Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR)

Highlights from the Manual

Edited by the RelEx team
José Noijons, Jana Bérešová, Gilles Breton and Gábor Szabó

European Centre for Modern Languages

Council of Europe Publishing

French edition:

Relier les examens de langues au Cadre européen commun de référence pour les langues : Apprendre, enseigner, évaluer (CECR)
Les points essentiels du Manuel
ISBN: 978-92-871-7168-9

Cover: Georg Gross
Layout: Stenner Medienproduktion
Copy-editing: Robert Blackwell

http://book.coe.int
Council of Europe Publishing
F-67075 Strasbourg Cedex

European Centre for Modern Languages / Council of Europe
Nikolaiplatz 4
A-8020 Graz
www.ecml.at

ISBN: 978-92-871-7169-6

Printed in Austria

Table of contents

This publication is the result of a project of the European Centre for Modern Languages entitled "Training in relating examinations to the Common European Framework of Reference for Languages" (RelEx).

For further information and materials relating to this publication, visit the websites http://relex.ecml.at and http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp.

Foreword

Waldemar Martyniuk
Executive Director, European Centre for Modern Languages

This publication is a result of RelEx – an ECML project on relating language examinations to the Common European Framework of Reference for Languages (CEFR) elaborated by the Council of Europe. Responding to a growing need of the users of the CEFR, the Language Policy Division of the Council of Europe has developed a manual and a set of accompanying tools to be used in order to establish a link between local language examinations and the common reference levels of language proficiency in a reliable and responsible way. One of the planned outputs of the RelEx project has been the production of a user friendly introduction to the Manual on relating language examinations to the CEFR. These Highlights from the Manual are aimed at policy makers, assessment experts at examination centres, curriculum developers, teacher trainers and other educationalists less familiar with the technicalities of the linking process. The message of the RelEx project has been that linking foreign language examinations to the CEFR is an activity worth doing. However, it is also stressed that linking requires several stages to be carried out with due care.

The project team has found that an introduction to the Manual like the present one, which highlights the most important aspects of the linking process, can play a useful role in making the Manual better known to stakeholders. These Highlights have proved to be useful for teacher trainers participating in the project. Practising and future teachers are made aware that claims of links to the CEFR need to be validated. When considering using existing assessment instruments teachers can look for information and evidence for such validation. Teachers can also apply some of the procedures as outlined in (these Highlights from) the Manual when developing their own classroom-based tests.

The Highlights from the Manual have been used in workshops organised by the ECML with participants from all its member states. Participants have commented positively on their conciseness and their accessibility for the non-specialist. It must be noted that these Highlights have all been taken exclusively from the Manual, and that no text has been changed or added to, except in a few cases where a technical term has been explained.

The ECML believes that this publication will constitute a useful new tool for linking assessment instruments of various kinds to the Common European Framework of Reference for Languages. Synergies also exist between this publication and other tools developed in this area by the Centre. The publication *Pathways through assessing,*

learning and teaching in the CEFR resulting from the ECML's "Encouraging the culture of evaluation among professionals" (ECEP) project represents a complementary approach. Both publications are available for download from the ECML website, http://www.ecml.at.

Chapter 1: The CEFR and 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 of Reference for Languages (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 in demonstrating the validity of those claims.

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

- describing the examination coverage, administration and analysis procedures;
- relating results reported from the examination to the CEFR common reference levels;
- providing 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 co-operation 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 to achieve 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 argumentation as to 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 to how to construct good language tests or examinations. There are several useful guides that do this, 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 (for example: the American Educational Research Association (AERA) (AERA/APA/NCME 1999), EALTA (www.ealta.org) and the ALTE (www.alte.org)), which are a source of further support and advice on language testing and linking procedures.

Despite the above, the earlier 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 test results 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;

- 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;
- 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.

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.

To ensure coherence, the numbering system applied throughout the publication for tables, forms and diagrams reflects the references used in the Manual.

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 subdivide, 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 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 mutual recognition of language qualifications awarded by all relevant bodies is a 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 prerequisite 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.

There is no suggestion that different examinations that have been linked to the CEFR by following procedures such as those proposed in the 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.

Neither is it the intention of the Manual to tell language professionals what their standards should be, or how they should prove linkage to them. Both the CEFR and the 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 co-operate autonomously; it is their privilege and responsibility to choose approaches appropriate to their purpose and context.

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

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. 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 CEFR provides the content and performance level descriptors (PLDs). 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 below.

