express the lack of consistency is to express the proportion of variance due to between item variance. This proportion is called the intra-class correlation, and is a number between zero and one, the latter being the ideal situation. Here is how to compute this intra-class correlation:

- Compute the variance on all the numbers in the table. This variance is called the *total variance*.
- Compute for each row in the table the average value. Compute the variance of these averages. This variance is the variance due to the rows (items).
- The ratio of the two is the intra-class correlation, indicated here symbolically as  $\rho_{ic}$

The difference  $1 - \rho_{ic}$  is the proportion of variance *not due* to differences between the items. This variance may be due to *systematic differences* between judges or to interaction between items and judges and random noise. To distinguish between the latter two we may compute easily the variance between the columns (judges), by computing the average of each column, and then computing the variance of these column averages.

**Table 7.7: Outcome of a Tucker-Angoff Procedure**

Items/judges	1	2	3	Average
1	38	32	24	31.3
2	27	31	38	32.0
3	42	33	50	41.7
4	51	49	47	49.0
5	52	60	62	58.0
6	63	58	71	64.0
7	71	68	75	71.3
8	82	77	92	83.7

<b>Average</b>	53.3	51.0	57.4
----------------	------	------	------

In Table 7.7 an artificial example is given with eight items and three judges. The numbers in the table represent the number of borderline persons out of 100 who would, according to the judges, give a correct response to each of the items. The rightmost column contains the row averages and the bottom row the column averages.

In Table 7.8 the decomposition of the total variance into three components is shown. The residual variance (interaction and error) is obtained by subtracting the item component and the judges' component from the total variance.

**Table 7.8: Variance Decomposition**

<b>Source</b>	
<b>Items</b>	308.91
<b>Judges</b>	6.97
<b>Residual</b>	17.89
<b>Total</b>	333.78

From this table, we learn that:

- The intra-class correlation is  $308.91/333.78 = 0.926$ , meaning that only about 7.5% of the total variance is due to the different way in which judges treat the items.
- The variance due to systematic differences between judges is 6.97, which compared to the total variance is about 2.1%.
- The remaining proportion (amounting to 5.4%) is what one really could call inconsistency.
- In this (artificial) example the intra-class correlation is very high, but this is not necessarily to be attributed to the quality of the judges or the standard setting process in some absolute sense. The items (row averages in Table 7.7) show a rather high variability, and what the results in Table 7.8 really tell is that the inconsistency of the raters is *relatively* small compared to the variability between the items.

This splitting up of the total variance can easily be accomplished (e.g. in an Excel spreadsheet) and is useful in guiding the subsequent discussion sessions as well as in reporting on the internal validity of the standard setting.

**Using Classical Test Theory:** Classical Test Theory offers a nice index of consistency in Cronbach's alpha. To apply this procedure here, we use the Angoff ratings as given in Table 7.7 as test data, where the items (the rows) of the table take the role of students and the judges take the role of items. So for Table 7.7 this would mean that it contains the scores of eight students to three items. The value of alpha in this example equals 0.97.

Note that the value of alpha does not change if the unit of measurement is changed. Concretely, the result will be the same if the data in Table 7.7 are expressed as percentages or as proportions<sup>36</sup>. More details on Cronbach's alpha are given in Section C of the Reference Supplement.

Using Classical Test Theory also offers another advantage. Using the item-total correlations in the present context gives an indication on how well each judge (having taken the role of item) agrees with the average judge, thus giving a nice way to detect outlying panel members. In the example in Table 7.7 all three correlations equal 0.98.

**Ordinal Measures:** The methods discussed in the preceding sections are applicable whenever one has observations that can be arranged in a two way table, mostly items by judges in test-centred methods of standard setting or students by judges in examinee centred methods like the Body of Work method. A problem can arise, however, when one has to decide what exactly to put in the two-way table and how to interpret the values in the table.

<sup>36</sup> On the condition, of course, that one is consistent throughout the table: using percentages for half of the columns and proportions for the other half will lead to strange and completely useless results.

We take the Item-descriptor Matching method as an example. The basic judgments given by the panel members are CEFR levels, ranging say, from A1 to C2. One could fill in these levels in the table, as labels, but then one cannot apply the above methods since these require a table with numbers. What one can do in such a case is replace the labels A1 through C2 by the numbers 1 to 6, and then proceed as above. In the literature, different positions are taken towards such a procedure, some authors arguing that it is not permissible since the numbers used to fill the table (1 to 6) are not measures on an interval scale. This is a correct argument, but it does not, however, make the use of techniques of variance decomposition or the use of techniques from Classical Test Theory useless. Applying them may give useful information, even if the interpretation is not ultra-orthodox. On top of that, one can also have recourse to indices of consistency that completely rely on the ordinal character of the data. A good candidate is Kendall's coefficient of concordance  $W^{37, 38}$ .

### 7.4.3. Accuracy and Consistency of the Standard Setting Method

Whatever one does do in the training sessions and in the discussion rounds, if one insists that panel members can give their judgments freely, independently and without fear of sanctions, it is unavoidable that there will remain variability in the judgments. This is not necessarily a bad thing, because panel members are not invited in their individual capacity, but to come to a reasonable and well considered group decision. Moreover, if the selection process of the panel members has been carried out with great care, such that the actual panel members are representative for their peers, this means that with another sample of the same size, one would observe results quite similar to the ones one has actually observed.

#### 7.4.3.1. Standard Error of the Cut Score

An appropriate question to ask then is what the cut scores would be if one could involve the whole population of judges who are considered as expert in the subject matter, i.e. the population mean. If we take the mean judgment (cut score) from the panel members in the sample, their averaged cut score is an estimate of this population mean, and the standard error ( $SE_S$ ) of this estimate is given by the standard deviation ( $SD_S$ ) of the individual cut scores in the sample divided by the square root of the number of panel members  $n$ :

$$SE_S = \frac{SD_S}{\sqrt{n}}$$

In the literature this standard error is usually compared to the standard error of measurement of the test, and it is generally agreed that the standard error of the standards must not exceed the standard error of measurement. But some authors are stricter. Cohen et al (1999) require that the standard error of the standards should be at most half of the standard error of measurement, while Jaeger (1991) requires it to be at most one quarter. Norcini et al (1987) advise that the standard error of the standards should not be more than two items out of a hundred. This means that for a test of 50 items, the standard error of the cut score should be at most one.

Standard 2.14 of the AERA/APA/NCME (1999) states:

“Where cut scores are specified for selection or classification, the standard errors of measurement should be reported in the vicinity of each cut score.”

Simple applications of Classical Test Theory usually report a single value for the standard error of measurement, implying that scores (as indicators of the true score) are equally precise independently of the value of the (true) score. In applying IRT, however, it is well known that the standard error of an ability estimate depends on the value of the variable itself (see the concept of test information in Appendix G of the Reference Supplement).

<sup>37</sup> For a good introduction, see Siegel and Castellan (1988).

<sup>38</sup> There also exist valuable techniques to do quantitative analyses on tables that contain data at the nominal level, i.e. where A1, ..., C2 are just considered as labels. These techniques are known under several names, like homogeneity analysis or multiple correspondence analysis. A practical reference is OECD (2005), Chapter 10.

Within the framework of Classical Test Theory, there have been attempts to arrive at different values for the standard error of measurement which depend on the score level (Feldt et al 1985). A suitable formula for expressing the standard error at different score levels for tests consisting of binary items is from Keats (1957):

$$SEM(X) = \sqrt{\frac{X(k-X)}{k-1} \times \frac{1-\rho_{xx'}}{1-KR_{21}}}$$

In this formula:

- $X$  represents the score;
- $k$  is the number of items;
- $\rho_{xx'}$  is the reliability of the test;
- $KR_{21}$  is one of the Kuder-Richardson formulae, which expresses the reliability of a homogeneous test for items of (about) equal difficulty. The formula for the  $KR_{21}$  is given by

$$KR_{21} = \frac{k}{k-1} \left[ 1 - \frac{k\bar{p}\bar{q}}{SD_x^2} \right]$$

where  $\bar{p}$  is the average  $p$ -value and  $\bar{q} = 1 - \bar{p}$ .

Notice that  $SEM(X)$  gives a different outcome, depending on or conditional on the score  $X$ . Therefore it is usually called the conditional standard error of measurement. Its values are largest for scores near the middle of the score range, and decrease as the score either becomes lower or higher. This means that if one chooses a criterion to judge the standard error of the cut score, such as the requirement that it is not higher than half the standard error of measurement, this will lead to a requirement of a smaller standard error the further the cut score is away from the middle of the score range.

#### 7.4.3.2. A Paradoxical Situation

It is well known that in applications of IRT, one obtains the most accurate estimates of the latent ability from students having about half of the items correct, meaning a score about halfway between the lowest and highest possible scores, while the results presented on the conditional standard error of measurement indicate just the opposite. To understand this seemingly contradictory result, one has to realise that the score range on a test is bounded from below, the minimal score being usually zero, and from above. With 50 items, each one worth one point, the maximum score is 50. In IRT, to the contrary, the basic concept is not the test score but an abstract latent variable that is conceived to be unbounded, i.e., it can accommodate all values from minus infinity to plus infinity.

A suitable way to express the relation between the latent variable and the score is the test characteristic function<sup>39</sup>. In Figure 7.2, a test characteristic curve for a test of 50 items is displayed. Although the curve has a general S-shape, it is not very regular; irregularities are caused by particular combinations of discrimination and difficulty parameters of the items<sup>40</sup>.

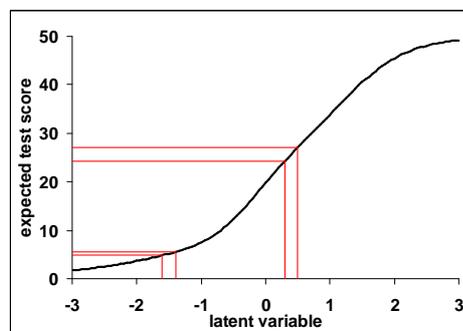


Figure 7.2: A Test Characteristic Curve

<sup>39</sup> More details on this function are given in Section 6.8.3.

<sup>40</sup> It is good practice, when using IRT, to construct the test characteristic curve: it makes the relation between an abstract concept (the latent variable) and observable facts (test scores) explicit. The parameters for the curve in Figure 7.2 were chosen to stress this irregularity.

On the horizontal axis two intervals are displayed, each one having a width of 0.2. The left one runs from -1.6 to -1.4, and the corresponding expected test scores for these two values are 4.82 and 5.54 respectively, i.e. a width of 0.72 score points. The second interval, having the same width on the horizontal axis (0.3 to 0.5) leads to an interval in the expected scores running from 24.26 to 27.00 score points, i.e., a width of 2.74 score points, about four times the width of the first interval.

If a standard setting method has been used where the cut-off point is determined on the latent scale, like in the Bookmark method or its Cito variation, the standard error is expressed on this scale. But for most users a cut-off point for the scores will be needed, and therefore an estimate of the standard error in the score domain has to be given as well. Use of the test characteristic curve may be helpful here<sup>41</sup>.

### 7.4.3.3. Accuracy and Consistency of Decisions

Setting standards, i.e., determining cut scores, implies making a decision on individual performances. If the cut score A2/B1 is set as 23/24, this implies the decision that any student obtaining a score lower than 24 on the examination will not be granted Level B1. By doing so, one intends to grant a certain level to a student if he/she really deserves it. But some decisions may be in error, and it is useful to distinguish several sources of error. A concrete example may help:

Suppose that student John obtained a score of 22 on the test.

- With a cut off of 23/24, John will not be granted B1. But if we replicated the standard setting procedure with a different sample of panel members, maybe we would arrive at a slightly different cut score for A2/B1 such that John would be categorised as B1 with a score of 22. So we are uncertain about our decisions because of the variability of average cut scores across replications of the standard setting procedure. This uncertainty is quantified in the standard error of the cut scores as discussed above.
- But even if we take the cut scores as determined in a single standard setting procedure, we might take the wrong decision on John, as it might have happened that John had a bad day when he took the test (resulting in a negative measurement error), while “on the average” he would pass the A2/B1 cut off. Variation between observed scores and true scores is well expressed by the reliability of the test (or by the related concept of the standard error of measurement). In the validation of a standard setting procedure, therefore, it is indispensable to relate characteristics of the standard setting and of the test itself, to have an accurate idea of the sources of error or inconsistency.
- The third kind of errors that can be made in standard setting consist of systematic errors. If panel members are too lenient as a group, this may lead to excessively low cut-off scores, categorising systematically students as B1 who do not really deserve it. Systematic errors directly influence the external validity of the procedure, and will be discussed in greater detail in the next section.

In the present section we will concentrate mainly on the second source of variability: the variation in decisions due to imperfect reliability of the test. We could have a good idea of the effects of lack of variability by making a sample of students to take the same test twice and by then constructing a bivariate frequency table to see how many students are categorised twice in the same category. Indices of agreement (absolute or Cohen’s kappa) would then give an indication of the consistency of the decisions.

Unfortunately, administrating the same test twice to the same students is seldom feasible in an examination context, and therefore one has recourse to psychometric models to derive measures of consistency from a single administration of the test. A fruitful approach is offered by the work of Livingston and Lewis (1995) that we discuss here briefly. Based on the work of Lord (1965), they assume a distribution of true scores

---

<sup>41</sup> But note (see Section 6.8.3.) that conversion of test scores to latent values via the test characteristic curve implies the use of the maximum likelihood estimate which can be severely biased when the cut-off scores are extreme.

which can be estimated from the distribution of observed scores in a representative sample of test takers, either using two parameters or four parameters<sup>42</sup>.

If the distribution is known (or estimated accurately), and if the cut-off scores are given, then:

- it can be determined what proportion of the population will be allocated to each category in a context of multiple standards;
- it can also be determined from the assumptions of the model and from the reliability of the test which proportion of the population will be categorised in each category (level) on the basis of the test score.

In the left hand panel of Table 7.9 an example of such a table is given for three categories (levels). The rows indicate the true category (level). From the column “marg” (meaning “marginal”) one can see that 16.04% of the population belongs to A2, 27.34% to B1 and 56.62% to B2. The reliability of the test has been estimated at 0.9. If a test with this reliability (not necessarily the test under study, but one with the same psychometric characteristics) is administered to the population, it is to be expected that 21.17% of the students will be categorised as A2 on the basis of their test score (to be read from the bottom row), and that 14.95% will be really in category A2 and also be categorised as such. From the diagonal of the table, we can determine the index of absolute agreement: it is  $0.1495 + 0.2002 + 0.4426 = 0.7922$ .

**Table 7.9: Decision Accuracy**

	A Test				The Test Under Study			
	A2	B1	B2	Marg	A2	B1	B2	Marg
T(A2)	0.1495	0.0109	0.0000	<b>0.1604</b>	0.1511	0.0102	0.0000	<b>0.1614</b>
T(B1)	0.0617	0.2002	0.0115	<b>0.2734</b>	0.0624	0.1874	0.0119	<b>0.2618</b>
T(B2)	0.0005	0.1232	0.4426	<b>0.5662</b>	0.0005	0.1154	0.4611	<b>0.5769</b>
Marg	<b>0.2117</b>	<b>0.3343</b>	<b>0.4540</b>	<b>1</b>	<b>0.2140</b>	<b>0.3130</b>	<b>0.4730</b>	<b>1</b>

The table in the left-hand panel of Table 7.9 has been estimated based on the observed score distribution of 1,000 students, of which 214, 313 and 473 have been categorised as A2, B1 and B2 respectively. From the left-hand panel, however, we see that the expected frequency of A2 classifications is not 214 but 211.7 (= 1,000 x 0.2117). To adapt this table such that the proportions in each category correspond exactly to what has been observed, one does the following: multiply each proportion in the table (not the margins) by the observed proportion and divide by the expected proportion in the column. For example, for the first row and first column, we find  $0.1495 \times 0.2140 / 0.2117 = 0.1511$ . The values for all nine cells are displayed in the right hand panel of Table 7.9. The row marginals are just the sums of the values in each row. The index of absolute agreement for this adjusted table is 0.7996.

Apart from giving valuable information on the accuracy of the decisions by an index of agreement, both tables also indicate a quite marked difference in the rate of false positives and false negatives: the proportion of false positives (being classified higher than one deserves) is about 2% while the rate of false negatives is about 18%.

To evaluate the consistency of the decisions, i.e. the extent to which the same or different decisions would be taken if two independent test administrations were used, two tables similar to the ones in Table 7.9 can be constructed. These tables are displayed in Table 7.10. The only difference between this table and Table 7.9 lies in the meaning of the rows. Whereas in Table 7.9 the rows indicate the classification on the basis of true score, in Table 7.10, the rows indicate classification on the basis of an independent administration of the test. So the left-hand panel indicates the joint classification probabilities based on two independent administrations (a test and another test with the same reliability), while the right-hand panel gives the joint probabilities for this administration and another test with the same reliability.

Since in the latter case measurement errors occur in both administrations, the index of agreement will be lower than in the case of accuracy testing. For both cases in Table 7.10, the index of agreement is about 0.77.

<sup>42</sup> In the two parameter model it is assumed that the true relative score (the proportion of items correct) follows a beta distribution; in the four parameter case, it is also assumed that the minimum and maximum relative true score can be different from zero and one respectively, and that these are also to be estimated from the observed data. The technical details of the model are quite complicated.

**Table 7.10: Decision Consistency<sup>43</sup>**

	A test				This test			
	A2	B1	B2	Marg	A2	B1	B2	Marg
A2	0.1663	0.0448	0.0007	<b>0.2117</b>	0.1681	0.0419	0.0007	<b>0.2107</b>
B1	0.0448	0.2212	0.0683	<b>0.3343</b>	0.0453	0.2071	0.0712	<b>0.3236</b>
B2	0.0007	0.0683	0.3851	<b>0.4540</b>	0.0007	0.0640	0.4012	<b>0.4658</b>
Marg	<b>0.2117</b>	<b>0.3343</b>	<b>0.4540</b>	<b>1</b>	<b>0.2140</b>	<b>0.3130</b>	<b>0.4730</b>	<b>1</b>

The most noteworthy difference between Table 7.9 and 7.10, however, is that in the latter case both tables are essentially symmetric, the proportion in the cell (A2, B1) being the same (approximately) as the proportion in the symmetric cell (B1, A2). For the left-hand table this symmetry is complete, and this is necessarily the case, because it is the outcome of two totally independent administrations of two parallel tests. This means that the difference between false negatives and false positives has no meaning in this case; they can only be considered in a meaningful way from the accuracy tables.

To see the influence of variation in cut-off scores, the accuracy tables can be recomputed with different cut scores, and the results can be meaningfully compared, especially with respect to their rates of false positives and false negatives.

A less sophisticated method to compute decision consistency is from Subkoviak (1988). An extensive discussion, together with the tables needed to apply the method can be found in Chapter 16 of Cizek and Bunch (2007). The method of Livingston and Lewis, however, is more versatile because it is applicable both with multiple standards and in cases where binary and partial credit items are used, equally or unequally weighted.

## 7.5. External Validation

The main outcome of a standard setting procedure is a decision rule to allocate students to a small number of CEFR levels on the basis of their performance in the examination. Usually test performance has been summarised already by a single number, the test score.

In the material presented in this Manual it has been stressed that the procedures to arrive at such a decision rule are complex and time consuming, that there are many possible pitfalls, and that the result is never perfect, due to measurement error in the test and residual variance in the judgment of the panel members. If all procedures have been followed with great care, if the examination has an adequate content validity and a high reliability, and if the standard error of the cut scores is low, one might think that the job is finished and summarise the results by a table showing the decision accuracy, like the left-hand panel of Table 7.9, which shows the limits of one's possibilities given that one has to use a fallible test.

The weak point in such reasoning, however, is that such an outcome depends completely on procedures carried out by 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 may be judged as too small a basis to warrant the truth, i.e. validity, of a claim such as: "if a student obtains a score of 39 or more on my test, he can deservedly be considered to be at Level B2". In general, the weakness resides in the contrast between the particularity of the procedures and the generality of the claim.

External validation then aims at providing evidence from *independent sources* which corroborate the results and conclusions of one's own procedures. Not all evidence provided, however, is independent from the

<sup>43</sup> Tables 7.9 and 7.10 have been computed with the computer program BB-CLASS developed by R.L. Brennan, and made freely available by the Center of Advanced Studies in Measurement and Assessment (CASMA) of the University of Iowa. The program can be downloaded from [www.education.uiowa.edu/casma/](http://www.education.uiowa.edu/casma/). When downloading it, an extended manual is included, together with the data and an input file to compute Tables 7.9 and 7.10. Although there are quite a lot of technical variations in the use of the program, the default values usually will give good results.

information one has used in the standard setting to the same degree, and not all evidence is necessarily equally convincing.

- Evidence may be provided from the results of the same students on another test or assessment procedure, or from results from other students on the same or another test.
- Evidence may be provided from another standard setting procedure using the same panel or an independent panel, led by the same staff members or by independent staff.

This is a summary of the kind of evidence that might be provided to justify the claim of generality emanating from the decision rules of one's own procedures of linking. One could take the attitude "Let's do it all", but this is unrealistic because the collection of some evidence may be fairly expensive, and not all studies giving corroborating results will be equally successful.

In this section some examples of external validation procedures will be discussed, and arguments as to their limits and persuasiveness (or lack of it) will be put forward. But first a general remark is in order. In test theory, the external validity problem is usually approached by showing the correspondence between test results and some external criterion. Sometimes the external criterion measures are considered as absolute in some sense. But actually no criterion is perfectly valid. Take educational success as an example. Obtaining a master's degree from university can in general be observed without measurement error, as this is mainly a clerical activity. As a criterion of mental abilities a master's degree is certainly useful but it is not absolute, because some students may fail at the university for reasons quite independent of their mental abilities and probably some students will succeed undeservedly, as no examination system is foolproof. Therefore it is preferable to consider all criterion measures as fallible in the same way that all tests are fallible, i.e. part of their variance is unwanted or irrelevant for showing the validity of a test procedure, such as the results of a standard setting.

### 7.5.1. Cross Validation

As discussed in Chapter 6, the main weakness of two popular examinee centred methods, the Contrasting Groups method and the Borderline Group method, is the fact that the information on the students involved stems in some sense from a non-disclosed source, the judgment of their own teacher. This judgment can (and should) be considered as a test result, but in general it is quite hard to get information on the psychometric qualities of such judgments. There is no opportunity for discussion of these results, as they are the private opinions of the teachers.

Moreover, in setting the standards with these methods, constructing decision tables has been advised so as to maximise the correspondence between test score and the judgments of the teachers. This implies that the standards arrived at are to a substantial degree dependent on the opinion of a small number of teachers on a (usually) small or at best moderately sized sample of students, so that the results may be dictated to an unknown degree by the peculiarities of this sample. In statistical terms this effect is known as *capitalising on chance*, and it is important to show how significant this effect is, by a technique called *cross validation*. In principle this technique is simple: use the results (cut-off scores) issuing from the standard setting procedure and apply them to an independent sample. Comparison of indices of quality on the original sample and the cross validation sample give an indication on the generalisability of the results. As an index of the quality here, the index of absolute agreement or Cohen's kappa may be used, as all students are allocated to a level by the teacher's judgment and by the decision rule issuing from the standard setting.

There are several ways to carry out such a cross validation:

- The original sample one has can be split (at random!) in two halves. One half is used to carry out the standard setting procedure, the other half is used for cross validation purposes. Or one could even proceed in a more balanced way by using each half sample for standard setting and the other half for cross validation, as the standard setting proper only consists in constructing tables and making decisions from them. Although such a procedure is certainly worthwhile and advisable, it is only meaningful if the total sample is large enough to yield two half samples of substantial size. Moreover, its power of

persuasion is rather limited. The criterion information originates from the same sources (the teachers), and if they happen to have a tendency to be too lenient, for example, this will not be detected in the cross validation.

- To control for this problem, one can split the sample of students so as to have all students from half of the teachers as a standard setting sample and the other half as the cross validation sample. Or, if sample sizes are large, one could even proceed with constructing four samples, first by splitting the teachers in two halves and then by splitting the sample of students of each teacher in two equivalent halves.
- The preceding procedure can be easily understood as a special case of genuine validation. If the sample size used for standard setting is not large enough to split, one can use the whole sample to set the standards and then collect data on a totally independent sample, coming from other schools. Validation will require administering the test (or examination) on this validation sample as well as asking judgments from the teachers on the CEFR level of the students in the sample. But in principle, this procedure does not differ from the previous one, as standard setting sample and validation sample can easily change roles.

In the standard setting methods discussed in Chapter 6, the Contrasting Groups method and the Borderline Group method have a special status stemming from the fact that a criterion measure (the judgment by the teachers) is a constituent part of the standard setting method itself. One might think that this is necessarily the case for all examinee centred methods, but this is not true. Take the Body of Work method as a good example. In this method all the information the panel members get on the students is their test performance, and some information of the ranking of the dossiers (although this is not strictly necessary). No information whatsoever as to the CEFR level the students are at is provided to the panel members. The method is confined completely to student performance on the examination. Much the same holds for all test centred methods discussed in Chapter 6: the standards arrived at are completely determined by the judgments of the panel members on the testing material. Even giving them impact information (as to the distribution of students across levels) confronts them only with the consequences of their own judgments and does not reflect a categorisation in CEFR levels coming from another source. Therefore, the concept of cross validation does not make much sense for these methods.

External validation of these standard setting procedures therefore will involve comparison of the results of the standard setting procedure (the decision rule) with the results of another decision rule. This comparison may take essentially two forms: using only marginal distributions or cross tabulations. These are discussed in turn.

### 7.5.2. Comparison of Marginal Distributions

Suppose data from a representative sample have been calibrated using an IRT model, and a decision rule to allocate students to four CEFR levels, say, has been derived using a Bookmark method. Then the students belonging to the calibration sample may be categorised in one of the four levels. If one has information on another sample, being representative for the same target population, and being categorised using another method, e.g., the judgment of their teacher, one could construct a two by four table as displayed in Table 7.11. In the table, Sample 1 refers to the calibration sample, and Sample 2 to an independent validation sample.

**Table 7.11: Marginal Distributions Across Levels (Frequencies)**

	A1	A2	B1	B2	Total
Sample 1	98	124	165	84	471
Sample 2	39	74	78	63	254
Total	137	198	243	147	725

As the two samples are of different size, comparison by mere inspection of the table is difficult. Converting the frequencies to row-wise percentages makes the comparison easier. The results are displayed in Table 7.12, showing that in the independent sample relatively more students are assigned to the Levels A2 and B2

and less to A1 and B1 than in the calibration sample. One can test this difference statistically by a chi-square test. The test statistic in this example is 7.94 and its associated p-value is 0.047 (with three degrees of freedom), meaning that there is a significant difference in level allocation due to the two methods<sup>44</sup>.

**Table 7.12: Marginal Distributions Across Levels (Percentages)**

	A1	A2	B1	B2	Total
<b>Sample 1</b>	20.8	26.3	35.0	17.8	100.0
<b>Sample 2</b>	15.4	29.1	30.7	24.8	100.0

This example, however simple it may be, already illustrates how difficult the process of validation is. On statistical grounds (the chi-square test), it may be concluded that there are systematic differences in allocation to the CEFR levels based on the two methods, but from this finding it does not follow why these differences are there. Take Level B2, the case with the largest difference in the percentage of allocation, as an example. It may be the case that the Bookmark method has led to too severe a standard for B1/B2. However, this cannot be deduced from the table, because it may also be the case that the teachers have been too lenient in assigning a B2 qualification. Finding out what is really happening here may require a lot of further study and data collection. Interviewing the teachers on their reasoning and rationale to give a B2 qualification might reveal that they have not well understood the CEFR description of B2. Alternatively they have been one sided in their judgments, just paying attention to only a few of the typical B2 “Can Do” statements, while ignoring others, which have perhaps been of much importance during the discussions in the bookmark standard setting method. Conversely, the examination used to set the standards may be too one-sided, and have neglected a number of aspects which experienced teachers take into account when asked to give a holistic judgment on the level of their students. A table like Table 7.12 can be used to point to the problem and at best to suggest a possible explanation; a lot of creativity, however, will be needed to detect the real causes of the differences.

### 7.5.3. Bivariate Decision Tables

More information may be gained if two sets of decision rules can be applied to the same sample of students. The results of a standard setting method (the decision rules) can usually be applied directly to a sample of students, e.g., a calibration sample. If one has another set of decision rules, either coming from holistic judgments of teachers, or from another standard setting method, and these rules can be applied to *the same sample* of students then one can construct a bivariate decision table giving the joint probabilities (or frequencies) of allocation to all possible pairs of levels. These tables are comparable to the right-hand panel of Table 7.10, with this – *essential* – difference: the columns refer to the allocations based on the method under study (as is the case in Table 7.10), but the rows are based on the allocation to levels based on an *independent* set of decision rules, and not on some model assumptions as was the case in judging decision consistency. If the independent set of decision rules really means the same as the decision rules arrived at in the standard setting method, i.e., if both have the same construct validity and the same reliability, then the bivariate decision table should essentially be the same as the right-hand panel in Table 7.10. Therefore, constructing and comparing both tables may reveal useful information:

- Marginal distributions may be compared much in the same way as discussed above with independent samples.
- Indices of agreement (absolute agreement, adjacent agreement and Cohen’s kappa) may be computed on both tables and be compared.
- Most relevant for validation is the comparison of the off diagonal cells in both tables. It has been said above that in judging the decision consistency, the bivariate decision table will be essentially symmetric. In the case of validation with another set of decision rules (the criterion decision rules) the symmetry or lack of symmetry is a purely empirical finding, and may be helpful in understanding the validity of the standard setting method. The concept of false positives and false negatives becomes important here, but

<sup>44</sup> The chi-square test must be carried out using the frequencies (Table 7.11), not the percentages as in Table 7.12.

one has to clearly define what is meant by these terms in a validation context. It may be helpful to define *false negatives* as the cases where the decisions following from the standard setting under study lead to a *lower* level than the criterion rules; *false positives* refer to the cases where the standard setting rules lead to a *higher* level. If in the validation study, the rate of false positives is higher than that of the false negatives, this means that the standard setting under study is more lenient than the criterion rules; in the opposite case it is more harsh<sup>45</sup>.

**Worked Example.** A worked example may help illustrate how bivariate decision tables may be used to relate test results to other assessment data, for example holistic ratings by teachers of a CEFR level. The principle of using bivariate tables is not complex in itself. The main problem with the use of teacher holistic judgments as an external criterion in this way is not the analysis. It is the fact that it requires that the teachers really do intimately know (a) the CEFR levels and (b) the competence of the individuals concerned; this may not be practical with teachers in mainstream education who see classes of 30 only a couple of times a week.

North (2000b) reports using class teacher judgments as an external criterion to reference item banks for English, German, French and Spanish onto the Eurocentres scale, which distinguishes nine levels. Provisional standard setting had been done previously with a simplified variant of the Bookmark method. In the external validation study, teacher ratings were used to verify through independent external validation the standard setting carried out during the development of an item bank for German. Class teachers were asked to allocate each student in their class to a level for the area tested by the item bank: knowledge of the language system. Figure 7.3 shows the relation between the performance standards (on the X-axis) and teacher judgments (criterion) on the Y-axis.

Teacher Judgment	9								
	8						3	1	2
	7				1	8	8		
	6				2				
	5			4	8	2			
	4		5		1				
	3		5	6	4				
	2		1	2					
	1	4	1						
	1	2	3	4	5	6	7	8	9
	Performance Standards								

**Figure 7.3: Bivariate Decision Table Using Nine Levels**

The relationship between classification by the performance standards and teacher classifications appears regular and balanced, with a correlation of 0.93. Nevertheless, only 28 of the 68 subjects (41%) have actually received exactly the same grade, despite the high correlation. There are eight learners placed at Level 7 by the teacher(s) and Level 6 by the program. This was caused by a single lenient teacher. However, even if these eight test takers were in “the right place” on the chart still only about 50% of the students would have received exactly the same rating from the teacher and the test. The index of adjacent agreement, however, is  $67/68 = 0.985$ : only one student has been placed two categories higher by teacher judgment than by using the standards.

The Eurocentres scale splits the CEFR levels into two (apart from Level A1). If a bivariate decision table is created using only CEFR levels, as shown in Table 7.13, the proportion of correct classifications increases

<sup>45</sup> More sophisticated analyses may be done here, as for example using the versatile family of log-linear analyses to find more locally situated significant differences. For more information, see, e.g., Fienberg (1977) for an easily accessible introduction or Fienberg et al (1975) for a thorough treatment.

considerably – from 41% to 73.5%, since 50 of the 68 learners now receive the same CEFR level from both standard setting and teachers<sup>46</sup>. The index of adjacent agreement equals one.

		Performance standards					
		A1 (1)	A2 (2 & 3)	B1 (4 & 5)	B2 (6 & 7)	C1 (8 & 9)	Total
Criterion (Teachers)	C1 (8 & 9)				3	3	5
	B2 (6 & 7)			3	16		18
	B1 (4 & 5)		5	13	2		20
	A2 (2 & 3)		14	4			19
	A1 (1)	4	1				6
	Total	4	20	20	21	3	68

**Figure 7.4: Bivariate Decision Table Using Five Levels**

If teacher assessments are used, it is good practice to consider such a judgmental procedure as a form of test and to pay attention to its internal validity as a test. Some relevant and challenging questions in this respect are listed below:

- If the judgment is a single holistic judgment, how then can one assess its reliability? From a psychometric point of view, this amounts to using a one item test; so there is no room to compute indices of internal consistency. In such a case, one should devise some retesting procedure, and one has to consider all the problems with the feasibility of a repeated judgment.
- Even with judgments with checklists of descriptors, the rater needs to know well the competence of the candidate. As with the examinee-centred standard setting procedures discussed in Chapter 6, this in turn implies that the rater can only judge a limited number of test takers (his/her own students). An added problem in using teachers to rate their own learners is that they may then exaggerate the differences between their stronger and weaker learners.

Multiple judges, giving judgments on more easily observed samples of behaviour like written texts can help to avoid the last two problems outlined above. The use of judges that are independent of the standard setting process, properly trained, and given appropriate rating instruments (see Section B of the Reference Supplement) is an option that has been used successfully in Finland. Rater variance could then be studied with a G study (see Section E of the Reference Supplement) or a many-faceted Rasch analysis (Linacre 1989), e.g. as operationalised in the program FACETS (Linacre 2008). This model takes into account a third facet (the rater), estimates rater severity/lenience, and takes account of it in arriving at ability estimates for the test takers.

#### 7.5.4. Some Scenarios

It has been pointed out in the previous paragraphs that all validation procedures aim at comparing different sets of decision rules, either using independent samples of students or the same sample. In this subsection a few scenarios will be described, which may be helpful in making a decision on what is a rational (or wise) comparison.

<sup>46</sup> Expressing Eurocentres levels in CEFR terms can be justified because a considerable number of the CEFR descriptors originate from the Eurocentres scale. This is because Eurocentres descriptors survived the qualitative validation process better than those from most other source scales, since Eurocentres formulations tended to be concrete and positive. The correlation of rank order placement for 73 common descriptors for interaction and production is 0.88. The shared classification shown by a Decision Table is 70%. (See North 2000a: 337.)

An important distinction between standard setting methods is the difference between examinee centred and test centred methods. It seems natural therefore to focus the validation of a method belonging to one class on a comparison with a method belonging to the other class. One should, however, not be overoptimistic as to the possibilities in accomplishing such a comparison. Let us take the Bookmark method (or its Cito variation) and the Body of Work method as an example of a suitable pair of contrasting standard setting methods. There are a couple of arguments pleading against such a scenario:

- The Bookmark method is suited for tests or examinations which can be successfully calibrated using IRT, i.e., highly itemised tests, while the Body of Work method, aiming at holistic judgments, is particularly suited for examinations which are usually not well suited for IRT modelling, like speaking or writing tests. The consequence will be that at least one of the methods will suffer from some kind of inappropriateness, which makes comparisons void.
- Even if an examination has a degree of complexity such that it allows for standard setting methods as different as the Bookmark method and the Body of Work method, implementing both of them, may be unrealistic from a practical point of view, as both methods require their specific training<sup>47</sup> and as a rule are time consuming. Lack of resources or boredom and fatigue of the panel members may be prohibitive for such a complex approach.

On the other hand, it is always possible (if resources are sufficient) to apply two different standard setting methods using two independent panels of judges, and implement the two methods at different times. The implementation of two high cost procedures may not be appropriate in local standard setting contexts, but it may be relevant in projects with far reaching and internationally relevant consequences.

An attractive compromise may be found in combining a test centred method with the Contrasting Groups method or the Borderline Group method, if the panel members can give holistic judgments on a sufficient large number of students who have taken the test under study. But see the worked example above.

#### **7.5.4.1. Taking Advantage of IRT Calibration**

Using an IRT-model to relate items or tasks to each other offers a number of opportunities to validate different standard setting methods against each other. In these approaches, advantage is taken of the fact that the relation of the items to the underlying (latent) ability is known (to a sufficient degree of accuracy) from a calibration study. Here we will describe a scenario which uses this relation explicitly in a validation study of a particular standard setting procedure. To exemplify the point, we take the Cito variation of the Bookmark method as the preferred method.

The method implies two standard setting procedures using a different collection of items to be presented to the panel members. Panel members in both procedures may be the same or different persons. In the latter case one should take care that the panels composed to do the standard setting are comparable across the two sessions – selected to present two parallel representative groups. The collection for the first standard setting procedure may consist of all the items (or a subset of them) to be used in Examination A, while the second set of items contains items (all or a subset) from Examination B. As is the case in all IRT-based standard setting procedures, the cut-off points are defined in the latent variable domain. Using techniques discussed in Section 6.8.3, these latent variable standards may be translated into cut-off scores of any test whose item characteristics are known. In particular, one might translate to a test which mainly consists of items used in the standard setting or to a test which mainly consists of items not used in the standard setting. The situation is depicted as a summary in Table 7.13.

The shaded cells are the conditions where the item material used for the standard setting have a close relationship (being identical to or being a large subset of) the items used in the examination. The blank cells are the more vulnerable ones: the items used to set the standards are items other than the ones really used in the examination.

---

<sup>47</sup> Usually, panel members invited for participating in standard setting for language examinations do not know very much about IRT. Giving them an introduction into this area, which is at the same time correct and simple, is a difficult and time consuming task, which should not be underestimated.

**Table 7.13: Design for a Paired Standard Setting**

	Standard setting based on items belonging to	
	Examination A	Examination B
Cut-off scores for ex. A		
Cut-off scores for ex. B		

As (virtually) nobody takes both examinations, empirical comparisons are only sensible within single rows of Table 7.13, meaning essentially that for the same examination two sets of cut-off scores have been set, and that the extent to which they lead to the same or different conclusions may be checked empirically by constructing decision tables as described earlier in this section and exemplified with Table 7.12. This evaluation procedure may be applied to both rows of Table 7.13, offering an opportunity to check whether an explanation of differences in outcomes (due to the two different standard setting procedures; i.e., one row in Table 7.13) is consistent with the differences found with the same standard setting methods on another occasion (i.e., the complementary row in Table 7.13).

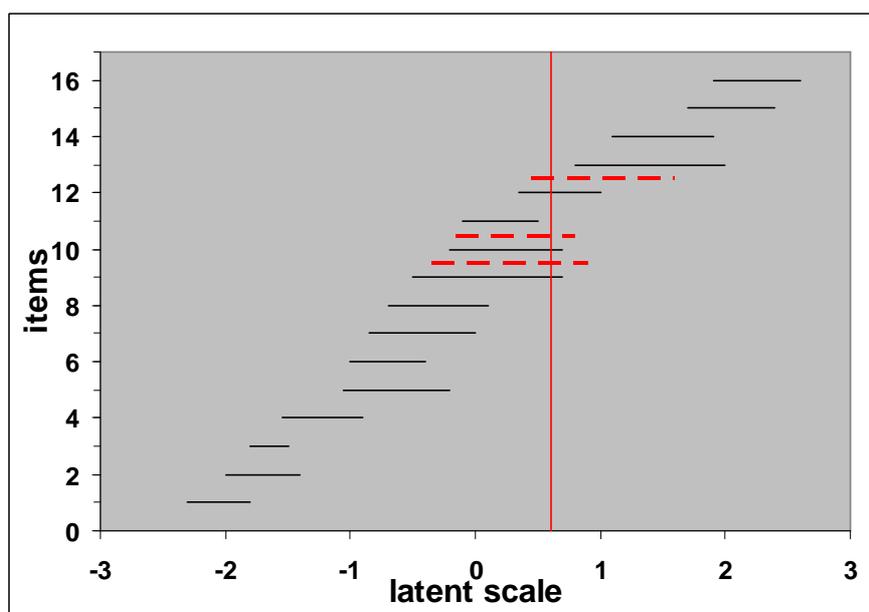
#### 7.5.4.2. Using “Can Do” Statements.

A method to exploit the CEFR very directly in external validation is to rate the candidates who will be providing the data for the test under study on European Language Portfolio-style checklists made up of 30–50 relevant CEFR descriptors. In this way, each descriptor can be included as a separate item in the IRT analysis alongside the test items, and so be calibrated onto the same latent ability scale. The source of the judgments could be class teachers, or the candidates themselves through self-assessment.

In combination with the Cito variant of the Bookmark method, this information can be used to validate the standard setting as exemplified in Figure 7.5. This Figure is the same as Figure 6.5, with the only exception that three “Can Do” statements (calibrated as items) have been added to the display. Suppose the standard for A2/B1 has been set as indicated by the vertical line in the display by the method described in Section 6.9. Suppose further that the three dashed lines in the display represent three “Can Do” statements for Level B1. For the bottom two one sees that at the performance standard “Full mastery” has almost been reached, while for the top one there is not much more than borderline mastery. This information (collected preferably with more than three “Can Do” statements) together with the content of these “Can Do” statements gives a quite detailed picture of what the standard means directly in terms of CEFR descriptors.

This way of validating the standard setting can be used in at least two different ways. A figure like Figure 7.5 can be constructed after the standard setting procedure has been finished to judge the validity of the standard setting outcomes. Such an approach conceives of standard setting and validation as a linear process. The judgments in relation to descriptors are used as an external criterion in a study of external validity. But if the results of the validation are disappointing – showing for example that the panel members have been too lenient as a group – the whole standard setting procedure can be seen as a failure and a waste of time and resources. A more effective approach is to incorporate this kind of information into the standard setting procedure itself – for example between two judgments rounds – as useful information about the consequences of setting the standards and as arguments to adapt earlier judgments made.