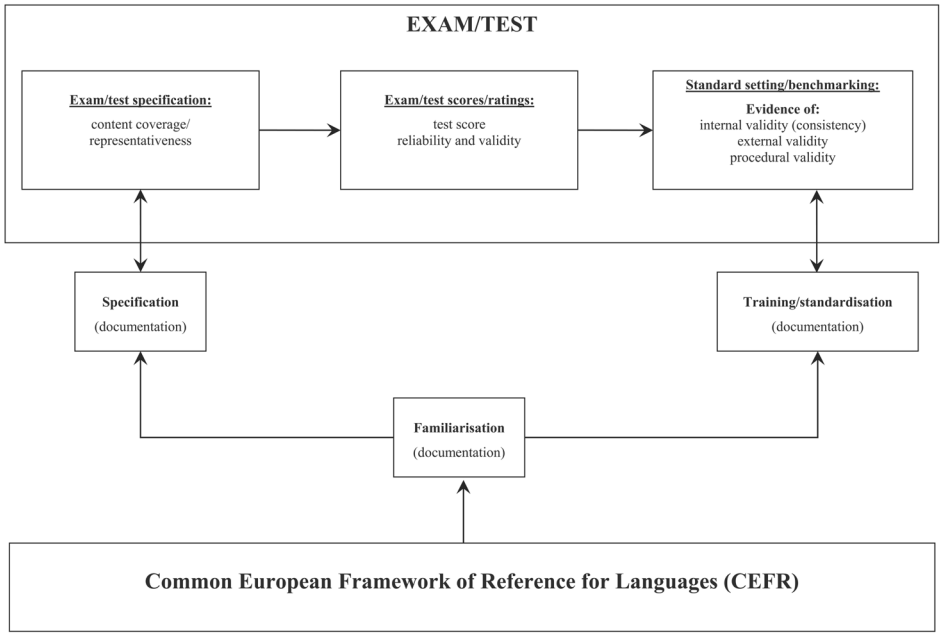


Figure 2.1: Validity evidence of linkage of examination/test results to the CEFR

The approach adopted in the Manual is such that thorough familiarity with the CEFR is a fundamental requirement.

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.

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 indeed reflect the common interpretation illustrated by so called illustrative samples. In the case of the revision or development of an exami-

nation, 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 it 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).

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;
- 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 as 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 to A20 in Chapter 4 focus on content analysis and the relationship of content to the CEFR;
- 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;

- 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 scores, borderline performances. The preceding stages of familiarisation, specification and standardisation can be seen as preparatory activities that lead to valid and rational decisions. Chapter 6 describes procedures to arrive at the final decision of setting cut scores;
- validation (Chapter 7): while the preceding stages of familiarisation, specification, standardisation and standard setting can be conceived of as roughly representing a chronological order of activities, it would be naive 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?"

2.4. Use of the CEFR

A common framework of reference enables different examinations be to 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.

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.

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;
- co-ordination of the participants in the local linking process;
- thoughtful application of the procedures;
- reliable recording of results;
- accurate, transparent and detailed reporting of conclusions.

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, etc.);
- how feasible the different sets of procedures are in their context in terms of time and money;
- 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, authorities such as school principals and inspectors;
- to what extent linking of the examination or test is required by the law.

Chapter 3: Familiarisation

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. In particular, while most are familiar with the more global CEFR tables (Table 1: Global scale and Table 2: 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.

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 sub-scales plus the summary Tables 1, 2 and 3) and the CEFR "plus levels" (only found on sub-scales).

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 the standardisation process.

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. 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.

It is highly recommended that the co-ordinator 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) two or three weeks before the seminar.

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 relevant sections of the CEFR

The task that participants are given is to read about levels in the CEFR in order to be able to identify the salient features for each level and in order to ascertain at 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). See Table 3.1 at the end of this chapter.

b) Consideration of a selection of CEFR question boxes

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 in scope of the CEFR.

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

- a 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;

Users of the Framework for the purpose of analysing and assessing speaking performances may wish to consider and, where appropriate, state:	Relevant/ why?
▪ how the physical conditions under which the learner will have to communicate will affect what he/she is required to do; ▪ how the number and nature of the interlocutors will affect what the learner is required to do; ▪ within which time constraints the learner will have to operate; ▪ to what extent the learners will need to adjust to the interlocutor's mental context; ▪ 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.	