It is true that in the latter approach, the validation is not genuinely independent of the standard setting procedure itself – as the information on the “Can Do” items is used in the procedure itself – but it can save time consuming repetition of the whole procedure. Good documentation of the results of all the judgment rounds can be just as convincing as a validity argument as a completely independent validation (external validation in a classic sense).



**Figure 7.5: Item Map with Test Items and “Can Do” Statements**

The source of the data for “Can Do” approaches can be either teacher assessments or self-assessments. The choice between self-assessment and teacher assessment is in principle problematic. Reliance solely on self-assessment data may lead to a – possibly incorrect – conclusion that the standard setting has been too strict. Therefore it is good practice to collect both teacher and self-assessment data and to add in this way to the strength of the validity argument.

#### **7.5.4.3. Cross Language Standard Setting**

In Chapter 6 (Section 6.8.3), a general procedure has been described for linking different examinations or tests (e.g. a test of French and a test of English) to the CEFR. The procedure leans heavily on the plurilingualism of the panel members, as it cannot be taken for granted that any student is equally proficient in two different languages. The vulnerable aspect of this procedure is whether it can be taken for granted that all panel members involved in the standard setting procedure are sufficiently proficient in the languages concerned.

To have a clear idea of what the results mean, one has to implement a procedure that controls for this. Taking English and French as example languages, one has to take care of the following:

- In the standard setting involving the two languages, a balance has to be found as to the background of the panel members. This might mean that half of them are native English speakers and the other half are French native speakers, while everybody has the other language as main specialisation.
- A monolingual standard setting procedure for each language may also be advisable, since a plurilingual context may create special settings (caused by the unusual context for example), which make the results unsuitable for generalisation.

These two considerations already imply a quite complicated design to test the validity of the standard setting suitable for cross language validation. Ideally we need:

- a mixed language standard setting procedure in which half of the panel members have English as their native language and French as first specialisation, and the other half have French as native language and English as first specialisation;

- a monolingual standard setting for French in which half of the panel members are native French-speakers and the other half have French as their first specialisation;
- a monolingual standard setting for English in which half of the panel members are native English-speakers and the other half have English as their first specialisation;

Preferably, the three conditions sketched above should consist of independent panels. Implementing such a design offers a possibility to compare standards across languages, and, through gaining and sharing expertise in cross-language standard setting, to offer suggestions on how to improve or even discard the procedure. Experience was gained with the Cross-language benchmarking seminar held at Sèvres in June 2008 (Breton et al forthcoming).

## 7.6. Conclusion

The discussion on external validation in this chapter may look disappointing in a number of respects, as 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 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. In contrast, we believe that what constitutes a “B1” is essentially a practical convention, but formulated so clearly and consistently that if two language professionals refer to it, they mean essentially the same thing, even if their own cultural and linguistic background is different and they are referring to different target languages. The CEFR constitutes a frame of reference intended to make such statements possible. From the perspective of validation studies, this means that every validation study can, in principle, offer constructive criticism that may lead to a refined, more elaborated and balanced frame of reference. This is true of all empirical testing of hypotheses, constructs and theories.

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, and as a consequence, it is not realistic to expect a definite verdict in any particular case.

Is this good news or bad news? We think it is just the state of the art. More definite conclusions may be drawn from a well designed meta-analysis, which can summarise the results of a large number of well designed validation studies conducted over the next few years. It is the responsibility of the present generation to provide the necessary data and documentation for such a meta-analysis to be meaningful. (See Plake 2008 for a good review of challenges and a set of thoughtful recommendations.)

Thus, it is to be hoped that many standard setting endeavours, under way or planned in the future, drawing on the information provided in this Manual, the Reference Supplement and other relevant sources, are conducted and reported in a transparent manner. By analysing and comparing them, standard setting know-how will increase, the defensibility of decisions on standards will improve and the awareness of the consequences of standard setting will be heightened.

*Users of the Manual may wish to consider:*

- *how the required validity evidence can best be obtained*
- *what techniques they will be able to apply and to what extent they may need outside technical support*
- *whether they can “build a validity argument” about the quality of the test and procedures associated with it (internal validity), the quality of the procedures followed in the linking project and in particular in the standard setting (procedural validity), and the corroboration of the result from independent analyses (external validity)*
- *how they ensure that standards are comparable across languages, if this is relevant*
- *whether, in particular, there is sufficient evidence supporting the validity of the established cut-off score*
- *how they will make their detailed findings available to professional colleagues*

## References

- AERA/APA/NCME (1999): American Educational Research Association, American Psychological Association, National Council on Measurement in Education: *Standards for Educational and Psychological Testing*. Washington, D.C.: American Educational Research Association. (ISBN 0-935302-25-5)
- Alderson, J. C. (2005): *Diagnosing Foreign Language Proficiency*. London: Continuum.
- Alderson, J. C., Clapham, C. and Wall, D. (1995): *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.
- Alderson, J. C., Figueras, N., Kuijpers, H., Nold, G., Takala, S. and Tardieu, C. (2006): Analysing Tests of Reading and Listening in relation to the CEFR: the experience of the Dutch CEFR Construct Project. *Language Assessment Quarterly* 3 (1): 3–30.
- American Educational Research Association (1999): *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.
- Angoff, W. H. (1971): Scales, Norms and Equivalent Scores. In: Thorndike, R. L. (ed.) *Educational Measurement* (2nd Edition), pp. 508–600. Washington, D.C.: American Council on Education.
- Beacco, J-C. and Porquier, R. (2008): *Niveau A2 pour le français : Un référentiel*. Paris: Didier.
- Beacco, J-C., Porquier, R. and Bouquet, S. (2004): *Niveau B2 pour le français : Un référentiel*. Paris: Didier. (2 vols)
- Beacco, J-C., De Ferrari, M., Lhote, G. and Tagliante, C. (2006): *Niveau A1.1 pour le français / référentiel DILF livre*. Paris: Didier.
- Beacco, J-C., Porquier, R. and Bouquet, S. (2007): *Niveau A1 pour le français : Un référentiel*. Paris: Didier.
- Berk, R.A. (1986): A Consumer's Guide to Setting Performance Standards on Criterion Referenced Tests. *Review of Educational Research*, 56, 137–172.
- Bolton, S., Glaboniat, M., Lorenz, H., Müller, M., Perlmann-Balme, M. and Steiner, S. (2008): *Mündlich: Mündliche Produktion und Interaktion Deutsch: Illustration der Niveaustufen des Gemeinsamen europäischen Referenzrahmens*. Berlin: Langenscheidt.
- Breton, Jones, Laplannes, Lepage and North, (forthcoming): *Séminaire interlangues / Cross language benchmarking seminar, CIEP Sèvres, 23–25 June 2008: Report*. Strasbourg: Council of Europe.
- Cizek, G. J. (ed.) (2001): *Setting Performance Standards: concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum.
- Cizek, G.J. and Bunch, M.B. (2007): *Standard Setting: a guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage.
- Cohen, A., Kane, M. and Crooks, T. (1999): A Generalized Examinee-Centered Method for Setting Standards on Achievement Tests. *Applied Measurement in Education*, 12, 343–366.
- Council of Europe (2001a): *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Council of Europe (2001b): *Cadre européen commun de référence pour les langues: Apprendre, enseigner, évaluer*. Paris: Didier.
- Council of Europe (2002): *Seminar on Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF), Helsinki, 30 June 30–2 July 2002: Report*. DG IV / EDU / LANG (2002) 15. Strasbourg: Council of Europe.
- Council of Europe (2003): *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment ("CEFR")* DGIV/EDU/LANG (2003) 5. Strasbourg: Council of Europe.
- Davidson, F. and Lynch, B. (1993): Criterion-referenced language test development: a prolegomenon. In: Huhta, A., Sajavaara, K. & Takala, S. (eds.), *Language Testing: New Openings*. Jyväskylä, Finland: University of Jyväskylä, pp.73–89.
- Davidson, F. and Lynch, B. (2002): *Testcraft: A Teacher's Guide to Writing and Using Language Test Specifications*. Yale University Press.
- Downing, S. M. and Haladyna, T. M. (eds.) (2006): *Handbook of Test Development*. Earlbaum.
- Ebel, R. L. and Frisbee, O. A. (1986): *Essentials of Educational Measurement (4<sup>th</sup> edition)*. Englewood Cliffs, N.J.: Prentice Hall.
- Feldt, L. S., Steffen, M. and Gupta, N. C. (1985): A Comparison of Five Methods for Estimating the Standard Error of Measurement at Specific Score Levels. *Applied Psychological Measurement*, 9, 351–361.

- Ferrara, S., Perie, M. and Johnson, E. (2002): *Matching the Judgmental Task with Standard Setting Panelist Expertise: the item-descriptor (ID) matching procedure*. Washington DC: American Institutes for Research.
- Fienberg, S. E. (1977): *The Analysis of Cross-classified Categorical Data*. Cambridge, Massachusetts: The MIT Press.
- Fienberg, S.E., Bishop, Y. M. M. and Holland, P. W. (1975): *Discrete Multivariate Analysis*. Cambridge (Massachusetts): The MIT Press.
- Glaboniat, M., Müller, M., Schmitz, H., Rusch, P., Wertenschlag, L., (2002/5): *Profile Deutsch*. Berlin: Langenscheidt, ISBN 3-468-49463-7.
- Hambleton, R.K. and Pitoniak, M.J. (2006): Setting Performance Standards. In Brennan, R.L. (ed.) *Educational Measurement* (4th edition). Westport, CT: American Council on Education/Praeger, pp. 433–470.
- Instituto Cervantes (2007): *Niveles de Referencia para el español, Plan Curricular del Instituto Cervantes*. Madrid: Biblioteca Nueva.
- Jaeger, R. M. (1991): Selection of Judges for Standard-setting. *Educational Measurement: Issues and Practice*, 10, 3–6.
- Kaftandjieva, F. (2007): Quantifying the Quality of Linkage between Language Examinations and the CEF. In Carlsen, C. and Moe, E. (eds.) *A Human Touch to Language Testing*. Oslo: Novus Press, 34–42.
- Keats, J. A. (1957): Estimation of Error Variances of Test Scores. *Psychometrika* 22, 29–41.
- Kingston, N. M., Kahl, S. R., Sweeny, K. P. and Bay, L. (2001): Setting Performance Standards using the Body of Work Method. In Cizek G. J. (ed.), *Setting Performance Standards: Concepts, methods and perspectives*. Mahwah, NJ: Erlbaum, pp. 219–248.
- Kolen, M. L. and Brennan, R-L. (2004): *Test Equating, Scaling and Linking*. New York: Springer.
- Lepage, S. and North, B. (2005): *Guide for the organisation of a seminar to calibrate examples of spoken performance in line with the scales of the Common European Framework of Reference for Languages*. Strasbourg: Council of Europe DGIV/EDU/LANG (2005) 4.
- Linacre, J. M. (1989): *Multi-faceted Measurement*. Chicago: MESA Press.
- Linacre, J. M. (2008): *A User's Guide to FACETS. Rasch Model Computer Program*. ISBN 0-941938-03-4. www.winsteps.com.
- Livingston, S. A. and Lewis, C. (1995): Estimating the Consistency and Accuracy of Classification based on Test Scores. *Journal of Educational Measurement*, 32, 179–197.
- Lord, F. (1965): A Strong True-score Theory, with Applications. *Psychometrika*, 30, 239–270.
- Lynch, B. and Davidson, F. (1994): Criterion-referenced language test development: linking curricula, teachers and tests. *TESOL Quarterly* 28:4, pp. 727–743.
- Lynch, B. and Davidson, F. (1998): Criterion Referencing. In: Clapham, C. & Dorson, D. (eds.) *Language Testing and Assessment*, Volume 7, Encyclopedia of Language and Education. Dordrecht: Kluwer Academic Publishers, pp. 263–273.
- Milanovic, M. (2002): *Language Examining and Test Development*. Strasbourg: Language Policy Division, Council of Europe.
- Mitzel, H. C., Lewis, D. M., Patz, R. J. & Green, D. R. (2001): The Bookmark Procedure: psychological perspectives. In Cizek G. J. (ed.) *Setting Performance Standards: concepts, methods and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.
- Norcini, J., Lipner, R., Langdon, L., and Strecker, C. (1987): A Comparison of Three Variations on a Standard-Setting Method. *Journal of Educational Measurement*, 24, 56–64.
- North, B. (2000a): *The Development of a Common Framework Scale of Language Proficiency*. New York: Peter Lang.
- North, B. (2000b): Linking Language Assessments: an example in a low-stakes context. *System* 28, 555–577.
- North, B. and Schneider, G. (1998): Scaling descriptors for language proficiency scales. *Language Testing* 15/2: 217–262.
- OECD (2005): *Pisa 2003 Technical Report*. Paris: OECD.
- Parizzi, F. and Spinelli, B. (forthcoming): *Profilo della Lingua Italiana*, Firenze: La Nuova Italia.
- Plake, B. S. (2008): Standard Setters: Stand Up and Take a Stand! *Educational Measurement: Issues and Practice* 27/1: 3–9.
- Reckase, M. D. (2006a): A Conceptual Framework for a Psychometric Theory for Standard Setting with Examples of Its Use for Evaluating the Functioning of Two Standard Setting Methods. *Educational Measurement: Issues and Practice*, 2006, 25(2), 4–18.

- Reckase, M. D. (2006b): Rejoinder: Evaluating Standard Setting Methods Using Error Models Proposed by Schulz. *Educational Measurement: Issues and Practice*, 2006, 25 (3), 14–17.
- Schneider, G. and North, B. (2000): *Fremdsprachen können – was heisst das? Skalen zur Beschreibung, Beurteilung und Selbsteinschätzung der fremdsprachlichen Kommunikationsfähigkeit*. Chur/Zürich: Ruediger Verlag.
- Siegel, S. and Castellan, N. J. (1988): *Non-parametric statistics for the behavioral sciences*. New York: McGraw-Hill.
- Subkoviak, M. J. (1988): A Practitioner's Guide to Computation and Interpretation of Reliability for Mastery Tests. *Journal of Educational Measurement*, 13, 265–276.
- Thorndike, R.L. (ed.) (1971): *Educational Measurement* (2nd Edition), pp. 508–600. Washington, D.C.: American Council on Education.
- Van der Schoot, F. (2001): *Standaarden voor Kerndoelen Basisonderwijs* [Standards for Primary Objectives in Primary Education]. PhD thesis. Arnhem: Cito.
- van Ek, Jan A. (1976): *The Threshold level in a European Unit/credit System for Modern Language Learning by Adults*. Strasbourg: Council of Europe.
- van Ek, J. A. and Trim, J. L. M., (2001a): *Waystage*. Cambridge: CUP, ISBN 0-521-56707-6
- van Ek, J. A. and Trim, J. L. M., (2001b): *Threshold 1990*. Cambridge: CUP, ISBN 0-521-56707-8
- van Ek, J. A. and Trim, J. L. M., (2001c): *Vantage*. Cambridge: CUP, ISBN 0-521-56705-X
- Verhelst, N. D. and Verstralen, H. H. F. M. (2008): Some Considerations on the Partial Credit Model. *Psicológica*, 29, 229–254.
- Weir, C. (1993): *Understanding and Developing Language Tests*. Hemel Hempstead UK: Prentice Hall.

# **Appendices**

## **A. Forms and Scales for Description and Specification Chapters 1 & 4**

**Section A1: Salient Characteristics of CEFR Levels (Chapter 1)**

**Section A2: Forms for Describing the Examination (Chapter 4)**

**Section A3: Specification: Communicative Language Activities (Chapter 4)**

**Section A4: Specification: Communicative Language Competence (Chapter 4)**

**Section A5: Specification: Outcome of the Analysis (Chapter 4)**

## **B. Content Analysis Grids Chapter 4**

**Section B1: CEFR Content Analysis Grid for Listening & Reading**

**Section B2: CEFR Content Analysis Grids for Writing and Speaking Tasks**

## **C. Forms and Scales for Standardisation & Benchmarking Chapter 5**

## Section A1: Salient Characteristics of CEFR Levels Chapter 1

Level		Table A1. Salient Characteristics: Interaction & Production (CEFR Section 3.6, simplified)
Proficient User		It cannot be overemphasised that Level C2 is not intended to imply native speaker competence or even near native speaker competence. Both the original research and a project using CEFR descriptors to rate mother-tongue as well as foreign language competence (North 2002: CEFR Case Studies volume) showed the existence of ambilingual speakers well above the highest defined level (C2). Wilkins had identified a seventh level of “Ambilingual Proficiency” in his 1978 proposal for a European scale for unit-credit schemes.
	C2	<b>Level C2</b> is intended to characterise the degree of <b>precision, appropriateness and ease with the language</b> which typifies the speech of those who have been highly successful learners. Descriptors calibrated here include: <i>convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices; has a good command of idiomatic expressions and colloquialisms with awareness of connotative level of meaning; backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.</i>
	C1	<b>Level C1</b> is characterised by a <b>broad range of language, which allows fluent, spontaneous communication</b> , as illustrated by the following examples: <i>Can express him/herself fluently and spontaneously, almost effortlessly. Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. There is little obvious searching for expressions or avoidance strategies; only a conceptually difficult subject can hinder a natural, smooth flow of language.</i> The discourse skills appearing at B2+ are more evident at C1, with an emphasis on more fluency, for example: <i>select a suitable phrase from a fluent repertoire of discourse functions to preface his remarks in order to get the floor, or to gain time and keep it whilst thinking; produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.</i>
Independent User	B2+	B2+ represents a strong B2 performance. The focus on argument, effective social discourse and on language awareness which appears at B2 continues. However, the focus on argument and social discourse can also be interpreted as a new focus on discourse skills. This new degree of discourse competence shows itself in conversational management (co-operating strategies): <i>give feedback on and follow up statements and inferences by other speakers and so help the development of the discussion; relate own contribution skilfully to those of other speakers.</i> It is also apparent in relation to coherence/cohesion: <i>use a variety of linking words efficiently to mark clearly the relationships between ideas; develop an argument systematically with appropriate highlighting of significant points, and relevant supporting detail.</i>
	B2	<b>Level B2</b> represents a break with the content so far. Firstly there is a focus on <b>effective argument</b> : <i>account for and sustain his opinions in discussion by providing relevant explanations, arguments and comments; explain a viewpoint on a topical issue giving the advantages and disadvantages of various options; develop an argument giving reasons in support of or against a particular point of view; take an active part in informal discussion in familiar contexts, commenting, putting point of view clearly, evaluating alternative proposals and making and responding to hypotheses.</i> Secondly, at this level one can hold your own in <b>social discourse</b> : e.g. <i>understand in detail what is said to him/her in the standard spoken language even in a noisy environment; 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; interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without imposing strain on either party.</i> Finally, there is a <b>new degree of language awareness</b> : <i>correct mistakes if they have led to misunderstandings; make a note of “favourite mistakes” and consciously monitor speech for it/them; generally correct slips and errors if he/she becomes conscious of them.</i>
	B1+	B1+ is a strong B1 performance. The same two main features at B1 continue to be present, with the addition of a number of descriptors which focus on the exchange of <b>quantities</b> of information, for example: <i>provide concrete information required in an interview/consultation (e.g. describe symptoms to a doctor) but does so with limited precision; explain why something is a problem; summarise and give his or her opinion about a short story, article, talk, discussion interview, or documentary and answer further questions of detail; carry out a prepared interview, checking and confirming information, though he/she may occasionally have to ask for repetition if the other person’s response is rapid or extended; describe how to do something, giving detailed instructions; exchange accumulated factual information on familiar routine and non-routine matters within his field with some confidence.</i>
	B1	<b>Level B1</b> reflects the <b>Threshold Level</b> specification and is perhaps most categorised by two features. The first feature is the ability to <b>maintain interaction and get across what you want to</b> , for example: <i>generally follow the main points of extended discussion around him/her, provided speech is clearly articulated in standard dialect; express the main point he/she wants to make comprehensibly; keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.</i> The second feature is the ability to <b>cope flexibly with problems in everyday life</b> , for example <i>cope with less routine situations on public transport; deal with most situations likely to arise when making travel arrangements through an agent or when actually travelling; enter unprepared into conversations on familiar topics.</i>
Basic User	A2+	A2+ represents a strong A2 performance with <b>more active participation in conversation</b> given some assistance and certain limitations, for example: <i>understand enough to manage simple, routine exchanges without undue effort; make him/herself understood and exchange ideas and information on familiar topics in predictable everyday situations, provided the other person helps if necessary; deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words; plus significantly more ability to sustain monologues</i> , for example: <i>express how he feels in simple terms; give an extended description of everyday aspects of his environment e.g. people, places, a job or study experience; describe past activities and personal experiences; describe habits and routines; describe plans and arrangements; explain what he/she likes or dislikes about something.</i>
	A2	<b>Level A2</b> has the <b>majority of descriptors stating social functions</b> like <i>use simple everyday polite forms of greeting and address; greet people, ask how they are and react to news; handle very short social exchanges; ask and answer questions about what they do at work and in free time; make and respond to invitations; discuss what to do, where to go and make arrangements to meet, make and accept offers.</i> Here too are to be found descriptors on <b>getting out and about</b> : <i>make simple transactions in shops, post offices or banks; get simple information about travel; use public transport: buses, trains, and taxis, ask for basic information, ask and give directions, and buy tickets; ask for and provide everyday goods and services.</i>
	A1	<b>Level A1</b> is the lowest level of generative language use – the point at which the learner can <i>interact in a simple way, ask and answer simple questions about themselves, where they live, people they know, and things they have, initiate and respond to simple statements in areas of immediate need or on very familiar topics</i> , rather than relying purely on a very finite rehearsed, lexically organised repertoire of situation-specific phrases.

**Table A2. Salient Characteristics: Reception**

	<b>Setting</b>	<b>Action</b>	<b>What is understood</b>	<b>Source</b>	<b>Restrictions</b>
<b>C1</b>	<ul style="list-style-type: none"> <li>Abstract and complex topics encountered in social, academic and professional life, whether or not they relate to own field/speciality</li> </ul>	<ul style="list-style-type: none"> <li>Follow, maybe with a little difficulty</li> </ul>		<ul style="list-style-type: none"> <li>Films with a considerable degree of slang and idiomatic usage</li> <li>Poor quality, audially distorted public announcements</li> </ul>	May occasionally need to: <ul style="list-style-type: none"> <li>confirm details (with dictionary, from speaker) if outside field</li> <li>re-read difficult sections</li> </ul>
		<ul style="list-style-type: none"> <li>Understand</li> </ul>	<ul style="list-style-type: none"> <li>Finer points of detail</li> <li>Implied as well as stated opinions</li> <li>A wide range of idiomatic expressions and colloquialisms</li> <li>Register shifts</li> <li>Implied attitudes and relationships</li> </ul>	<ul style="list-style-type: none"> <li>Lengthy, complex texts of various kinds</li> <li>Extended speech – lectures, discussions, debates –even when not clearly structured</li> <li>Complex interactions between third parties in interaction and debate</li> <li>A wide range of recorded and broadcast texts, including some non-standard</li> <li>Any correspondence</li> </ul>	
<b>B2+</b>	<ul style="list-style-type: none"> <li>A wide range of familiar and unfamiliar topics encountered in social, academic and professional life</li> </ul>	<ul style="list-style-type: none"> <li>Follow, maybe with a little difficulty</li> </ul>		<ul style="list-style-type: none"> <li>Animated conversation between native speakers</li> </ul>	<ul style="list-style-type: none"> <li>Standard, non-idiomatic:</li> <li>Adequate discourse structure</li> <li>Low background noise</li> <li>May occasionally need to confirm details (with dictionary, from speaker)               <ul style="list-style-type: none"> <li>if outside field</li> <li>if above conditions not met</li> </ul> </li> </ul>
		<ul style="list-style-type: none"> <li>Understand</li> </ul>		<ul style="list-style-type: none"> <li>Spoken language, live broadcast</li> <li>Specialised texts (highly specialised if within field)</li> </ul>	
<b>B2</b>	<ul style="list-style-type: none"> <li>Reasonably familiar concrete and abstract topics related to field of interest/speciality</li> </ul>	<ul style="list-style-type: none"> <li>Follow, maybe with a little difficulty</li> </ul>	<ul style="list-style-type: none"> <li>Much of what is said</li> </ul>	<ul style="list-style-type: none"> <li>Discussion around him/her by native speakers</li> </ul>	<ul style="list-style-type: none"> <li>Standard</li> <li>Clearly signposted/signalled with explicit markers</li> <li>If native speakers talking together modify language</li> <li>If can re-read difficult sections</li> </ul>
		<ul style="list-style-type: none"> <li>Scan quickly</li> </ul>	<ul style="list-style-type: none"> <li>Relevance</li> <li>Whether closer study is worthwhile</li> <li>Specific details</li> </ul>	<ul style="list-style-type: none"> <li>Long and complex texts</li> <li>News items, articles and reports</li> </ul>	
		<ul style="list-style-type: none"> <li>Understand (with a large degree of independence)</li> </ul>	<ul style="list-style-type: none"> <li>Main ideas</li> <li>Essentials/essential meaning</li> <li>Complex lines of argument</li> <li>Speaker/writer mood, tone etc.</li> </ul>	<ul style="list-style-type: none"> <li>Extended speech: lectures, talks, presentations, reports, discussions</li> <li>Propositionally and linguistically complex text</li> <li>Technical discussions; lengthy, complex instructions; details on conditions or warnings</li> <li>Most TV and current affairs programmes</li> <li>TV documentaries, interviews, talk shows, highly specialised sources</li> <li>Announcements and messages</li> <li>Most radio documentaries, recorded audio materials</li> <li>Correspondence</li> </ul>	
<b>B1+</b>	<ul style="list-style-type: none"> <li>Common everyday or job-related topics</li> <li>Topics in his/her field of (personal) interest</li> </ul>	<ul style="list-style-type: none"> <li>Follow, though not necessarily in detail</li> </ul>	<ul style="list-style-type: none"> <li>Line of argument in treatment of the issue</li> </ul>	<ul style="list-style-type: none"> <li>Argumentative text</li> </ul>	<ul style="list-style-type: none"> <li>Standard – (Familiar accent)</li> <li>Straightforward</li> <li>Clearly signposted/signalled with explicit markers</li> </ul>
		<ul style="list-style-type: none"> <li>Scan</li> </ul>	<ul style="list-style-type: none"> <li>Desired information</li> </ul>	<ul style="list-style-type: none"> <li>Longer texts</li> <li>Different texts, different parts of a text</li> </ul>	
		<ul style="list-style-type: none"> <li>Understand</li> </ul>	<ul style="list-style-type: none"> <li>Straightforward factual information content</li> <li>General message</li> <li>Main conclusions</li> <li>Specific details</li> </ul>	<ul style="list-style-type: none"> <li>Argumentative text</li> <li>Lectures and talks within own field</li> <li>Large part of many TV programmes: interviews, short lectures, news reports</li> <li>Majority of recorded and broadcast audio material</li> </ul>	

**Table A2. Salient Characteristics: Reception (continued)**

	<b>Setting</b>	<b>Action</b>	<b>What is understood</b>	<b>Source</b>	<b>Restrictions</b>
<b>B1</b>	<ul style="list-style-type: none"> <li>Familiar topics regularly encountered in a school, work or leisure context</li> <li>Topics in his/her field of (personal) interest</li> </ul>	<ul style="list-style-type: none"> <li>Follow, though not necessarily in detail</li> </ul>	<ul style="list-style-type: none"> <li>Significant points</li> </ul>	<ul style="list-style-type: none"> <li>Extended discussion around him/her</li> <li>Many films in which visuals and action carry much of the story line</li> <li>TV programmes: interviews, short lectures, news reports</li> <li>Straightforward newspaper articles</li> </ul>	<ul style="list-style-type: none"> <li>Clear</li> <li>Standard</li> <li>Straightforward</li> <li>Relatively slow</li> </ul>
		<ul style="list-style-type: none"> <li>Understand with satisfactory comprehension</li> </ul>	<ul style="list-style-type: none"> <li>Main points</li> <li>Relevant information</li> </ul>	<ul style="list-style-type: none"> <li>Straightforward factual texts</li> <li>Short narratives</li> <li>Descriptions of events, feelings, wishes</li> <li>Detailed directions</li> <li>Short talks</li> <li>Radio news bulletins and simpler recorded materials</li> <li>Everyday written materials: letters, brochures, short official documents</li> <li>Simple technical information e.g. operating instructions</li> </ul>	
<b>A2+</b>	<ul style="list-style-type: none"> <li>Familiar topics of a concrete type</li> </ul>	<ul style="list-style-type: none"> <li>Identify</li> </ul>	<ul style="list-style-type: none"> <li>Main points</li> </ul>	<ul style="list-style-type: none"> <li>TV news items reporting events, accidents etc. in which visuals support the commentary</li> </ul>	<ul style="list-style-type: none"> <li>Clearly and slowly articulated</li> </ul>
		<ul style="list-style-type: none"> <li>Understand enough to meet needs</li> </ul>		<ul style="list-style-type: none"> <li>Basic types of standard letters, faxes (enquiries, orders, confirmations)</li> <li>Short texts with simpler, high frequency everyday and job-related language</li> <li>Regulations, e.g. safety</li> </ul>	<ul style="list-style-type: none"> <li>Expressed in simple language</li> </ul>
<b>A2</b>	<ul style="list-style-type: none"> <li>Predictable everyday matters</li> <li>Areas of most immediate priority: basic personal, family, shopping, local area, employment</li> </ul>	<ul style="list-style-type: none"> <li>Identify</li> </ul>	<ul style="list-style-type: none"> <li>Specific, predictable information</li> <li>Topic of discussion</li> <li>Changes of topic</li> <li>An idea of the content</li> </ul>	<ul style="list-style-type: none"> <li>Simpler everyday material: advertisements, menus, reference lists, timetables, brochures, letters</li> <li>Discussion around him/her</li> <li>Short newspaper articles describing events</li> <li>Factual TV news items</li> </ul>	<ul style="list-style-type: none"> <li>Clearly and slowly articulated</li> </ul>
		<ul style="list-style-type: none"> <li>Understand</li> </ul>	<ul style="list-style-type: none"> <li>Main point</li> <li>Essential information</li> </ul>	<ul style="list-style-type: none"> <li>Short simple texts containing the highest frequency vocabulary including a proportion of shared international vocabulary items</li> <li>Simple directions relating to how to get from A to B</li> <li>Simple clear messages, announcements, recorded passages</li> <li>Simple instructions on equipment encountered in everyday life (e.g. telephone)</li> <li>Short simple personal letters</li> <li>Everyday signs and notices: directions, instructions, hazards</li> </ul>	
<b>A1</b>	<ul style="list-style-type: none"> <li>The most common everyday situations</li> </ul>	<ul style="list-style-type: none"> <li>Identify</li> </ul>	<ul style="list-style-type: none"> <li>Familiar words, phrases, names</li> <li>An idea of the content</li> </ul>	<ul style="list-style-type: none"> <li>Simple notices</li> <li>Simpler informational material</li> </ul>	<ul style="list-style-type: none"> <li>Very slow, carefully articulated, with long pauses to allow assimilation of meaning</li> <li>Familiar names, words and basic phrases</li> <li>A chance to re-read/get repetition</li> </ul>
		<ul style="list-style-type: none"> <li>Understand</li> </ul>	<ul style="list-style-type: none"> <li><i>(Main point)</i></li> </ul>	<ul style="list-style-type: none"> <li>Very short simple texts with visual support, a single phrase at a time:                             <ul style="list-style-type: none"> <li>messages on postcards</li> <li>directions</li> <li>descriptions</li> </ul> </li> </ul>	

## Section A2: Forms for Describing the Examination (Chapter 4)

GENERAL EXAMINATION DESCRIPTION			
<b>1. General Information</b>			
Name of examination	_____		
Language tested	_____		
Examining institution	_____		
Versions analysed (date)	_____		
Type of examination	<input type="checkbox"/> International <input type="checkbox"/> National <input type="checkbox"/> Regional <input type="checkbox"/> Institutional		
Purpose	_____		
Target population	<input type="checkbox"/> Lower Sec <input type="checkbox"/> Upper Sec <input type="checkbox"/> Uni/College Students <input type="checkbox"/> Adult		
No. of test takers per year	_____		
<b>2. What is the overall aim?</b>			
_____			
<b>3. What are the more specific objectives? If available describe the needs of the intended users on which this examination is based.</b>			
_____			
<b>4. What is/are principal domain(s)?</b>	<input type="checkbox"/> Public <input type="checkbox"/> Personal <input type="checkbox"/> Occupational <input type="checkbox"/> Educational		
<b>5. Which communicative activities are tested?</b>	<input type="checkbox"/> 1 Listening comprehension <input type="checkbox"/> 2 Reading comprehension <input type="checkbox"/> 3 Spoken interaction <input type="checkbox"/> 4 Written interaction <input type="checkbox"/> 5 Spoken production <input type="checkbox"/> 6 Written production <input type="checkbox"/> 7 Integrated skills <input type="checkbox"/> 8 Spoken mediation of text <input type="checkbox"/> 9 Written mediation of text <input type="checkbox"/> 10 Language usage <input type="checkbox"/> 11 Other: (specify): _____	<b>Name of Subtest(s)</b>	<b>Duration</b>
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
		_____	_____
<b>6. What is the weighting of the different subtests in the global result?</b>	_____		

Form A1: General Examination Description (part)



3. If no external organisation was involved, what other factors determined design and development of examination?	<input type="checkbox"/> A needs analysis <input type="checkbox"/> Internal description of examination aims <input type="checkbox"/> Internal description of language level <input type="checkbox"/> A syllabus or curriculum <input type="checkbox"/> Profile of candidates
4. In producing test tasks are specific features of candidates taken into account?	<input type="checkbox"/> Linguistic background (L1) <input type="checkbox"/> Language learning background <input type="checkbox"/> Age <input type="checkbox"/> Educational level <input type="checkbox"/> Socio-economic background <input type="checkbox"/> Social-cultural factors <input type="checkbox"/> Ethnic background <input type="checkbox"/> Gender
5. Who writes the items or develops the test tasks?	
6. Have test writers guidance to ensure quality?	<input type="checkbox"/> Training <input type="checkbox"/> Guidelines <input type="checkbox"/> Checklists <input type="checkbox"/> Examples of valid, reliable, appropriate tasks: <input type="checkbox"/> Calibrated to CEFR level description <input type="checkbox"/> Calibrated to other level description: _____
7. Is training for test writers provided?	<input type="checkbox"/> Yes <input type="checkbox"/> No
8. Are test tasks discussed before use?	<input type="checkbox"/> Yes <input type="checkbox"/> No
9. If yes, by whom?	<input type="checkbox"/> Individual colleagues <input type="checkbox"/> Internal group discussion <input type="checkbox"/> External examination committee <input type="checkbox"/> Internal stakeholders <input type="checkbox"/> External stakeholders
10. Are test tasks pretested?	<input type="checkbox"/> Yes <input type="checkbox"/> No
11. If yes, how?	
12. If no, why not?	
13. Is the reliability of the test estimated?	<input type="checkbox"/> Yes <input type="checkbox"/> No
14. If yes, how?	<input type="checkbox"/> Data collection and psychometric procedures <input type="checkbox"/> Other: specify: _____
15. Are different aspects of validity estimated?	<input type="checkbox"/> Face validity <input type="checkbox"/> Content validity <input type="checkbox"/> Concurrent validity <input type="checkbox"/> Predictive validity <input type="checkbox"/> Construct validity
16. If yes, describe how.	

**Form A2: Test Development (continued)**

<b>Marking: Subtest</b>	<b>Complete a copy of this form for each subtest. Short description and/or reference</b>
1. How are the test tasks marked?	For receptive test tasks: <input type="checkbox"/> Optical mark reader <input type="checkbox"/> Clerical marking For productive or integrated test tasks: <input type="checkbox"/> Trained examiners <input type="checkbox"/> Teachers
2. Where are the test tasks marked?	<input type="checkbox"/> Centrally <input type="checkbox"/> Locally: <input type="checkbox"/> By local teams <input type="checkbox"/> By individual examiners
3. What criteria are used to select markers?	
4. How is accuracy of marking promoted?	<input type="checkbox"/> Regular checks by co-ordinator <input type="checkbox"/> Training of markers/raters <input type="checkbox"/> Moderating sessions to standardise judgments <input type="checkbox"/> Using standardised examples of test tasks: <input type="checkbox"/> Calibrated to CEFR <input type="checkbox"/> Calibrated to another level description <input type="checkbox"/> Not calibrated to CEFR or other description
5. Describe the specifications of the rating criteria of productive and/or integrative test tasks.	<input type="checkbox"/> One holistic score for each task <input type="checkbox"/> Marks for different aspects for each task <input type="checkbox"/> Rating scale for overall performance in test <input type="checkbox"/> Rating Grid for aspects of test performance <input type="checkbox"/> Rating scale for each task <input type="checkbox"/> Rating Grid for aspects of each task <input type="checkbox"/> Rating scale bands are defined, but not to CEFR <input type="checkbox"/> Rating scale bands are defined in relation to CEFR
6. Are productive or integrated test tasks single or double rated?	<input type="checkbox"/> Single rater <input type="checkbox"/> Two simultaneous raters <input type="checkbox"/> Double marking of scripts / recordings <input type="checkbox"/> Other: specify: _____
7. If double rated, what procedures are used when differences between raters occur?	<input type="checkbox"/> Use of third rater and that score holds <input type="checkbox"/> Use of third marker and two closest marks used <input type="checkbox"/> Average of two marks <input type="checkbox"/> Two markers discuss and reach agreement <input type="checkbox"/> Other: specify: _____
8. Is inter-rater agreement calculated?	<input type="checkbox"/> Yes <input type="checkbox"/> No
9. Is intra-rater agreement calculated?	<input type="checkbox"/> Yes <input type="checkbox"/> No

**Form A3: Marking**

<b>Grading: Subtest _____</b>	<b>Complete a copy of this form for each Subtest. Short description and/or reference</b>
1. Are pass marks and/or grades given?	<input type="checkbox"/> Pass marks <input type="checkbox"/> Grades
2. Describe the procedures used to establish pass marks and/or grades and cut scores	
3. If only pass/fail is reported, how are the cut-off scores for pass/fail set?	
4. If grades are given, how are the grade boundaries decided?	
5. How is consistency in these standards maintained?	

*Form A4: Grading*

<b>Results</b>	<b>Short description and/or reference</b>
1. What results are reported to candidates?	<input type="checkbox"/> Global grade or pass/fail <input type="checkbox"/> Grade or pass/fail per subtest <input type="checkbox"/> Global grade plus profile across subtests <input type="checkbox"/> Profile of aspects of performance per subtest
2. In what form are results reported?	<input type="checkbox"/> Raw scores <input type="checkbox"/> Undefined grades (e.g. "C") <input type="checkbox"/> Level on a defined scale <input type="checkbox"/> Diagnostic profiles
3. On what document are results reported?	<input type="checkbox"/> Letter or email <input type="checkbox"/> Report card <input type="checkbox"/> Certificate / Diploma <input type="checkbox"/> On-line
4. Is information provided to help candidates to interpret results? Give details.	
5. Do candidates have the right to see the corrected and scored examination papers?	
6. Do candidates have the right to ask for remarking?	

*Form A5: Reporting Results*

<b>Data analysis</b>	<b>Short description and/or reference</b>
1. Is feedback gathered on the examinations?	<input type="checkbox"/> Yes <input type="checkbox"/> No
2. If yes, by whom?	<input type="checkbox"/> Internal experts (colleagues) <input type="checkbox"/> External experts <input type="checkbox"/> Local examination institutes <input type="checkbox"/> Test administrators <input type="checkbox"/> Teachers <input type="checkbox"/> Candidates
3. Is the feedback incorporated in revised versions of the examinations?	<input type="checkbox"/> Yes <input type="checkbox"/> No
4. Is data collected to do analysis on the tests?	<input type="checkbox"/> On all tests <input type="checkbox"/> On a sample of test takers: How large?: _____. How often?: _____ <input type="checkbox"/> No
5. If yes, indicate how data are collected?	<input type="checkbox"/> During pretesting <input type="checkbox"/> During live examinations <input type="checkbox"/> After live examinations
6. For which features is analysis on the data gathered carried out?	<input type="checkbox"/> Difficulty <input type="checkbox"/> Discrimination <input type="checkbox"/> Reliability <input type="checkbox"/> Validity
7. State which analytic methods have been used (e.g. in terms of psychometric procedures).	
8. Are performances of candidates from different groups analysed? If so, describe how.	
9. Describe the procedures to protect the confidentiality of data.	
10. Are relevant measurement concepts explained for test users? If so, describe how.	