- the co-ordinators 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 co-ordinators may draw on the work done by the participants in this activity when discussing the sorting exercises ((f) and (g)) in section 3.4.

c) Accessing the CEFTrain website

The CEFTrain project (www.helsinki.fi/project/ceftrain/index.php.35.html, www.webcef.eu/?q=node/49) 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.

3.3. Introductory activities at the seminar

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 co-ordinator 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 1

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 3.1 below, which simplifies CEFR section 3.6. Level references should be eliminated so that participants need to read the descriptors really carefully. The co-ordinator 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 3.1 will be distributed.
- The co-ordinator then asks the participants to share – in pairs or small groups – their views on the salient features of each of the CEFR levels. One way to do this in a concrete fashion is to ask the participants to highlight key elements in colour.

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. Table 3.2 at the end of this chapter is an important part of the ELP and often referred to as the ELP grid.

- Participants are asked to self-assess their ability in two foreign languages with the ELP grid (Table 3.2 at the end of this chapter). 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.
- 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, using Table 3 in the CEFR defining each level for linguistic range, grammatical accuracy, fluency, coherence and interaction. See Table 3.3 at the end of this chapter.

3.4. Qualitative analysis of the CEFR scales

Once the introductory activities phase has been completed, the familiarisation phase should proceed with discussion of CEFR levels in relation to the descriptors for the specific skill concerned.

f) Sorting the individual descriptors from a CEFR scale

This activity is effective because it forces participants to consider the descriptors in isolation from each other as independent criteria.

- The co-ordinator 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 (for example, 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.

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 (Table 3.2 in these Highlights). 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 Table 3.2, but with all the cells empty. Participants can then be asked to place the descriptors in 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 large groups and also with rooms without big tables.

A combination of this reconstruction activity with a self-assessment of own language level ((c), above) is effective if done as follows:

- Participants, in small groups, carefully read and discuss each descriptor to reconstruct the table. The co-ordinator supervises group work and helps to clarify doubts about the interpretation of the different descriptors.
- The co-ordinator distributes a copy of the completed and "whole" Table 3.2 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 Table 3.2.

3.5. Preparation for rating

The last phase of familiarisation involves preparing the participants in more detail for the rating of tasks and performances in the relevant skill(s).

h) Reconstructing the CEFR-based rating grid to be used

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 co-ordinator 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

The activity can only be carried out if the co-ordinator 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 co-ordinator plays the selected performance(s) once and asks the participants to assign a level to it according to Table 3.l.

Before discussion, participants are given Table 3.3 and are asked to confirm their initial level assignment individually.

- The co-ordinator then fosters discussion in groups of the level(s) assigned in relation to the descriptors in Table 3.3.
- The co-ordinator 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 Table 3.3.

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 section 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.

Level		Salient characteristics: interaction and 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	Level C2 is intended to characterise the degree of precision, appropriateness and ease with the language which typifies the speech of those who have been highly successful learners. Descriptors calibrated here include: *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.*
	C1	Level C1 is characterised by a broad range of language, which allows fluent, spontaneous communication, as illustrated by the following examples: *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.* The discourse skills appearing at B2+ are more evident at C1, with an emphasis on more fluency, for example: *select a suitable phrase from a fluent repertoire of discourse functions to preface his/her 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.*
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): *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*. It is also apparent in relation to coherence/cohesion: *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*.

Level		Salient characteristics: interaction and production (CEFR section 3.6, simplified)
Independent user	**B2**	Level B2 represents a break with the content so far. Firstly, there is a focus on effective argument: *account for and sustain his/her 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*. Secondly, at this level one can hold one's own in social discourse: for example, *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.* Finally, there is a new degree of language awareness: *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.*
	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 quantities of information, for example: *provide concrete information required in an interview/consultation (for example, describe symptoms to a doctor) but does so with limited precision; explain why something is a problem; summarise and give his/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/her field with some confidence.*
	B1	Level B1 reflects the Threshold Level specification and is perhaps most categorised by two features. The first feature is the ability to maintain interaction and get across what you want to, for example: *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.* The second feature is the ability to cope flexibly with problems in everyday life, for example *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.*
Basic user	**A2+**	A2+ represents a strong A2 performance with more active participation in conversation given some assistance and certain limitations, for example: *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, for example: *express how he/she feels in simple terms; give an extended description of everyday aspects of his/her environment, for example, 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.*

Level		Salient characteristics: interaction and production (CEFR section 3.6, simplified)
Basic user	**A2**	Level A2 has the majority of descriptors stating social functions like *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.* Here too are to be found descriptors on getting out and about: *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*.
	A1	Level A1 is the lowest level of generative language use – the point at which the learner can *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*, rather than relying purely on a very finite rehearsed, lexically organised repertoire of situation-specific phrases.