**Form A6: Data Analysis**

<b>Rationale for decisions (and revisions)</b>	<b>Short description and/or reference</b>
Give the rationale for the decisions that have been made in relation to the examination or the test tasks in question. Is there a review cycle for the examination? (How often? Who by? Procedures for revising decisions)	

**Form A7: Rationale for Decisions**

Initial Estimation of Overall CEFR Level		
<input type="checkbox"/> A1	<input type="checkbox"/> B1	<input type="checkbox"/> C1
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> A2	<input type="checkbox"/> B2	<input type="checkbox"/> C2
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Short rationale, reference to documentation		

*Form A8: Initial Estimation of Overall Examination Level*

## Section A3: Specification: Communicative Language Activities (Chapter 4)

### A3.1 Reception

#### Listening Comprehension

	Short description and/or reference
1 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
4 What text-types and what length of text are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
5 After reading the scale for Overall Listening Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for listening comprehension in CEFR 4.4.2.1 listed after the scale might be of help as a reference.	<b>Level:</b>
	<b>Justification (incl. reference to documentation)</b>

*Form A9: Listening Comprehension*

<b>OVERALL LISTENING COMPREHENSION</b>	
<b>C2</b>	<i>Has no difficulty in understanding any kind of spoken language, whether live or broadcast, delivered at fast native speed.</i>
<b>C1</b>	<i>Can understand enough to follow extended speech on abstract and complex topics beyond his/her own field, though he/she may need to confirm occasional details, especially if the accent is unfamiliar. Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts. Can follow extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly.</i>
<b>B2</b>	<i>Can understand standard spoken language, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life. Only extreme background noise, inadequate discourse structure and/or idiomatic usage influences the ability to understand.</i>
	<i>Can understand the main ideas of propositionally and linguistically complex speech on both concrete and abstract topics delivered in a standard dialect, including technical discussions in his/her field of specialisation. Can follow extended speech and complex lines of argument provided the topic is reasonably familiar, and the direction of the talk is sign-posted by explicit markers.</i>
<b>B1</b>	<i>Can understand straightforward factual information about common everyday or job related topics, identifying both general messages and specific details, provided speech is clearly articulated in a generally familiar accent.</i>
	<i>Can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives.</i>
<b>A2</b>	<i>Can understand enough to be able to meet needs of a concrete type provided speech is clearly and slowly articulated.</i>
	<i>Can understand phrases and expressions related to areas of most immediate priority (e.g. very basic personal and family information, shopping, local geography, employment) provided speech is clearly and slowly articulated.</i>
<b>A1</b>	<i>Can follow speech which is very slow and carefully articulated, with long pauses for him/her to assimilate meaning.</i>

<b>Relevant Subscales for Listening Comprehension</b>	<b>English</b>
➤ Understanding conversation between native speakers	Page 66
➤ Listening as a member of an audience	Page 67
➤ Listening to announcements and instructions	Page 67
➤ Listening to audio media and recordings	Page 68
➤ Watching TV and film	Page 71
➤ Identifying cues and inferring	Page 72
➤ Notetaking	Page 96

### Reading Comprehension

	<b>Short description and/or reference</b>
1 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
4 What text-types and what length of text are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	

### *Form A10: Reading Comprehension (part)*

<p>5 After reading the scale for Overall Reading Comprehension, given below, indicate and justify at which level(s) of the scale the subtest should be situated.</p> <p>The subscales for reading comprehension in CEFR 4.4.2.2 listed after the scale might be of help as a reference.</p>	<p><b>Level</b></p> <p><b>Justification (incl. reference to documentation)</b></p>
---	--

**Form A10: Reading Comprehension (continued)**

<b>OVERALL READING COMPREHENSION</b>	
<b>C2</b>	<i>Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.</i>
<b>C1</b>	<i>Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.</i>
<b>B2</b>	<i>Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low-frequency idioms.</i>
<b>B1</b>	<i>Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.</i>
<b>A2</b>	<i>Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.</i>
<b>A1</b>	<i>Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.</i>

<b>Relevant Subscales for Reading Comprehension</b>	<b>English</b>
➤ Reading correspondence	Page 69
➤ Reading for orientation	Page 70
➤ Reading for information and argument	Page 70
➤ Reading instructions	Page 71
➤ Identifying cues and inferring	Page 72
➤ Notetaking	Page 96

**A3.2 Interaction**

<b>Spoken Interaction</b>	<b>Short description and/or reference</b>
1 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	

**Form A11: Spoken Interaction (part)**

4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
5 After reading the scale for Overall Spoken Interaction, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for spoken interaction in CEFR 4.4.3.1 listed after the scale might be of help as a reference.	<b>Level</b> <b>Justification (incl. reference to documentation)</b>

**Form A11: Spoken Interaction (continued)**

<b>OVERALL SPOKEN INTERACTION</b>	
<b>C2</b>	<i>Has a good command of idiomatic expressions and colloquialisms with awareness of connotative levels of meaning. Can convey finer shades of meaning precisely by using, with reasonable accuracy, a wide range of modification devices. Can backtrack and restructure around a difficulty so smoothly the interlocutor is hardly aware of it.</i>
<b>C1</b>	<i>Can express him/herself fluently and spontaneously, almost effortlessly. Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions. There is little obvious searching for expressions or avoidance strategies; only a conceptually difficult subject can hinder a natural, smooth flow of language.</i>
<b>B2</b>	<i>Can use the language fluently, accurately and effectively on a wide range of general, academic, vocational or leisure topics, marking clearly the relationships between ideas. Can communicate spontaneously with good grammatical control without much sign of having to restrict what he/she wants to say, adopting a level of formality appropriate to the circumstances.</i>
	<i>Can interact with a degree of fluency and spontaneity that makes regular interaction, and sustained relationships with native speakers quite possible without imposing strain on either party. Can highlight the personal significance of events and experiences, account for and sustain views clearly by providing relevant explanations and arguments.</i>
<b>B1</b>	<i>Can communicate with some confidence on familiar routine and non-routine matters related to his/her interests and professional field. Can exchange, check and confirm information, deal with less routine situations and explain why something is a problem. Can express thoughts on more abstract, cultural topics such as films, books, music etc.</i>
	<i>Can exploit a wide range of simple language to deal with most situations likely to arise whilst travelling. Can enter unprepared into conversation of familiar topics, express personal opinions and exchange information on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events).</i>
<b>A2</b>	<i>Can interact with reasonable ease in structured situations and short conversations, provided the other person helps if necessary. Can manage simple, routine exchanges without undue effort; can ask and answer questions and exchange ideas and information on familiar topics in predictable everyday situations.</i>
	<i>Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters to do with work and free time. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of his/her own accord.</i>
<b>A1</b>	<i>Can interact in a simple way but communication is totally dependent on repetition at a slower rate of speech, rephrasing and repair. Can ask and answer simple questions, initiate and respond to simple statements in areas of immediate need or on very familiar topics.</i>

<b>Relevant Subscales for Spoken Interaction</b>	<b>English</b>
➤ Understanding a native-speaker interlocutor	Page 75
➤ Conversation	Page 76
➤ Informal discussion	Page 77
➤ Formal discussion and meetings	Page 78
➤ Goal-oriented cooperation	Page 79

➤ Transactions to obtain goods and services	Page 80
➤ Information exchange	Page 81
➤ Interviewing and being interviewed	Page 82

Written Interaction	Short description and/or reference
1 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
5 After reading the scale for Overall Written Interaction, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for written interaction in CEFR 4.4.3.4 listed after the scale might be of help as a reference.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

**Form A12: Written Interaction**

OVERALL WRITTEN INTERACTION	
<b>C2</b>	<i>As C1</i>
<b>C1</b>	<i>Can express him/herself with clarity and precision, relating to the addressee flexibly and effectively.</i>
<b>B2</b>	<i>Can express news and views effectively in writing, and relate to those of others.</i>
<b>B1</b>	<i>Can convey information and ideas on abstract as well as concrete topics, check information and ask about or explain problems with reasonable precision.</i>
	<i>Can write personal letters and notes asking for or conveying simple information of immediate relevance, getting across the point he/she feels to be important.</i>
<b>A2</b>	<i>Can write short, simple formulaic notes relating to matters in areas of immediate need.</i>
<b>A1</b>	<i>Can ask for or pass on personal details in written form.</i>

Relevant Subscales for Written Interaction	English
➤ Correspondence	Page 83
➤ Notes, messages and forms	Page 84

### A3.3 Production

Spoken Production	Short description and/or reference
1 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
5 After reading the scale for Overall Spoken Production, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for spoken production in CEFR 4.4.1.1 listed after the scale might be of help as a reference.	<b>Level</b> <b>Justification (incl. reference to documentation)</b>

#### *Form A13: Spoken Production*

OVERALLSPOKEN PRODUCTION	
<b>C2</b>	<i>Can produce clear, smoothly flowing well-structured speech with an effective logical structure which helps the recipient to notice and remember significant points.</i>
<b>C1</b>	<i>Can give clear, detailed descriptions and presentations on complex subjects, integrating sub themes, developing particular points and rounding off with an appropriate conclusion.</i>
<b>B2</b>	<i>Can give clear, systematically developed descriptions and presentations, with appropriate highlighting of significant points, and relevant supporting detail.</i>
	<i>Can give clear, detailed descriptions and presentations on a wide range of subjects related to his/her field of interest, expanding and supporting ideas with subsidiary points and relevant examples.</i>
<b>B1</b>	<i>Can reasonably fluently sustain a straightforward description of one of a variety of subjects within his/her field of interest, presenting it as a linear sequence of points.</i>
<b>A2</b>	<i>Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes etc. as a short series of simple phrases and sentences linked into a list.</i>
<b>A1</b>	<i>Can produce simple mainly isolated phrases about people and places.</i>

Relevant Subscales for Spoken Production	English
➤ Sustained monologue: Describing experience	Page 59
➤ Sustained monologue: Putting a case (e.g. in debate)	Page 59
➤ Public announcements	Page 60
➤ Addressing audiences	Page 60

<b>Written Production</b>	<b>Short description and/or reference</b>
1 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
2 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
3 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
4 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
5 After reading the scale for Overall Written Production, given below, indicate and justify at which level(s) of the scale the subtest should be situated. The subscales for written production in CEFR 4.4.1.2 listed after the scale might be of help as a reference.	<b>Level</b> <b>Justification (incl. reference to documentation)</b>

**Form A14: Written Production**

<b>OVERALL WRITTEN PRODUCTION</b>	
<b>C2</b>	<i>Can write clear, smoothly flowing, complex texts in an appropriate and effective style and a logical structure which helps the reader to find significant points.</i>
<b>C1</b>	<i>Can write clear, well-structured texts of complex subjects, underlining the relevant salient issues, expanding and supporting points of view at some length with subsidiary points, reasons and relevant examples, and rounding off with an appropriate conclusion.</i>
<b>B2</b>	<i>Can write clear, detailed texts on a variety of subjects related to his/her field of interest, synthesising and evaluating information and arguments from a number of sources.</i>
<b>B1</b>	<i>Can write straightforward connected texts on a range of familiar subjects within his/her field of interest, by linking a series of shorter discrete elements into a linear sequence.</i>
<b>A2</b>	<i>Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because".</i>
<b>A1</b>	<i>Can write simple isolated phrases and sentences.</i>

<b>Relevant Subscales for Written Production</b>	<b>English</b>
➤ Creative writing	Page 62
➤ Reports and essays	Page 62

### A3.4 Integrated Skills

What combinations of skills occur in the examination subtests?

Indicate in Form A15 and then complete a copy of Form A16 for each combination

<b>Integrated Skills Combinations</b>		<b>Subtest it occurs in</b>
1 Listening and Note-taking	<input type="checkbox"/>	
2 Listening and Spoken Production	<input type="checkbox"/>	
3 Listening and Written Production	<input type="checkbox"/>	
4 Reading and Note-taking	<input type="checkbox"/>	
5 Reading and Spoken Production	<input type="checkbox"/>	
6 Reading and Written Production	<input type="checkbox"/>	
7 Listening and Reading, plus Note-taking	<input type="checkbox"/>	
8 Listening and Reading, plus Spoken Production	<input type="checkbox"/>	
9 Listening and Reading, plus Written Production	<input type="checkbox"/>	

*Form A15: Integrated Skills Combinations*

	<b>Complete for each combination</b>
<b>Integrated Skills</b>	<b>Short description and/or reference</b>
1 Which skills combinations occur? Refer to your entry in Form A15.	
2 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	
3 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
4 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
7 After reading the scales for Processing Text, given below, plus Comprehension and Written Production given earlier, indicate and justify at which level(s) of the scale the subtest should be situated. The subscale for Note-taking in CEFR 4.6.3 might also be of help as a reference.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

*Form A16: Integrated Skills*

<b>PROCESSING TEXT</b>	
<b>C2</b>	<i>Can summarise information from different sources, reconstructing arguments and accounts in a coherent presentation of the overall result.</i>
<b>C1</b>	<i>Can summarise long, demanding texts.</i>
<b>B2</b>	<i>Can summarise a wide range of factual and imaginative texts, commenting on and discussing contrasting points of view and the main themes. Can summarise extracts from news items, interviews or documentaries containing opinions, argument and discussion. Can summarise the plot and sequence of events in a film or play.</i>
<b>B1</b>	<i>Can collate short pieces of information from several sources and summarise them for somebody else. Can paraphrase short written passages in a simple fashion, using the original text wording and ordering.</i>
<b>A2</b>	<i>Can pick out and reproduce key words and phrases or short sentences from a short text within the learner's limited competence and experience. Can copy out short texts in printed or clearly handwritten format.</i>
<b>A1</b>	<i>Can copy out single words and short texts presented in standard printed format.</i>

### **A3.5 Mediation**

<b>Spoken Mediation</b>	<b>Short description and/or reference</b>
1 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	
2 Which type of mediating activities are tested? The list in CEFR 4.4.4.1 might be of help as a reference.	
3 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
4 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
7 There is no scale for Translation in the CEFR. Generalising from the scales for Listening Comprehension, Processing Text and Spoken Production, indicate and justify at which level(s) the subtest should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

<b>Written Mediation</b>	<b>Short description and/or reference</b>
1 Which text-to-text activities occur? Table 6 in CEFR 4.6.4 might be of help as a reference.	
2 Which type of mediating activities are tested? The list in CEFR 4.4.4.2 might be of help as a reference	
3 In what contexts (domains, situations, ...) are the test takers to show ability? Table 5 in CEFR 4.1 might be of help as a reference.	
4 Which communication themes are the test takers expected to be able to handle? The lists in CEFR 4.2 might be of help as a reference.	
5 Which communicative tasks, activities and strategies are the test takers expected to be able to handle? The lists in CEFR 4.3, 4.4.2.1, 7.1, 7.2 and 7.3 might be of help as a reference.	
6 What kind of texts and text-types are the test takers expected to be able to handle? The lists in CEFR 4.6.2 and 4.6.3 might be of help as a reference.	
7 There is no scale for Translation in the CEFR. Generalising from the scales for Reading Comprehension, Processing Text and Written Production, indicate and justify at which level(s) the subtest should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

*Form A18: Written Mediation*

## Section A4: Specification: Communicative Language Competence (Chapter 4)

Forms concerning competence are again provided in the following order:

1. Reception
2. Interaction
3. Production
4. Mediation

### A4.1 Reception

Those CEFR scales most relevant to Receptive skills have been used to create Table A3, which can be referred to in this section. Table A3 does not include any descriptors for “plus levels”. The original scales consulted, some of which do define plus levels, include:

#### Linguistic Competence

- General Linguistic Range English: page 110
- Vocabulary Range English: page 112

#### Socio-linguistic Competence

- Socio-linguistic Appropriateness English: page 122

#### Pragmatic Competence

- Thematic Development English: page 125
- Cohesion and Coherence English: page 125
- Propositional Precision English: page 129

#### Strategic Competence

- Identifying Cues/Inferring English: page 72

<b>Linguistic Competence</b>	<b>Short description and/or reference</b>
1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	
2 After reading the scale for Linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>
<b>Socio-linguistic Competence</b>	<b>Short description and/or reference</b>
3 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	
4 After reading the scale for Socio-linguistic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

**TABLE A3: RELEVANT QUALITATIVE FACTORS FOR RECEPTION**

	<b>LINGUISTIC</b> Edited from General Linguistic Range; <b>Vocabulary Range</b>	<b>SOCIO-LINGUISTIC</b> Edited from Socio-linguistic Appropriateness	<b>PRAGMATIC</b> Edited from Thematic Development and Propositional Precision	<b>STRATEGIC</b> <b>Identifying Cues and Inferring</b>
<b>C2</b>	<i>Can understand a very wide range of language precisely, appreciating emphasis and, differentiation. No signs of comprehension problems. Has a good command of a very broad lexical repertoire including idiomatic expressions and colloquialisms; shows awareness of connotative levels of meaning.</i>	<i>Has a good command of idiomatic expressions and colloquialisms with awareness of connotative levels of meaning. Appreciates fully the socio-linguistic and sociocultural implications of language used by native speakers and can react accordingly.</i>	<i>Can understand precisely finer shades of meaning conveyed by a wide range of qualifying devices (e.g. adverbs expressing degree, clauses expressing limitations). Can understand emphasis and differentiation without ambiguity.</i>	<i>As C1.</i>
<b>C1</b>	<i>Has a good command of a broad lexical repertoire. Good command of idiomatic expressions and colloquialisms.</i>	<i>Can recognise a wide range of idiomatic expressions and colloquialisms, appreciating register shifts; may, however, need to confirm occasional details, especially if the accent is unfamiliar. Can follow films employing a considerable degree of slang and idiomatic usage. Can understand language effectively for social purposes, including emotional, allusive and joking usage.</i>	<i>Can understand elaborate descriptions and narratives, recognising sub-themes, and points of emphasis. Can understand precisely the qualifications in opinions and statements that relate to degrees of, for example, certainty/uncertainty, belief/doubt, likelihood etc.</i>	<i>Is skilled at using contextual, grammatical and lexical cues to infer attitude, mood and intentions and anticipate what will come next.</i>
<b>B2</b>	<i>Has a sufficient range of language to be able to understand descriptions, viewpoints and arguments on most topics pertinent to his everyday life such as family, hobbies and interests, work, travel, and current events.</i>	<i>Can with some effort keep up with fast and colloquial discussions.</i>	<i>Can understand description or narrative, identifying main points from relevant supporting detail and examples. Can understand detailed information reliably.</i>	<i>Can use a variety of strategies to achieve comprehension, including listening for main points; checking comprehension by using contextual clues.</i>
<b>B1</b>	<i>Has enough language to get by, with sufficient vocabulary to understand most texts on topics such as family, hobbies and interests, work, travel, and current events.</i>	<i>Can respond to a wide range of language functions, using their most common exponents in a neutral register. Can recognise salient politeness conventions. Is aware of, and looks out for signs of, the most significant differences between the customs, usages, attitudes, values and beliefs prevalent in the community concerned and those of his or her own.</i>	<i>Can reasonably accurately understand a straightforward narrative or description that is a linear sequence of points. Can understand the main points in an idea or problem with reasonable precision.</i>	<i>Can identify unfamiliar words from the context on topics related to his/her field and interests. Can extrapolate the meaning of occasional unknown words from the context and deduce sentence meaning provided the topic discussed is familiar.</i>
<b>A2</b>	<i>Has a sufficient vocabulary for coping with everyday situations with predictable content and simple survival needs.</i>	<i>Can handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, apologies etc.</i>	<i>Can understand a simple story or description that is a list of points. Can understand a simple and direct exchange of limited information on familiar and routine matters.</i>	<i>Can use an idea of the overall meaning of short texts and utterances on everyday topics of a concrete type to derive the probable meaning of unknown words from the context.</i>
<b>A1</b>	<i>Has a very basic range of simple expressions about personal details and needs of a concrete type.</i>	<i>Can understand the simplest everyday polite forms of: greetings and farewells; introductions; saying please, thank you, sorry etc.</i>	<i>No descriptor available.</i>	<i>No descriptor available.</i>

<b>Pragmatic Competence</b>	<b>Short description and/or reference</b>
5 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.	
6 After reading the scale for Pragmatic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>
<b>Strategic Competence</b>	<b>Short description and/or reference</b>
7 What are the strategic competences that the test takers are expected to be able to handle? The discussion in CEFR 4.4.2.4. might be of help as a reference.	
8 After reading the scale for Strategic Competence in Table A3, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

*Form A19: Aspects of Language Competence in Reception (continued)*

## **A4.2 Interaction**

Those CEFR scales most relevant to Interaction have been used to create Table A4 which can be referred to in this section. Table A4 does not include any descriptors for “plus levels”. The original scales consulted, some of which do define plus levels, include:

### **Linguistic Competence**

- General Linguistic Range                      English: page 110
- Vocabulary Range                                English: page 112
- Vocabulary Control                              English: page 112
- Grammatical Accuracy                         English: page 114

### **Socio-linguistic Competence**

- Socio-linguistic Appropriateness          English: page 122

### **Pragmatic Competence**

- Flexibility                                         English: page 124
- Turntaking                                         English: page 124
- Spoken Fluency                                 English: page 129
- Propositional Precision                        English: page 129

### **Strategic Competence**

- Turntaking (repeated)                         English: page 86
- Cooperating                                      English: page 86
- Asking for Clarification                       English: page 87
- Compensating                                  English: page 64
- Monitoring and Repair                         English: page 65

<b>Linguistic Competence</b>	<b>Short description and/or reference</b>
------------------------------	---

1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	
2 What is the range of phonological and orthographic competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.4 and 5.2.1.5 might be of help as a reference.	
3 After reading the scales for Range and Accuracy in Table A4, indicate and justify at which level(s) of the scale the examination should be situated. The scales for Phonological Control in CEFR 5.2.1.4 and for Orthographic Control in 5.2.1.5 might also be of help as a reference.	<b>Level</b> <b>Justification (incl. reference to documentation)</b>
<b>Socio-linguistic Competence</b>	<b>Short description and/or reference</b>
4 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	
5 After reading the scale for Socio-linguistic Competence in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b> <b>Justification (incl. reference to documentation)</b>
<b>Pragmatic Competence</b>	<b>Short description and/or reference</b>
6 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.	
7 After reading the scale for Fluency in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b> <b>Justification (incl. reference to documentation)</b>

*Form A20: Aspects of Language Competence in Interaction (part)*

<b>Strategic Competence</b>	<b>Short description and/or reference</b>
-----------------------------	---