Table 3.1: Salient characteristics of CEFR levels

		A1	A2	B1	B2	C1	C2
Understanding	**Listening**	I can recognise familiar words and very basic phrases concerning myself, my family and immediate concrete surroundings when people speak slowly and clearly.	I can understand phrases and the highest frequency vocabulary related to areas of most immediate personal relevance (for example, very basic personal and family information, shopping, local area, employment). I can catch the main point in short, clear, simple messages and announcements.	I can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure, etc. I can understand the main point of many radio or TV programmes on current affairs or topics of personal or professional interest when the delivery is relatively slow and clear.	I can understand extended speech and lectures and follow even complex lines of argument provided the topic is reasonably familiar. I can understand most TV news and current affairs programmes. I can understand the majority of films in standard dialect.	I can understand extended speech even when it is not clearly structured and when relationships are only implied and not signalled explicitly. I can understand television programmes and films without too much effort.	I have no difficulty in understanding any kind of spoken language, whether live or broadcast, even when delivered at fast native speed, provided I have some time to get familiar with the accent.
	Reading	I can understand familiar names, words and very simple sentences, for example on notices and posters or in catalogues.	I can read very short, simple texts. I can find specific, predictable information in simple everyday material such as advertisements, prospectuses, menus and timetables and I can understand short, simple personal letters.	I can understand texts that consist mainly of high frequency everyday or job-related language. I can understand the description of events, feelings and wishes in personal letters.	I can read articles and reports concerned with contemporary problems in which the writers adopt particular attitudes or viewpoints. I can understand contemporary literary prose.	I can understand long and complex factual and literary texts, appreciating distinctions of style. I can understand specialised articles and longer technical instructions, even when they do not relate to my field.	I can read with ease virtually all forms of the written language, including abstract, structurally or linguistically complex texts such as manuals, specialised articles and literary works.

		A1	A2	B1	B2	C1	C2
Speaking	**Spoken interaction**	I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I am trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics.	I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I cannot usually understand enough to keep the conversation going myself.	I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared into conversation on topics that are familiar, of personal interest or pertinent to everyday life (for example, family, hobbies, work, travel and current events).	I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views.	I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision and relate my contribution skilfully to those of other speakers.	I can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. I can express myself fluently and convey finer shades of meaning precisely. If I do have a problem I can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it.
	Spoken production	I can use simple phrases and sentences to describe where I live and people I know.	I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.	I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.	I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.	I can present a clear, smoothly flowing description or argument in a style appropriate to the context and with an effective logical structure which helps the recipient to notice and remember significant points.

		A1	A2	B1	B2	C1	C2
Writing	**Writing**	I can write a short, simple postcard, for example sending holiday greetings. I can fill in forms with personal details, for example entering my name, nationality and address on a hotel registration form.	I can write short, simple notes and messages relating to matters in areas of immediate needs. I can write a very simple personal letter, for example thanking someone for something.	I can write simple connected text on topics which are familiar or of personal interest. I can write personal letters describing experiences and impressions.	I can write clear, detailed text on a wide range of subjects related to my interests. I can write an essay or report, passing on information or giving reasons in support of or against a particular point of view. I can write letters highlighting the personal significance of events and experiences.	I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind.	I can write clear, smoothly flowing text in an appropriate style. I can write complex letters, reports or articles which present a case with an effective logical structure which helps the recipient to notice and remember significant points. I can write summaries and reviews of professional or literary works.

Table 3.2: Self-assessment grid

	Range	Accuracy	Fluency	Interaction	Coherence
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (for example, in forward planning, in monitoring others' reactions).	Can express him/herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using non-verbal and intonational cues apparently effortlessly. Can interweave his/her contribution into the joint discourse with fully natural turn-taking, referencing, allusion-making, etc.	Can create coherent and cohesive discourse making full and appropriate use of a variety of organisational patterns and a wide range of connectors and other cohesive devices.
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/ she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can select a suitable phrase from a readily available range of discourse functions to preface his/her 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.
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he 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.

	Range	Accuracy	Fluency	Interaction	Coherence
B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circum-locutions on topics such as family, hobbies and interests, work, travel and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2+					
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself under-stood 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".
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical struc-tures 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 3.3: Oral assessment criteria grid (CEFR Table 3)

Chapter 4: Specification

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 specification procedures outlined in this 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 (for example, Alderson et al. 1995: Chapter 2; Davidson and Lynch 1993 and 2002; Lynch and Davidson 1994 and 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. 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 of the Manual 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 sub-tests and subsections and the information and results made available to test users.

Forms A2 to 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 sub-tests (Form A3);
- the grading and standard setting procedures for different sub-tests (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 sub-tests (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.

Initial estimation of overall CEFR level		
☐ A1 ☐ ☐ A2 ☐	☐ B1 ☐ ☐ B2 ☐	☐ C1 ☐ ☐ C2 ☐
Short rationale, reference to documentation		

Form A8: Initial estimation of overall examination level

The detailed specification process is reported in Forms A9 to A22. 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 Manual. 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 (for example, 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 and 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 the 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 section 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 section is quite extensive. The forms in this section can be downloaded from www.coe.int/lang.

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

4.3.2. Content analysis grids

The CEFR content analysis grid for listening and reading and the CEFR content analysis grids for speaking and 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 sub-scales, and the associated tables in Appendix A of the Manual.

The CEFR content analysis grid for listening and reading is an online tool that 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 (for example, 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.

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, 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

The most recent copies of the grids plus illustrative samples using completed grids can be downloaded at www.coe.int/portfolio.

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 reader is referred to the Manual.

4.4. Procedures

The procedures involve consulting the CEFR, the appendices to the Manual and other sources referred to above, before systematically completing the series of forms provided in Appendix A of the Manual, which are also available for download from 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 co-

ordinator. 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 to A7);
   - Form A8 (initial estimation of overall examination level);
   - those forms – ranging from A9 to A22 – that are relevant to the examination or test tasks in question;
   - 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 to 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 sub-test and the relevant CEFR sub-scale.

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	✓	

Table 4.1: Forms and scales for communicative language activities

	Reception		Interaction		Production		Mediation	
	Listening	Reading	Spoken interaction	Written interaction	Spoken production	Written production	Spoken mediation	Written mediation
Linguistic competence								
General linguistic range	✓	✓	✓	✓	✓	✓	✓	✓
Vocabulary range	✓	✓	✓	✓	✓	✓	✓	✓
Vocabulary control			✓	✓	✓	✓	✓	✓
Grammatical accuracy			✓	✓	✓	✓	✓	✓
Phonological control			✓		✓		✓	
Orthographic control				✓		✓		✓
Sociolinguistic competence								
Sociolinguistic appropriateness	✓	✓	✓	✓	✓	✓	✓	✓
Pragmatic competence								
Flexibility			✓	✓			✓	✓
Turn-taking			✓					
Thematic development	✓	✓		✓	✓	✓	✓	✓
Cohesion and coherence	✓	✓			✓	✓	✓	✓
Spoken fluency			✓		✓		✓	
Propositional precision	✓	✓			✓	✓	✓	✓

	Reception		Interaction		Production		Mediation	
Strategic competence								
Identifying cues/inferring	✓	✓					✓	✓
Turn-taking (repeated)			✓					
Co-operating			✓	✓				
Asking for clarification			✓	✓				
Planning					✓	✓		✓
Compensating			✓	✓	✓	✓	✓	✓
Monitoring and repair			✓	✓	✓	✓	✓	✓

Table 4.2: CEFR scales for aspects of communicative language competence

5. Communicative language competence: next, the forms for aspects of communicative language competence should be completed (Forms A19 to 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.