<p>8 What are the interaction strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.3.5 might be of help as a reference.</p>	
<p>9 After reading the scale for Interaction in Table A4, indicate and justify at which level(s) of the scale the examination should be situated.</p>	<p><b>Level</b></p> <hr/> <p><b>Justification (incl. reference to documentation)</b></p>

*Form A20: Aspects of Language Competence in Interaction (continued)*

### A4.3 Production

Those CEFR scales most relevant to Production have been used to create Table A5, which can be referred to in this section. Table A5 does not include any descriptors for “plus levels”. The original scales consulted, some of which do define plus levels, include:

#### Linguistic Competence

- General Linguistic Range English: page 110
- Vocabulary Range English: page 112
- Vocabulary Control English: page 112
- Grammatical Accuracy English: page 114

#### Socio-linguistic Competence

- Socio-linguistic Appropriateness English: page 122

#### Pragmatic Competence

- Flexibility English: page 124
- Thematic Development English: page 125
- Cohesion and Coherence English: page 125
- Spoken Fluency English: page 129
- Propositional Precision English: page 129

#### Strategic Competence

- Planning English: page 64
- Compensating English: page 64
- Monitoring and Repair English: page 65

<b>Linguistic Competence</b>	<b>Short description and/or reference</b>
<p>1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.</p>	
<p>2 What is the range of phonological and orthographic competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.4 and 5.2.1.5 might be of help as a reference.</p>	

*Form A21: Aspects of Language Competence in Production (part)*

<p>3 After reading the scales for Range and Accuracy in Table A5 indicate and justify at which level(s) of the scale the examination should be situated. The scales for Phonological Control in CEFR 5.2.1.4 and for Orthographic Control in 5.2.1.5 might also be of help as a reference.</p>	<p><b>Level</b></p> <hr/> <p><b>Justification (incl. reference to documentation)</b></p>
<p><b>Socio-linguistic Competence</b></p>	<p><b>Short description and/or reference</b></p>
<p>4 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.</p>	
<p>5 After reading the scale for Socio-linguistic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.</p>	<p><b>Level</b></p> <hr/> <p><b>Justification (incl. reference to documentation)</b></p>
<p><b>Pragmatic Competence</b></p>	<p><b>Short description and/or reference</b></p>
<p>6 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.</p>	
<p>7 After reading the scale for Pragmatic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.</p>	<p><b>Level</b></p> <hr/> <p><b>Justification (incl. reference to documentation)</b></p>
<p><b>Strategic Competence</b></p>	<p><b>Short description and/or reference</b></p>
<p>8 What are the production strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.1.3 might be of help as a reference.</p>	
<p>9 After reading the scale for Strategic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.</p>	<p><b>Level</b></p> <hr/> <p><b>Justification (incl. reference to documentation)</b></p>

*Form A21: Aspects of Language Competence in Production (continued)*

**TABLE A4: RELEVANT QUALITATIVE FACTORS FOR SPOKEN INTERACTION**

	<b>LINGUISTIC RANGE</b> Edited from General Linguistic Range; Vocabulary Range, Flexibility	<b>LINGUISTIC ACCURACY</b> Edited from Grammatical Accuracy and Vocabulary Control	<b>SOCIO-LINGUISTIC</b> Edited from Socio-linguistic Appropriateness	<b>FLUENCY</b> Fluency, Flexibility	<b>INTERACTION</b> Edited from Turntaking, Cooperating, Asking for Clarification
<b>C2</b>	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).	Appreciates fully the socio-linguistic and sociocultural implications of language used by speakers and can react accordingly. Can mediate effectively between speakers of the target language and that of his/her community of origin taking account of sociocultural and socio-linguistic differences.	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 turntaking, referencing, allusion making etc.
<b>C1</b>	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 use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.	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.
<b>B2</b>	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 with some effort keep up with and contribute to group discussions even when speech is fast and colloquial. Can sustain relationships with native speakers without unintentionally amusing or irritating them or requiring them to behave other than they would with a native speaker.	Can adjust to the changes of direction, style and emphasis normally found in conversation. Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or 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.
<b>B1</b>	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 perform and respond to basic language functions, such as information exchange and requests and express opinions and attitudes in a simple way. Is aware of the salient politeness conventions and acts appropriately.	Can exploit a wide range of simple language flexibly to express much of what he/she wants. 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.
<b>A2</b>	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 handle very short social exchanges, using everyday polite forms of greeting and address. Can make and respond to invitations, apologies etc.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can expand learned phrases through simple recombinations of their elements.	Can indicate when he/she is following but is rarely able to understand enough to keep conversation going of his/her own accord. Can ask for attention.
<b>A1</b>	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited grammatical control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can establish basic social contact by using the simplest everyday polite forms of: greetings and farewells; introductions; saying please, thank you, sorry etc.	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 interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.

**TABLE A5: RELEVANT QUALITATIVE FACTORS FOR PRODUCTION**

	<b>LINGUISTIC RANGE</b> General Linguistic Range; Vocabulary Range	<b>LINGUISTIC ACCURACY</b> Grammatical Accuracy, Vocabulary Control, Phonological Control	<b>SOCIO- LINGUISTIC</b> Socio-linguistic Appropriateness	<b>PRAGMATIC</b> Fluency, Flexibility	<b>PRAGMATIC</b> Thematic Development, Propositional Precision, Coherence and Cohesion	<b>STRATEGIC</b> Compensating, Monitoring and Repair
<b>C2</b>	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).	Appreciates fully the socio-linguistic and sociocultural implications of language used by speakers and can react accordingly.	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 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.	Can substitute an equivalent term for a word he/she can't recall so smoothly that it is scarcely noticeable.
<b>C1</b>	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 use language flexibly and effectively for social purposes, including emotional, allusive and joking usage.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices. Can give elaborate descriptions and narratives, integrating sub themes, developing particular points and rounding off with an appropriate conclusion.	Can backtrack when he/she encounters a difficulty and reformulate what he/she wants to say without fully interrupting the flow of speech.
<b>B2</b>	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 express him or herself appropriately in situations and avoid crass errors of formulation.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can develop a clear description or narrative, expanding and supporting his/her main points with relevant supporting detail and examples. 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.	Can use circumlocution and paraphrase to cover gaps in vocabulary and structure. Can make a note of "favourite mistakes" and consciously monitor speech for it/them.
<b>B1</b>	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.	<i>No descriptor available</i>	Can exploit a wide range of simple language flexibly to express much of what he/she wants. 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 link a series of shorter, discrete simple elements in order to reasonably fluently relate a straightforward narrative or description as a linear sequence of points.	Can use a simple word meaning something similar to the concept he/she wants to convey and invites "correction". Can start again using a different tactic when communication breaks down.
<b>A2</b>	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.	<i>No descriptor available</i>	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can expand learned phrases through simple recombinations of their elements.	Can link groups of words with simple connectors like "and", "but" and "because".	<i>No descriptor available</i>
<b>A1</b>	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.	<i>No descriptor available</i>	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 link words or groups of words with very basic linear connectors like "and" or "then".	<i>No descriptor available</i>

## A4.4 Mediation

The question of which CEFR scales are most relevant to Mediation depends upon the type of mediation involved.

In a foreign language context, one will naturally focus on the foreign language skill. Thus the language competences required in mediating from the foreign language to mother tongue will be primarily those required for reception, whilst for mediating from the mother tongue to the foreign language those for production will be necessary. For Mediation entirely in the foreign language, aspects of competence for both reception and production will be required.

<b>Language Variables:</b>	<b>Type of Language Competences:</b>	<b>Descriptors:</b>
a. within a foreign language	For Reception and Production	Tables A3 and A5
b. from one foreign language to another	For Reception and Production	Tables A3 and A5
c. from foreign language to mother tongue	For Reception	Table A3
d. from mother tongue to foreign language	For Production	Table A5

Other factors to consider are skill variables (spoken or written reception to spoken or written production) and task variables – with formal or informal register – as outlined in CEFR 4.4.4.1 (oral mediation) and 4.4.4.2 (written mediation).

Thus, although there are no descriptors for Mediation as such in the CEFR, all the descriptor scales in CEFR Chapter 5, plus the scales for Receptive and Productive Strategies (included in Tables A3 and A5 respectively) are relevant. If the examination includes Mediation, please consult Tables A3, A4, and/or A5 as appropriate in completing Form A22.

<b>Linguistic Competence</b>	<b>Short description and/or reference</b>
1 What is the range of lexical and grammatical competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.1 and 5.2.1.2 might be of help as a reference.	
2 What kind of semantic relationships are the test takers expected to be able to handle? The list in CEFR 5.2.1.3 might be of help as a reference.	
3 What is the range of phonological or orthographic competence that the test takers are expected to be able to handle? The lists in CEFR 5.2.1.4 and 5.2.1.5 might be of help as a reference.	

*Form A22: Aspects of Language Competence in Mediation (part)*

4 The scale for Orthographic Control in CEFR 5.2.1.5 might also be of help as a reference.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>
<b>Socio-linguistic Competence</b>	<b>Short description and/or reference</b>
5 What are the socio-linguistic competences that the test takers are expected to be able to handle: linguistic markers, politeness conventions, register, adequacy, dialect/accent, etc.? The lists in CEFR 5.2.2 might be of help as a reference.	
6 After reading the scale for Socio-linguistic Competence in Table A3 and A4, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>
<b>Pragmatic Competence</b>	<b>Short description and/or reference</b>
7 What are the pragmatic competences that the test takers are expected to be able to handle: discourse competences, functional competences? The lists in CEFR 5.2.3 might be of help as a reference.	
8 After reading the scale for Pragmatic Competence in Table A5, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>
<b>Strategic Competence</b>	<b>Short description and/or reference</b>
9 What are the reception and production strategies that the test takers are expected to be able to handle? The discussion in CEFR 4.4.2.4 and 4.4.1.3 might be of help as a reference.	
10 After reading the scales for Strategic Competence in Tables A3 and A5, indicate and justify at which level(s) of the scale the examination should be situated.	<b>Level</b>
	<b>Justification (incl. reference to documentation)</b>

*Form A22: Aspects of Language Competence in Mediation (continued)*

**Section A5: Specification: Outcome of the Analysis (Chapter 4)**

Form A23 provides a graphic profile of the coverage of the examination in relation to CEFR categories and levels. It should be completed at the end of the Specification process.

<b>C2</b>								
<b>C1</b>								
<b>B2.2</b>								
<b>B2</b>								
<b>B1.2</b>								
<b>B1</b>								
<b>A2.2</b>								
<b>A2</b>								
<b>A1</b>								
<b>Overall</b>	<b>Activity 1</b>	<b>Activity 2</b>	<b>Activity 3</b>	<b>Activity 4</b>	<b>Activity 5</b>	<b>Socio-linguistic</b>	<b>Pragmatic</b>	<b>Linguistic</b>

*Form A23: Graphic Profile of the Relationship of the Examination to CEFR Levels*

<b>Confirmed Estimation of Overall CEFR Level</b>		
<input type="checkbox"/> <b>A1</b>	<input type="checkbox"/> <b>B1</b>	<input type="checkbox"/> <b>C1</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/> <b>A2</b>	<input type="checkbox"/> <b>B2</b>	<input type="checkbox"/> <b>C2</b>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<p><b>Short rationale, reference to documentation. If this form presents a different conclusion to the initial estimation in Form A8, please comment on the principal reasons for the revised view.</b></p>		

*Form A24: Confirmed Estimation of Overall Examination Level*

## Appendix B

### Content Analysis Grids Chapter 4

#### Section B1: CEFR Content Analysis Grid for Listening & Reading

The *CEFR Content Analysis Grid for Listening & Reading*<sup>48</sup> allows test developers to analyse tests of reading and listening in order to relate them to the CEFR. Information about each task, text and item in the test is entered into the Grid by specifying their characteristics (e.g. text source, discourse type, estimated difficulty level, etc.) from a range of options derived from the CEFR. The analyst must, however, be fully familiar with the CEFR in order to use the Grid effectively. For further guidance the system also includes a familiarisation component. The Grid is designed to be used on the web but a paper version is given here. New categories could be added if necessary.

While the Grid was developed to analyse tests of reading and listening, it can also be used as a tool in planning such tests.

A link to the on-line version of the Grid is also available on [www.coe.int/portfolio](http://www.coe.int/portfolio) The direct link is [www.lancs.ac.uk/fss/projects/grid](http://www.lancs.ac.uk/fss/projects/grid)

In this section, the same form has been presented in three versions:

1. A blank version;
2. A version which has been filled in after the panel has analysed the tests, resulting in provisional cut-off scores;
3. A third version in which provisional item classifications have been revised on the basis of confronting pre-estimates of difficulty with empirical information on their difficulty, and similar adjustments have been made to cut-offs.

---

<sup>48</sup> The Grid was produced by a working group consisting of J. Charles Alderson (Project Coordinator) Neus Figueras, Henk Kuijpers, Günther Nold, Sauli Takala and Claire Tardieu. With further funding from the Dutch Ministry of Education the group developed a computerised version which is available at [www.lancs.ac.uk/fss/projects/grid](http://www.lancs.ac.uk/fss/projects/grid) A report on the project is available on request from the Project Coordinator at [c.alderson@lancaster.ac.uk](mailto:c.alderson@lancaster.ac.uk)

**Blank Form Listening**

<b>Listening/reading Comprehension in ... (language)...</b>					
<b>Target level in the curriculum:</b>					
<b>Item types</b>					
<b>Source</b>					
<b>Length (total 45 mins)</b>					
<b>Authenticity</b>					
<b>Discourse type</b>					
<b>Domain</b>					
<b>Topic</b>					
<b>Curriculum linkage (an optional new category)</b>					
<b>Number of speakers</b>					
<b>Pronunciation</b>					
<b>Content</b>					
<b>Grammar</b>					
<b>Vocabulary</b>					
<b>Nr of listening</b>					
<b>Input text comprehensible at level</b>					
<b>Items comprehensible at level (enter item codes)</b>					
<b>A1</b>					
<b>A1/A2</b>					
<b>A2</b>					
<b>A2/B1</b>					
<b>B1</b>					
<b>B1/B2</b>					
<b>B2</b>					
<b>B2/C1</b>					
<b>C1</b>					
<b>C1/C2</b>					
<b>C2</b>					

### Sample Specification of a Listening Test

<b>Test</b>	Listening Comprehension in English				
<b>Target level in the curriculum: B2.1</b>					
<b>Item types</b>	30 multiple-choice items				5 constructed response items
<b>Source</b>	Interview	Interview	Presentation	Radio programme	News
<b>Length (total 45 mins)</b>	7	12	7	9	10
<b>Authenticity</b>	Modified	Modified	Authentic	Authentic	Modified (cut)
<b>Discourse type</b>	Narrative	Argumentative	Descriptive	Descriptive	Narrative
<b>Domain</b>	Personal	Personal	Public	Public	Public
<b>Topic</b>	Pop culture	Environment	Business/trade	Entertainment	Society
<b>Curriculum linkage</b>	Grade 8	Grade 9	Grade 10	Grade 11	Grade 10
<b>Number of speakers</b>	2	2 + 1	2	1	1
<b>Pronunciation</b>	Standard BrE	Standard AmE	Standard BrE	Standard AmE	Standard BrE
<b>Content</b>	Concrete	Concrete	Somewhat abstract	Somewhat abstract	Somewhat abstract
<b>Grammar</b>	Simple	Somewhat complex	Rather complex	Somewhat complex	Somewhat complex
<b>Vocabulary</b>	Only frequent	Mostly frequent	Rather extensive	Rather extensive	Rather extensive
<b>Nr of listening</b>	2	2	2	1	1
<b>Input text comprehensible at level</b>	B1	B1	B2	B1	B1
<b>Items comprehensible at level (enter ratings using item codes)</b>					
<b>A1</b>					
<b>A1/A2</b>					
<b>A2</b>					
<b>A2/B1</b>					
<b>B1</b>	1, 2, 3, 4, 5			25, 27	
<b>B1/B2</b>		6, 7, 8, 10, 12, 14, 15	17	24, 26	Constructed response: 1, 2
<b>B2</b>		9, 11, 13, 16	18, 19, 20	21, 22, 23	Constructed response: 4
<b>B2/C1</b>				28, 29, 30	Constructed response: 3, 5
<b>C1</b>					
<b>C1/C2</b>					
<b>C2</b>					

Preliminary cut-offs:

< B1: 0; B1: 1–19; B2: 20–30; >B2: 31–35

### Sample Grid to be used after Test Administration

<b>Test</b>	Listening Comprehension in English				
<b>Target level in the curriculum: B2.1</b>					
<b>Item types</b>	30 multiple-choice items				5 open-ended
<b>Source</b>	Interview	Interview	Presentation	Radio programme	News
<b>Length (total 45 mins)</b>	7	12	7	9	10
<b>Authenticity</b>	Modified	Modified	Authentic	Authentic	Modified (cut)
<b>Discourse type</b>	Narrative	Argumentative	Descriptive	Descriptive	Narrative
<b>Domain</b>	Personal	Personal	Public	Public	Public
<b>Topic</b>	Pop culture	Environment	Business/trade	Entertainment	Society
<b>Curriculum linkage</b>	Grade 8	Grade 9	Grade 10	Grade 11	Grade 10
<b>Number of speakers</b>	2	2 + 1	2	1	1
<b>Pronunciation</b>	Standard BrE	Standard AmE	Standard BrE	Standard AmE	Standard BrE
<b>Content</b>	Concrete	Concrete	Somewhat abstract	Somewhat abstract	Somewhat abstract
<b>Grammar</b>	Simple	Somewhat complex	Rather complex	Somewhat complex	Somewhat complex
<b>Vocabulary</b>	Only frequent	Mostly frequent	Rather extensive	Rather extensive	Rather extensive
<b>Nr of listening</b>	2	2	2	1	1
<b>Input text comprehensible at level (enter data after standard setting)</b>					
<b>Items comprehensible at level (enter item codes after standard setting)</b>					
<b>A1</b>					
<b>A1/A2</b>					
<b>A2</b>					
<b>A2/B1</b>					
<b>B1</b>					
<b>B1/B2</b>					
<b>B2</b>					
<b>B2/C1</b>					
<b>C1</b>					
<b>C1/C2</b>					
<b>C2</b>					

Final cut-offs:

### Sample Blank Grid for a Reading test

Characteristic	Text 1	Text 2	Text 3	Text 4	Text 5
Text source					
Authenticity					
Discourse type					
Domain					
Topic					
Nature of content					
Text length					
Vocabulary					
Grammar					
<b>Text likely to be comprehensible by learner/user at CEFR level:</b>					

Items comprehensible to a learner/user at CEFR level (enter item code)					
A1					
A2					
B1					
B2					
C1					
C2					

Preliminary cut-offs:  
(Final cut-offs):



## Section B2: CEFR Content Analysis Grids for Writing and Speaking Tasks

These Grids were designed by the ALTE Manual Special Interest Group with the aim of assisting test providers in their work with the CEFR and the Manual. The ALTE Manual Special Interest Group also endeavours to update the Grids according to the feedback they obtain from users. For this reason, users are advised to download the latest versions from the Language Policy Division pages of the Council of Europe website: [www.coe.int/lang](http://www.coe.int/lang)

The aim in designing the Grids was to provide flexible instruments that could be of use in a variety of contexts and for a number of different uses.

There are two types of Grid:

- **Analysis:** used when panels are asked to make judgments about test tasks, such as in training sessions, sessions aiming to provide illustrative samples and in standard setting exercises.
- **Presentation:** used to present analysis already done, perhaps as exemplars for training and standardisation, to serve as a record, or to be presented externally.

As a single use or method was never intended for these Grids, it is not possible to give comprehensive instructions for their use here. For this reason, two examples of the way in which they have been used are provided.

### Example 1

#### **Grid used:**

CEFR Writing Grid: analysis, version 3.0, 2005

**Reason for use:** Benchmarking of writing performances for a suite of local examinations

#### **Procedure:**

At a benchmarking workshop with 11 raters, the Grid was used for an introductory activity. Raters were asked to complete it for one of the tasks and then reflect on and discuss the relevance that each of the categories in the Grid had for relating a task to a certain level. The aim of this activity was to focus the raters on the relationship between task and performance, and on different aspects of task difficulty.

A modification of the Grid was also proposed to the raters, namely the insertion of a category: “Genre of expected text”, to complement the category on “Genre of input”.

The Grid had been complemented by one column where raters could indicate which of the categories they found decisive for relating a task to a level. Raters were asked to state for categories 16 to 38 whether they found each individual category useful for this purpose. The categories nominated most often were: “Genre of expected text” (10 nominations), “Time permitted or suggested” (9), “Genre of input” (8), “Topic or theme of input” (8), “Number of words expected” (8). Some categories elicited discussions about (a) interpretation of the categories, and (b) their applicability across the levels (e.g. category “Genre of input” was found to be relevant only for the higher levels).

**Points to note, positive:** Benchmarking focuses on the linguistic qualities of a text, rather than on task fulfilment aspects. The Grid allowed some task fulfilment aspects to enter into the discussion of text quality, such as the time allowed for writing.

**Points to note, negative:** Some categories will tend to be interpreted differently by different people (e.g. how controlled is “semi-controlled”?).

**Recommendations:**

It would be worthwhile to use the Grid for a test authors' workshop, as it invites reflection on the level of language which will be elicited by a task, and thus on the characteristics a task should have in order to elicit the performance that is required.

One way to encourage a similar interpretation of the terms in the Grid would be for the organisers to provide illustrative examples, perhaps documented with a completed "Output" version reporting the conclusions reached in such an activity.

**Example 2****Grid used:**

CEFR Speaking Grid: analysis and presentation, version 01, 09/12/05

**Reason for use:** Benchmarking of speaking performances of a local examination

**Procedure:**

During the training stage, 12 experienced raters were shown videos of standardised performances selected previously at a benchmarking conference which had been organised in cooperation with the CoE for the language concerned. Each rater had to classify the tasks performed in the recordings to CEFR levels. Individual rating in which raters were asked to fill in the output Grid was followed by pair discussion and then plenary discussion.

The Grid was used to raise the raters' awareness of task difficulty and to show them what kind of categories may influence difficulty more than others. As the performance given by a candidate is closely connected to the expected response elicited by the task, it was useful to get an idea of the task difficulty before the judgment of performance samples started.

In a second stage, the Grids were used in the same way to classify the local tasks of spoken production and to judge the performances of samples of the local exam.

**Points to note, positive:** This method worked well for the following reasons: judges got a more precise idea about the different facets of task difficulty and the level of the related performance. This was helpful, especially for the judgment of the local tasks.

**Points to note, negative:** One of the difficulties of this method was that it takes some time to explain the 45 categories of the input Grid. Therefore, the Grid was translated to the native language and only a limited choice of categories were selected for use during the meeting. Part 1 on general information was left out, in Part 2 we focused on 15/16 control/guidance, 23 topic. Part 3 on response, however was used in its entirety.

**Recommendations:**

The Grid should be sent out to the judges before the standardisation meeting starts in order to familiarise them with the Grid.

This Grid has been developed by the ALTE Manual Special Interest Group in order to assist test providers in their work with the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* and the *Manual for Relating Language Examinations to the CEFR*, both available from the Language Policy Division of the Council of Europe.

There are two varieties of this Grid: the **Analysis** Grid and the **Presentation** Grid (a simplified version).

The **Analysis** Grid is intended to be used in workshops and benchmarking events.

- If a workshop is intended to analyse test content and specifications, the relevant stage of the Manual is *specification*. (Chapter 4.)
- If the Grid is used for benchmarking new, local samples, the relevant section of the Manual is Section 5.6.

The **Presentation** Grid provides a descriptive record of the analysis of test tasks in a previous benchmarking exercise. If completed Grids are used to document illustrative samples, they can be exploited in *standardisation training* (Chapter 5 of the Manual).

## Sample Test Tasks

Report on analysis of	
Target language of this test	
Target level (CEFR) of this test	
Task number/name	

### General Information – the whole test

1. Total test time
2. Purpose
3. Background to the examination
4. Candidature
5. Structure of the test

### General Information – the writing component

6. Number of tasks in the writing paper
7. Total component time
8. Integration of skills
9. Channel
10. CEFR level of this component
11. The writing component format
12. Specific information – example task
13. Mark distribution
14. Task rating
15. Effective level
16. Sample task:

– sample task here –

<b>i) Task input/prompt</b>		
17	Language of input/prompt	
18	CEFR level of input/prompt	
19	Time permitted or suggested for this task	minutes
20	Control/guidance	
21	Content	
22	Genre	
23	Rhetorical function(s) of input	
24	Imagined audience	
25	Mode of input/prompt	
26	Topic or theme of input	
27	Integration of skills for input	

<b>ii) Response (description of written response elicited by the prompt(s)/input)</b>		
28	Number of words expected	
29	Rhetorical function(s) expected	
30	Text purpose	
31	Register	
32	Domain	
33	Grammatical competence expected	
34	Lexical competence expected	
35	Discoursal competence expected	
36	Authenticity: situational	
37	Authenticity: interactional	
38	Cognitive processing	
39	Content knowledge required	

<b>iii) Rating of task</b>		
40	Known criteria	
41	Task rating method	
42	Assessment criteria	
43	Number and combination of raters	

iv) Feedback to candidates	
44	Quantitative feedback
45	Qualitative feedback

#### 46 Example answer

#### 47 Commentary

#### 48 Score allocated

**Notes:** Numbers below correspond to numbered items in the Grid.

- 2 The purpose of the test may be general proficiency, for a specific purpose. State the purpose if specific (English for Legal Purposes, German for Academic Purposes, etc.).
- 3 The description of test background may contain the reasons for developing the test, a description of the suite of which this test is a part, or other such details.
- 4 Describe the size and demographic profile of the candidature.
- 5 Describe the other components of the test (e.g. the speaking component, the reading component).
- 6 In the case that the number of tasks depends on which options are chosen, specify in the introductory text (point 5).
- 8 Skills, in addition to writing, which are involved in the completion of this task (regardless of whether they are explicitly recognised at the rating stage). Choose from: none, reading, speaking, listening, a combination.
- 9 The method by which the candidate's response is recorded. Choose from handwritten, word processed, either.
- 10 *CEFR*, Ch. 3.
- 11 The description may include information such as the number of subsections, task types in each subsection, time allowed for each subsection.
- 12 You may wish to include a short description of the task here. The description could include the aims of the task, what candidates have been asked to do and would constitute a full completion of the task.
- 13 Describe how marks are distributed in this section of the task and what candidates would need to include to achieve full marks on this task.
- 14 Explain how the task is rated (e.g. clerically, machine marked), what instruments are used and what aspects are considered when deciding the grade.
- 15 Describe the measures taken to ensure Writing tasks are set at the appropriate level. This description may include the process of question paper production and trialling.
- 16 Insert the sample task, including rubric and prompt/input.
- 18 Choose *CEFR* level: A1, A2, B1, B2, C1, C2.
- 19 If not specified, expected time.
- 20 The extent to which the rubric, prompt or input determines the nature and content of the response. Choose from: controlled, semi-controlled or open-ended.
- 21 Whether the content of the response is specified in the rubric. Choose from: specified or not specified.
- 22 Choose from: **letter (business), letter (personal), review, academic essay, composition, report, story, proposal, article, form**, other (specify).
- 23 The functions which might be expected in the response. Choose from: **describing (events), describing (processes), narrating, commentating, expositing, explaining, demonstrating, instructing, arguing, persuading, reporting events, giving opinions, making complaints, suggesting, comparing and contrasting, exemplifying, evaluating, expressing possibility/probability, summarising**, other (specify). *CEFR*, p125–130.

- 24 The imagined audience for the input. Choose from: **friend/acquaintance, teacher, employer, employee, committee, board, business, students, general public** (e.g. with a newspaper article), other (specify).
- 25 Choose from: **oral, written** or **visual**, or a **combination**.
- 26 The topic or theme. Choose from: **personal identification, house and home/environment, daily life, free time/entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, weather**, other (specify). *CEFR*, p51–53.
- 27 The language skills the candidate needs to understand the rubric and prompt/input. Choose from: **reading, listening**, or a **combination**.
- 29 The functions which might be expected in the response. Choose from: **describing (events), describing (processes), narrating, commentating, expositing, explaining, demonstrating, instructing, arguing, persuading, reporting events, giving opinions, making complaints, suggesting, comparing and contrasting, exemplifying, evaluating, expressing possibility/probability, summarising**, other (specify). *CEFR*, p125–130.
- 30 The expected purpose(s) of the response. Choose from: referential (to give “objective” facts about the world), emotive (to describe the emotional state of the writer), conative (to persuade the reader(s)), phatic (to establish or maintain social contact with the reader(s)), metalingual (to clarify or verify understanding), poetic (writing for aesthetic purposes).
- 31 The register the candidate is expected to adopt in their response. Choose from: **informal, unmarked to informal, unmarked, unmarked to formal, formal**. *CEFR*, p118–122.
- 32 The domain to which the expected response is imagined to belong. Choose from: **personal, public, occupational, educational/academic**. *CEFR*, p45–46.
- 33 Choose *CEFR* level: **A1, A2, B1, B2, C1, C2**. *CEFR*, p112–116.
- 34 Choose *CEFR* level: **A1, A2, B1, B2, C1, C2**. *CEFR*, p110–112.
- 35 Choose *CEFR* level: **A1, A2, B1, B2, C1, C2**. *CEFR*, p123–125.
- 36 The extent to which the task reflects a real-life activity a candidate could perform. Choose from **low, medium, or high**.
- 37 The extent to which interaction patterns are likely to mirror those in an equivalent, real-life task. Choose from **low, medium, or high**.
- 38 The difficulty in performing the task from a non-linguistic point-of-view. Choose from: **reproduction of known ideas, knowledge transformation**.
- 39 The kind of extra-linguistic knowledge required to complete the task. Choose from: **personal/everyday life knowledge areas, general/non-specialised knowledge areas, specialised knowledge areas** (scientific, study-related, etc.), a **wide range of knowledge areas**.
- 40 Describe the rating criteria made available to the candidate, either before or during the test. If the criteria are not available together with the paper, state where they can be viewed.
- 41 Choose from: **impressionistic/holistic, descriptive scale, analytical scale**.
- 42 State the criteria used in marking. Choose from: **grammatical range, grammatical accuracy, lexical range, lexical accuracy, cohesion and coherence, content/task fulfilment, development of ideas, orthography**, other (specify).
- 43 If clerically marked, the number of raters will be **one** or more. However, responses may only be second- or third-marked in some cases and by fellow raters, or by more senior raters. If so, insert ‘+ more in selected cases’ after the base number of raters.
- 44 Quantitative feedback routinely given (for the writing component). Choose from: **raw score, percentage score, ranking in candidature, CEFR level, exam-specific grade, pass/fail status**, other (specify).
- 45 Qualitative feedback routinely given (for the writing component). Choose from: **comments for each of the rating criteria, holistic comments**, other (specify).
- 46 Insert a sample response to the task.
- 47 An explanation or justification of the grade awarded to the sample response.
- 48 The grade (or score) awarded to this sample response.

# The CEFR Grid for Writing Tasks

v. 3.1

(Analysis)

This Grid has been developed by the ALTE Manual Special Interest Group in order to assist test providers in their work with the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* and the *Manual for Relating Language Examinations to the CEFR*, both available from the Language Policy Division of the Council of Europe.

There are two varieties of this Grid: the **Analysis** Grid and the **Presentation** Grid (a simplified version).

The **Analysis** Grid is intended to be used in workshops and benchmarking events.

- If a workshop is intended to analyse test content and specifications, the relevant stage of the Manual is *specification*. (Chapter 4.)
- If the Grid is used for benchmarking new, local samples, the relevant section of the Manual is Section 5.6.

The **Presentation** Grid provides a descriptive record of the analysis of test tasks in a previous benchmarking exercise. If completed Grids are used to document illustrative samples, they can be exploited in *standardisation training* (Chapter 5 of the Manual).

## Sample Test Tasks

Report on analysis of	
Target language of this test	
Target level (CEFR) of this test	
Task number/name	

### General Information – the whole test

1	Total test time	minutes
2	Purpose	general proficiency
		specific purpose (specify):

3 Background to the examination

4 Candidature

5 Structure of the test

### General Information – the writing component

6	Number of tasks in the writing paper	1	2	3	4 or more		
7	Total component time	minutes					
8	Integration of skills	none		reading			
		speaking		listening			
		a combination (specify):					
9	Channel	handwritten	word processed	either			
10	CEFR level of this component	A1	A2	B1	B2	C1	C2

11 The writing component format

12 Specific information – example task

13 Mark distribution

14 Task rating

15 Effective level

16 Sample task:

– sample task here –
----------------------

i) Task input/prompt							
17	Language of input/prompt						
18	CEFR level of input/prompt	A1	A2	B1	B2	C1	C2
19	Time permitted or suggested for this task	minutes					
20	Control/guidance	controlled		semi-controlled		open-ended	
21	Content	fully specified		specified to some extent		not specified	
22	Genre of input	letter (business)				letter (personal)	
		review				academic essay	
		composition				report	
		story				proposal	
		article				form	
		other (specify):					
23	Rhetorical function(s) of input	describing (events)			describing (processes)		
		narrating			commentating		
		expositing			explaining		
		demonstrating			instructing		
		arguing			persuading		
		reporting events			giving opinions		
		making complaints			suggesting		
		comparing and contrasting			exemplifying		
		evaluating			expressing possibility		
		expressing probability			summarising		
		other (specify):					
24	Imagined audience for input	friend(s)/acquaintance(s)			general public		
		employer(s)			employee(s)		
		teacher(s)			student(s)		
		committee			business(es)		
		other (specify):					

25	Mode of input/prompt	oral	written	
		visual	a combination	
26	Topic or theme of input	personal identification	house and home, environment	
		daily life	free time, entertainment	
		travel	relations with other people	
		health and body care	education	
		education	shopping	
		food and drink	services	
		places	language	
		weather		
		other (specify):		
27	Integration of skills for input	reading	listening	a combination

**ii) Response (description of written response elicited by the prompt(s)/input)**

28	Number of words expected	0 – 50	51 – 100	101 – 150
		151 – 200	201 – 250	251 – 300
		301 – 350	351 – 400	more than 400
29	Rhetorical function(s) expected	describing (events)	describing (processes)	
		narrating	commentating	
		expositing	explaining	
		demonstrating	instructing	
		arguing	persuading	
		reporting events	giving opinions	
		making complaints	suggesting	
		comparing and contrasting	exemplifying	
		evaluating	expressing possibility	
		expressing probability	summarising	
		other (specify):		
30	Text purpose	referential	emotive	
		conative	phatic	
		metalingual	poetic	
31	Register	informal	unmarked to informal	
		unmarked	unmarked to formal	
		formal		

32	Domain	personal			public		
		occupational			educational/academic		
33	Grammatical competence expected	A1	A2	B1	B2	C1	C2
34	Lexical competence expected	A1	A2	B1	B2	C1	C2
35	Discoursal competence expected	A1	A2	B1	B2	C1	C2
36	Authenticity: situational	low		medium		high	
37	Authenticity: interactional	low		medium		high	
38	Cognitive processing	reproduction of known ideas					
		knowledge transformation					
39	Content knowledge required	general/non-specialised			specialised knowledge		
		very specialised knowledge			a range of knowledge		

### iii) Rating of task

40	Known criteria						
41	Task rating method	impressionistic/holistic			descriptive scale		
		analytical scale			with compensation system		
		other (specify):					
42	Assessment criteria	grammatical range			grammatical accuracy		
		lexical range			lexical accuracy		
		cohesion and coherence			content/task fulfilment		
		development of ideas			orthography		
		other (specify):					
43	Number and combination of raters	1			2		
		3 or more			1 + more in selected cases		
		2 + more in selected cases			computer rated		

### iv) Feedback to candidates

44	Quantitative feedback	raw score			percentage score		
		ranking in candidature			CEFR level		
		exam-specific grade			pass/fail status		
		other (specify):					
45	Qualitative feedback	comments for each rating criteria					
		holistic comments					
		other (specify):					

## 46 Example answer

## 47 Commentary

## 48 Score allocated

### Notes:

All references to the *CEFR* are to the document on the Council of Europe Language Policy Division's website.

Numbers below correspond to numbered items in the Grid.

- 2 The purpose of the test may be general proficiency, for a specific purpose. State the purpose if specific (English for Legal Purposes, German for Academic Purposes, etc.).
- 3 The description of test background may contain the reasons for developing the test, a description of the suite of which this test is a part, or other such details.
- 4 Describe the size and demographic profile of the candidature.
- 5 Describe the other components of the test (e.g. the speaking component, the reading component).
- 6 In the case that the number of tasks depends on which options are chosen, specify in the introductory text (point 5).
- 8 Skills, in addition to writing, which are involved in the completion of this task (regardless of whether they are explicitly recognised at the rating stage). Choose from: none, reading, speaking, listening, a combination.
- 9 The method by which the candidate's response is recorded. Choose from handwritten, word processed, either.
- 10 *CEFR*, Ch. 3.
- 11 The description may include information such as the number of subsections, task types in each subsection, time allowed for each subsection.
- 12 You may wish to include a short description of the task here. The description could include the aims of the task, what candidates have been asked to do and would constitute a full completion of the task.
- 13 Describe how marks are distributed in this section of the task and what candidates would need to include to achieve full marks on this task.
- 14 Explain how the task is rated (e.g. clerically, machine marked), what instruments are used and what aspects are considered when deciding the grade.
- 15 Describe the measures taken to ensure Writing tasks are set at the appropriate level. This description may include the process of question paper production and trialling.
- 16 Insert the sample task, including rubric and prompt/input.
- 18 Choose *CEFR* level: A1, A2, B1, B2, C1, C2.
- 19 If not specified, expected time.
- 20 The extent to which the rubric, prompt or input determines the nature and content of the response. Choose from: controlled, semi-controlled or open-ended.
- 21 Whether the content of the response is specified in the rubric. Choose from: specified or not specified.
- 22 Choose from: **letter (business), letter (personal), review, academic essay, composition, report, story, proposal, article, form**, other (specify).
- 23 The functions which might be expected in the response. Choose from: **describing (events), describing (processes), narrating, commentating, expositing, explaining, demonstrating, instructing, arguing, persuading, reporting events, giving opinions, making complaints, suggesting, comparing and contrasting, exemplifying, evaluating, expressing possibility/probability, summarising**, other (specify). *CEFR*, p125–130.
- 24 The imagined audience for the input. Choose from: **friend/acquaintance, teacher, employer, employee, committee, board, business, students, general public** (e.g. with a newspaper article), other (specify).
- 25 Choose from: **oral, written or visual**, or **a combination**.
- 26 The topic or theme. Choose from: **personal identification, house and home/environment, daily life, free time/entertainment, travel, relations with other people, health and body care, education, shopping, food and drink, services, places, language, weather**, other (specify). *CEFR*, p 51 – 53.

- 27 The language skills the candidate needs to understand the rubric and prompt/input. Choose from: **reading, listening, or a combination.**
- 29 The functions which might be expected in the response. Choose from: **describing (events), describing (processes), narrating, commentating, expositing, explaining, demonstrating, instructing, arguing, persuading, reporting events, giving opinions, making complaints, suggesting, comparing and contrasting, exemplifying, evaluating, expressing possibility/probability, summarising**, other (specify). *CEFR*, p125–130.
- 30 The expected purpose(s) of the response. Choose from: referential (to give “objective” facts about the world), emotive (to describe the emotional state of the writer), conative (to persuade the reader(s)), phatic (to establish or maintain social contact with the reader(s)), metalingual (to clarify or verify understanding), poetic (writing for aesthetic purposes).
- 31 The register the candidate is expected to adopt in their response. Choose from: **informal, unmarked to informal, unmarked, unmarked to formal, formal.** *CEFR*, p118–122.
- 32 The domain to which the expected response is imagined to belong. Choose from: **personal, public, occupational, educational/academic.** *CEFR*, p45–46.
- 33 Choose *CEFR* level: **A1, A2, B1, B2, C1, C2.** *CEFR*, p112–116.
- 34 Choose *CEFR* level: **A1, A2, B1, B2, C1, C2.** *CEFR*, p110–112.
- 35 Choose *CEFR* level: **A1, A2, B1, B2, C1, C2.** *CEFR*, p123–125.
- 36 The extent to which the task reflects a real-life activity a candidate could perform. Choose from **low, medium, or high.**
- 37 The extent to which interaction patterns are likely to mirror those in an equivalent, real-life task. Choose from **low, medium, or high.**
- 38 The difficulty in performing the task from a non-linguistic point-of-view. Choose from: **reproduction of known ideas, knowledge transformation.**
- 39 The kind of extra-linguistic knowledge required to complete the task. Choose from: **personal/everyday life knowledge areas, general/non-specialised knowledge areas, specialised knowledge areas** (scientific, study-related, etc.), **a wide range of knowledge areas.**
- 40 Describe the rating criteria made available to the candidate, either before or during the test. If the criteria are not available together with the paper, state where they can be viewed.
- 41 Choose from: **impressionistic/holistic, descriptive scale, analytical scale.**
- 42 State the criteria used in marking. Choose from: **grammatical range, grammatical accuracy, lexical range, lexical accuracy, cohesion and coherence, content/task fulfilment, development of ideas, orthography**, other (specify).
- 43 If clerically marked, the number of raters will be **one** or more. However, responses may only be second- or third-marked in some cases and by fellow raters, or by more senior raters. If so, insert ‘+ more in selected cases’ after the base number of raters.
- 44 Quantitative feedback routinely given (for the writing component). Choose from: **raw score, percentage score, ranking in candidature, CEFR level, exam-specific grade, pass/fail status**, other (specify).
- 45 Qualitative feedback routinely given (for the writing component). Choose from: **comments for each of the rating criteria, holistic comments**, other (specify).
- 46 Insert a sample response to the task.
- 47 An explanation or justification of the grade awarded to the sample response.
- 48 The grade (or score) awarded to this sample response.

# The CEFR Grid for Speaking Tasks v. 3.1 (Presentation)

This Grid has been developed by the ALTE Manual Special Interest Group in order to assist test providers in their work with the *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)* and the *Manual for Relating Language Examinations to the CEFR*, both available from the Language Policy Division of the Council of Europe.

There are two varieties of this Grid: the **Analysis** Grid and the **Presentation** Grid (a simplified version).

The **Analysis** Grid is intended to be used in workshops and benchmarking events.

- If a workshop is intended to analyse test content and specifications, the relevant stage of the Manual is *specification*. (Chapter 4.)
- If the Grid is used for benchmarking new, local samples, the relevant section of the Manual is Section 5.6.

The **Presentation** Grid provides a descriptive record of the analysis of test tasks in a previous benchmarking exercise. If completed grids are used to document illustrative samples, they can be exploited in *standardisation training* (Chapter 5 of the Manual).

1	Report on analysis of	
2	Target language	

## 1 GENERAL INFORMATION (whole speaking test)

3	N°. of tasks in speaking component	
4	Integration of skills	
5	Total duration of speaking component	
6	Target performance level	
7	Channel	
8	Test purpose	

## 2 TASK INPUT/PROMPT for task n°.name

9	Language of instructions/rubric	
10	Channel	
11	Language level of instructions/rubric	
12	Task duration (minutes)	
13	N°. assessors present	

14	Recorded?	
15	Control/guidance by task	
16	Control/guidance by interlocutor	
17	Specification of content	
18	Interaction type	
19	Discourse mode (genre)	
20	Audience (real)	
21	Audience (imagined, as in role play)	
22	Type of prompt	
23	Topic	
24	Planning time	
25	Setting (imagined)	

**3 RESPONSE (the expected spoken response elicited by the prompt(s)/input)**

26	Length of response	
27	Text type	
28	Rhetorical function(s)	
29	Register	
30	Domain	
31	Grammatical level	
32	Lexical level	
33	Discourse features	
34	Situational authenticity	
35	Interactional authenticity	
36	Cognitive processing	

37	Content knowledge	
38	Task purpose	

#### 4 RATING OF TASK

39	Known criteria	
40	Rating method	
41	Assessment criteria	
42	N°. of raters	
43	Use of moderator	

#### 5 FEEDBACK TO CANDIDATE

44	Quantitative feedback	
45	Qualitative feedback	

# The CEFR Grid for Speaking Tasks v. 3.1 (Analysis)

This Grid is designed to elicit information pertaining to ONE task in the test indicated. The GENERAL INFORMATION section (Section 1) refers to the speaking test **as a whole**. Other sections refer to an individual task within the test.

For definitions (and translations) of terminology, users are referred to the **ALTE Multilingual Glossary of Testing Terms** (Cambridge University Press).

## 1 GENERAL INFORMATION (whole speaking test)

0	Name of test provider						
1	Name of test						
	Component	speaking component					
2	Target language						
3	N°. of tasks in the speaking component	1	2	3	4 or more		
4	Integration of skills <sup>49</sup> (circle at least one)	speaking (no other skill involved)	reading	writing	listening		
	Comment						
5	Total duration of speaking component (including preparation time)	approx. .... minutes (of which ..... minutes preparation time)					
6	Target performance level CEFR – General (p26, p58) (Also appendix D for ALTE “ <i>Can Dos</i> ” – p244) (circle at least one)	A1	A2	B1	B2	C1	C2
7	Channel	face to face	phone	computer	video conference	tape recorder	video recorder
				aud			
8	Test Purpose	general proficiency		specified purpose (Language for Specific Purposes):			

<sup>49</sup> To what extent the **whole speaking component** involves integration with another skill. Is this integration explicit or implicit? Bear in mind, that even a written prompt implies a degree of skill integration, which may or may not be recognised at the rating stage.

The following tables (sections 2–6) refer to ONE TASK in the subtest. Fill in the Grid in relation to **each one of tasks** on the subtest.

**2 TASK INPUT/PROMPT** – Rubric and prompts (verbal, iconic) or other forms of input designed to elicit the required response(s) in the target language.

0	<b>To which task in the speaking component of the test does the information relate?</b>					
9	Language of instructions/rubric	language of test provider	target language of test		other language ?	
10	Instructions spoken or written (channel)	spoken	written	recorded	pictorial/iconic	
11	Level of language of instructions/ rubric	much easier than level of test	easier than level of test	same as level of test	more difficult than level of test	
12	Task duration (minutes)	approx..... minutes				
13	N <sup>o</sup> . of assessors present	0	1	2		
14	Recorded?	yes – audio	yes – video	no		
15	Control/guidance by the task (flexibility of task frame <sup>50</sup> )	rigidly controlled	partially controlled		open format	
16	Control/guidance by interlocutor (flexibility of interlocutor frame <sup>51</sup> )	rigidly controlled format (e.g. list of questions to be asked)	partially controlled format (e.g. interview in controlled format with specified topic)		open format (e.g. undirected interview or discussion)	
17	Specification of content	specified			not specified	
18	Interaction type	dialogue: paired candidates	dialogue: grouped candidates	dialogue: candidate/ examiner	dialogue: simulated/ recorded prompts	monologue
		repetition of prompt	role play	reading aloud	react to a prompt	other:
19	Discourse mode (genre)	interview			story telling (narration)	
		speech, presentation			discussion/conversation	
20	Audience (real)	assessor	other candidate	teacher	none (e.g. tape recorder)	other:

<sup>50</sup> The extent to which the task frame guides or limits the response of the candidate.

<sup>51</sup> The extent to which the interlocutor frame controls the input from the examiner/assessor/interviewer in a way that determines the nature and content of the interaction. The input from interlocutor may be largely unguided, resulting in free or creative speaking. Is the content which is expected in the response specified by the interviewer examiner?

21	Audience (imagined, as in role play)	employer	committee, board	business, shop, etc.	teacher	answering machine	
		general public	family member	friend, acquaintance	other: (specify)		
22	Type of prompt (select at least one)	oral only (given orally by examiner)					
		textual (written)		written sentence, question, instructions			
				letters	e.g. to pen-friend		
				notes, messages, memos	e.g. office memo		
				adverts			
				programmes	e.g. theatre, football, etc.		
				forms	e.g. fill immigration form		
				excerpts	books/journals magazines/newspapers		
		iconic		graph	annotated/ not annotated		
				chart			
				table			
				diagram			
				map			
				sequence of diagrams			
		pictorial (non-verbal)		photo(s)			
				drawing, sketch			
sequence of pictures							
other (specify)							

23	Topic CEFR p52 (select at least one)	personal identification		current affairs	
		house/home/environment		shopping	
		daily life		food and drink	
		free time, entertainment		services	
		travel		places	
		relations with other people		language	
		health and body care		weather	
		education		celebrities	
		science and environment		work environment	
		other (please specify):			
24	Planning time	30 secs	1 min	2 mins	not applicable
		comments			
25	Setting (imagined)	workplace	social	educational	other

### 3 RESPONSE (the expected spoken response elicited by the prompt(s)/input)

26	Length of response expected	30 secs	1min	2 mins	3mins	4mins	5mins	>5mins
27	Text type	word level		phrase	discourse level			
28	Rhetorical function(s) CEFR p126	description (events) description (process) description (data) description (objects) description (pictures) narration commentary presentation explanation demonstration		instruction argumentation persuasion reporting events giving opinions making complaints suggestion comparison and contrast		exemplification synthesis analysis evaluation expressing possibility/probability summarising asking for information other: (specify) .....		
29	Register CEFR p120	informal		neutral		formal		
30	Domain CEFR p45	personal		public		occupational		educational/ academic

31	Grammatical level CEFR, p114	only simple grammatical structures	mainly simple structures	limited range of complex structures	wide range of complex grammatical structures	
32	Lexical level CEFR, 112	only frequent vocabulary	mainly frequent vocabulary	extended vocabulary	wide range of advanced vocabulary	wide range of advanced and specialised vocabulary
33	Discourse features (e.g. cohesion) CEFR, p125	extremely limited use	limited	competent use	advanced use	
34	Situational authenticity <sup>52</sup>	low	medium	high		
35	Interactional authenticity <sup>53</sup>	low	medium	high		
36	Cognitive processing <sup>54</sup>	reproduction of known ideas only		knowledge transformation		
37	Content knowledge	personal/daily life/basic communication needs	common, general, non-specialised	wide range of non-specialised knowledge areas	very wide range of knowledge areas (social, scientific, study-related, sometimes specialised, etc.)	
38	Task purpose	referential (telling)	emotive (reacting)	conative <sup>55</sup>	phatic <sup>56</sup>	

#### 4 RATING OF TASK

39	Known criteria	are the grading criteria available to the candidate ON THE PAPER and is s/he familiar with them? Y/N If no, where can these be viewed?					
40	Rating method	impressionistic/holistic	descriptive scale (band descriptors)	analytical method			
41	Assessment criteria	grammatical accuracy	cohesion and coherence	lexical control	content	interactive communication	development of ideas
		pronunciation (phonological)		pronunciation (intonation and stress)	other:		
42	No. of raters	1	2	3	computer rated		
		other (explain)					
43	Use of moderator <sup>57</sup>	YES			NO		

<sup>52</sup> To what extent does the task reflect a real life activity that the candidate is likely to perform?

<sup>53</sup> Conative refers to tasks which require that the candidate argue, persuade, discuss for and against, etc.

<sup>54</sup> How difficult the task is to perform from a non-linguistic point of view; e.g., how difficult the iconic prompts are to interpret if presented in graphic form, which may be unfamiliar to the candidate.

<sup>55</sup> Conative refers to tasks which require that the candidate argue, persuade, discuss for and against, etc.

<sup>56</sup> Phatic – intending to keep in touch with correspondent(s).

<sup>57</sup> A moderator checks that rating criteria are observed consistently and ensures that grades have been allocated correctly and fairly by examiners.

## 5 FEEDBACK TO CANDIDATE

44	Quantitative feedback <sup>58</sup>	raw score	score as %	ranking (e.g. quartile)	CEFR level	exam specific grade	pass/fail only	other:
		(tick here)						
		<input type="checkbox"/>						
45	Qualitative feedback	grammar	lexis	cohesion/coherence	content	development of ideas	task relevance	other:
		<input type="checkbox"/>						

---

<sup>58</sup> Information given to candidates regarding their performance on the task.

## Appendix C

### Forms and Scales for Standardisation & Benchmarking Chapter 5

Training Record Form			
<b>Location</b>		<b>Date:</b>	
<b>Coordinator</b>	<b>Name:</b>	<b>Institution/Project:</b>	
<b>Stage</b>	Familiarisation Training Benchmarking	<input type="checkbox"/>	<input type="checkbox"/>
<b>Area(s)</b>	Assessing Spoken Samples Assessing Written Samples Test Tasks/ Items: ▪ Listening ▪ Reading ▪ Linguistic Competence Other: _____	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
<b>Participants</b>	<b>Number:</b>	<b>Functions:</b>	
<b>Activities completed</b>	Familiarisation Illustration with exemplars Controlled/free practice with exemplars Benchmarking local performance samples Training with exemplar tasks Judging item difficulty Feedback of actual item difficulty Other _____	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
<b>Materials used</b>	CEFR exemplar samples CEFR rating instruments (Tables 5.4, 5.5, 5.8) Local performance samples Adapted rating instruments (to be appended) CEFR exemplar test tasks and items Local test tasks and items Other _____	<input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/> <input type="checkbox"/>	
<b>Information on tasks and items</b>			
<b>Additional comment</b>			
<b>Dissemination procedures planned</b>			

*Form C1 Training Record Form*

**LEARNER'S NAME**/Ihr Name/Votre nom: \_\_\_\_\_

Niveaus/Niveaux: **A1, A2, A2+, B1, B1+, B2, B2+, C1, C2**

1. Initial Impression

Einstufung mit der Globalskala  
Classement – échelle globale

2. Detailed Analysis with Grid / Beurteilung mit Raster / Estimation – grille

<b>RANGE</b> Spektrum Étendue	<b>ACCURACY</b> Korrektheit Correction	<b>FLUENCY</b> Flüssigkeit Aisance	<b>INTERACTION</b> Interaktion Interaction	<b>COHERENCE</b> Kohärenz Cohérence

3. Considered Judgment

Abschliessende Einstufung  
Classement final

**Form C2: Analytic Rating Form**

***Eurocentres (North 1991/1992) / Swiss Project (Schneider and North 2000)***

<b>Skill:</b> _____	<b>Level assigned</b>	<b>Comments</b>
Sample/Task 1		
Sample/Task 2		
Sample/Task 3		
Sample/Task 4		
Sample/Task 5		
Sample/Task 6		
Sample/Task 7		
Sample/Task 8		

*This is an example of a simple rating sheet which requires the participant to give one global judgment about the level of each sample or task. This rating sheet can be used to rate either performances or test items.*

**Form C3: Holistic Rating Form (DIALANG)**

	<i>Mickey Mouse</i>	<i>Donald Duck</i>	<i>Groover</i>	<i>Fred</i>	<i>Henry VII</i>	<i>Susi Q</i>	<i>Other code names</i>	<i>Other code names</i>	<i>Other code names</i>
<i>Item 1</i>									
<i>Item 2</i>									
<i>Item 3</i>									
<i>Item 4</i>									

**Form C4: Collation Global Rating Form (DIALANG)**

<b>Skill _____</b>	<b>Descriptor Operationalised (List subscale and level)</b>	<b>CEFR level assigned</b>	<b>Comments (Include references to Form A10)</b>
Item 1			
Item 2			
Item 3			
Item 4			
Item 5			
etc.			

**Form C5: Item Rating Form (DIALANG)**

**Table C1: GLOBAL ORAL ASSESSMENT SCALE**

<b>C2</b>	<p><b><i>Conveys finer shades of meaning precisely and naturally.</i></b></p> <p>Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions.</p>
<b>C1</b>	<p><b><i>Shows fluent, spontaneous expression in clear, well-structured speech.</i></b></p> <p>Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy; errors are rare.</p>
<b>B2+</b>	
<b>B2</b>	<p><b><i>Expresses points of view without noticeable strain.</i></b></p> <p>Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding.</p>
<b>B1 +</b>	
<b>B1</b>	<p><b><i>Relates comprehensibly the main points he/she wants to make.</i></b></p> <p>Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations.</p>
<b>A2+</b>	
<b>A2</b>	<p><b><i>Relates basic information on, e.g. work, family, free time etc.</i></b></p> <p>Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes.</p>
<b>A1</b>	<p><b><i>Makes simple statements on personal details and very familiar topics.</i></b></p> <p>Can make him/herself understood in a simple way, asking and answering questions about personal details, provided the other person talks slowly and clearly and is prepared to help. Can manage very short, isolated, mainly pre-packaged utterances. Much pausing to search for expressions, to articulate less familiar words.</p>
<b>Below A1</b>	Does not reach the standard for A1.
<ul style="list-style-type: none"> <li>• <b><i>Use this scale in the first 2–3 minutes of a speaking sample to decide approximately what level you think the speaker is.</i></b></li> <li>• <b><i>Then change to Table C2 (CEFR Table 3) and assess the performance in more detail in relation to the descriptors for that level.</i></b></li> </ul>	

**Table C2: ORAL ASSESSMENT CRITERIA GRID (CEFR Table 3)**

	<b>RANGE</b>	<b>ACCURACY</b>	<b>FLUENCY</b>	<b>INTERACTION</b>	<b>COHERENCE</b>
<b>C2</b>	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 turntaking, 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.
<b>C1</b>	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.
<b>B2+</b>					
<b>B2</b>	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 or 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.
<b>B1+</b>					
<b>B1</b>	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.
<b>A2+</b>					
<b>A2</b>	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 ask and 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".
<b>A1</b>	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".

**Table C3: SUPPLEMENTARY CRITERIA GRID: “Plus Levels”**

	RANGE	ACCURACY	FLUENCY	INTERACTION	COHERENCE
<b>C2</b>					
<b>C1</b>					
<b>B2+</b>	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	Shows good grammatical control; occasional “slips” or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.	Can intervene appropriately in discussion, exploiting a variety of suitable language to do so, and relating his/her own contribution to those of other speakers.	Can use a variety of linking words efficiently to mark clearly the relationships between ideas.
<b>B2</b>					
<b>B1+</b>	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influences.	Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and “cul-de-sacs”, he/she is able to keep going effectively without help.	Can exploit a basic repertoire of strategies to keep a conversation or discussion going. Can give brief comments on others’ views during discussion. Can intervene to check and confirm detailed information.	<i>No descriptor available</i>
<b>B1</b>					
<b>A2+</b>	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics, though he/she will generally have to compromise the message and search for words.	<i>No descriptor available</i>	Can adapt rehearsed memorised simple phrases to particular situations with sufficient ease to handle short routine exchanges without undue effort, despite very noticeable hesitation and false starts.	Can initiate, maintain and close simple, restricted face-to-face conversation, asking and answering questions on topics of interest, pastimes and past activities. Can interact with reasonable ease in structured situations, given some help, but participation in open discussion is fairly restricted.	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points.
<b>A2</b>					
<b>A1</b>					

**Table C4: WRITTEN ASSESSMENT CRITERIA GRID**

	<b>Overall</b>	<b>Range</b>	<b>Coherence</b>	<b>Accuracy</b>	<b>Description</b>	<b>Argument</b>
<b>C2</b>	Can write clear, <i>highly accurate</i> and smoothly flowing complex texts in an appropriate and effective <i>personal style conveying finer shades of meaning</i> . Can use a logical structure which helps the reader to find significant points.	Shows great flexibility in <i>formulating</i> ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Can create coherent and cohesive texts making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.	Maintains consistent and <i>highly accurate</i> grammatical control of <i>even the most complex language forms</i> . <i>Errors are rare and concern rarely used forms</i> .	Can write clear, smoothly flowing and fully engrossing stories and descriptions of experience in a style appropriate to the genre adopted.	Can produce clear, smoothly flowing, complex reports, articles and essays which present a case or give critical appreciation of proposals or literary works. Can provide an appropriate and effective logical structure which helps the reader to find significant points.
<b>C1</b>	Can write clear, well-structured and <i>mostly accurate</i> texts of complex subjects. Can <i>underline</i> the relevant salient issues, <i>expand and support</i> points of view at some length with subsidiary points, reasons and relevant examples, and <i>round off</i> with an appropriate conclusion.	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. <i>The flexibility in style and tone is somewhat limited</i> .	Can produce clear, smoothly flowing, well-structured text, showing controlled use of organisational patterns, connectors and cohesive devices.	Consistently maintains a high degree of grammatical accuracy; <i>occasional errors in grammar, collocations and idioms</i> .	Can write clear, detailed, well-structured and developed descriptions and imaginative texts in a mostly assured, personal, natural style appropriate to the reader in mind.	Can write clear, well-structured expositions of complex subjects, underlining the relevant salient issues. Can expand and support point of view with some subsidiary points, reasons and examples.
<b>B2</b>	Can write clear, detailed <i>official and semi-official</i> texts on a variety of subjects related to his field of interest, synthesising and evaluating information and arguments from a number of sources. <i>Can make a distinction between formal and informal language with occasional less appropriate expressions</i> .	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, using some complex sentence forms to do so. <i>Language lacks, however, expressiveness and idiomaticity and use of more complex forms is still stereotypic</i> .	Can use a number of cohesive devices to link his/her sentences into clear, coherent text, though there may be some "jumpiness" in a longer text.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstandings.	Can write clear, detailed descriptions of real or imaginary events and experiences marking the relationship between ideas in clear connected text, and following established conventions of the genre concerned. Can write clear, detailed descriptions on a variety of subjects related to his/her field of interest. Can write a review of a film, book or play.	Can write an essay or report that develops an argument systematically with appropriate highlighting of some significant points and relevant supporting detail. Can evaluate different ideas or solutions to a problem. Can write an essay or report which develops an argument, giving some reasons in support of or against a particular point of view and explaining the advantages and disadvantages of various options. Can synthesise information and arguments from a number of sources.
<b>B1</b>	Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. <i>The texts are understandable but occasional unclear expressions and/or inconsistencies may cause a break-up in reading</i> .	Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Can link a series of shorter discrete elements into a connected, linear text.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more <i>common</i> situations. <i>Occasionally makes errors that the reader usually can interpret correctly on the basis of the context</i> .	Can write accounts of experiences, describing feelings and reactions in simple connected text. Can write a description of an event, a recent trip – real or imagined. Can narrate a story. Can write straightforward, detailed descriptions on a range of familiar subjects within his field of interest.	Can write short, simple essays on topics of interest. Can summarise, report and give his/her opinion about accumulated factual information on a familiar routine and non-routine matters, within his field with some confidence. Can write very brief reports to a standard conventionalised format, which pass on routine factual information and state reasons for actions.
<b>A2</b>	Can write a series of simple phrases and sentences linked with simple connectors like "and", "but" and "because". <i>Longer texts may contain expressions and show coherence problems which makes the text hard to understand</i> .	Uses basic sentence patterns with memorized phrases, groups of a few words and formulae in order to communicate limited information mainly in everyday situations.	Can link groups of words with simple connectors like "and", "but" and "because".	Uses simple structures correctly, but still systematically makes basic mistakes. <i>Errors may sometimes cause misunderstandings</i> .	Can write very short, basic descriptions of events, past activities and personal experiences Can write short simple imaginary biographies and simple poems about people.	
<b>A1</b>	Can write simple isolated phrases and sentences. <i>Longer texts contain expressions and show coherence problems which make the text very hard or impossible to understand</i> .	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Can link words or groups of words with very basic linear connectors like "and" and "then".	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorized repertoire. <i>Errors may cause misunderstandings</i> .	Can write simple phrases and sentences about themselves and imaginary people, where they live and what they do, etc.	