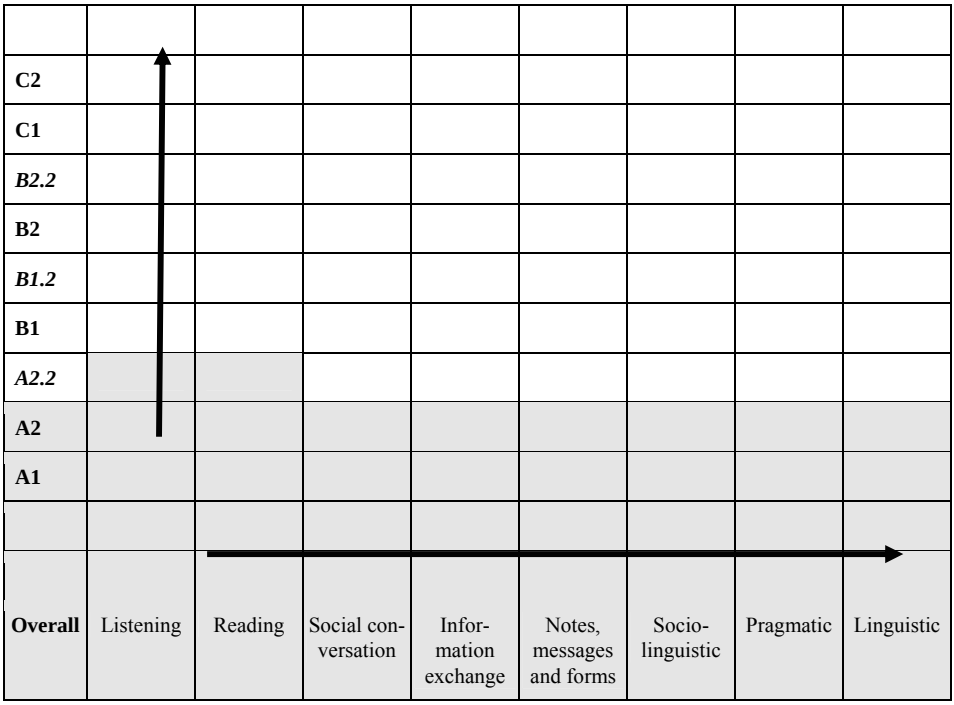
The forms are provided in the following order:

- 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 the 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 sub-scales for communicative language activities and for aspects of language competence (see the example of a completed Form A23 below).



Form A23: Graphic profile of the relationship of the examination to CEFR levels (example)

On the chart, the Y-axis (vertical) 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 shown above.

The labelling of the columns on Form A23 will not necessarily be the same as the names given to the sub-tests of the examination. Some columns may coincide with sub-tests, but other columns may also be added. For example, the examination might not have a separate sub-test 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 the 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		
☐ A1 ☐ ☐ A2 ☐	☐ B1 ☐ ☐ B2 ☐	☐ C1 ☐ ☐ C2 ☐
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.		

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

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.

The way that levels are assigned to test takers falls roughly into 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 test (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, assessing performance, 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 score". A cut score of 30, for example, says that a numerical score of 30 or more indicates performance at a particular level (for example, B1) or higher, while a lower score points to a level lower than the level of the cut score (here: B1). The process to arrive at a cut score is commonly referred to as standard setting. In the case of receptive skills (reading and listening) or underlying competences (grammar, vocabulary), cut scores need to be decided upon.

Both benchmarking and standard setting are procedures that 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.

5.2. The need for training

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.

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.

The appointed co-ordinator must be properly qualified and appropriately prepared to conduct the training.

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. This order is recommended as the most effective, but it will of course be modified according to the needs and constraints of the context.

Once training is completed, and common agreement on the assessment of illustrative samples is considered adequate, 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 co-ordinator is responsible for:

- the rationale to be followed, based upon the 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. Some 12 to 15 judges can be considered a minimum, and 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;
  - co-ordinating 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:
  - CEFR descriptors;
  - CEFR tables and rating instruments;
  - selection of illustrative CEFR performance samples and tasks;
  - selection of local performance samples and/or local test items − reporting forms and documents to record information on the sessions.
- 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 (for example, 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 co-ordinator 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 ongoing monitoring, dissemination and follow-up actions.

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 co-ordinator(s) is (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;
- 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 co-ordinators 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 co-ordinators 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 co-ordinator. 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 co-ordinator: (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 into 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.

The co-ordinator will need to calculate the percentage of participants who agree on the different ratings, or inter-rater correlation coefficients. The co-ordinator 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.

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 examples of spoken performance and continues (if appropriate) with illustrative scripts of written presentations. 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).

For the assessment of writing, it is also important to see samples of both written interaction (for example, notes and letters) and written production (for example, descriptions, stories and 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. 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 to make a judgment on 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.

The session is organised in three phases:

- *Phase 1: illustration* – the session starts with two or three CEFR illustrative performances that the co-ordinator uses to introduce the levels. The co-ordinator plays the sample and then invites participants to discuss the performance with neighbours. At an appropriate point the co-ordinator should bring the group together, and elicit from the group the way in which the performance illustrates the level, and why it is not the level described above or below;
- *Phase 2: practice* – in a second phase the role of the co-ordinator is to help individuals see if they are still tending to be too strict or too lenient. If voting is on paper, the co-ordinator will use the collation form to record the ratings onto a transparency or on a flip chart. Throughout this phase, the co-ordinator 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, 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.

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

5.5.2. Written performance

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

- Phase 1: illustration – the session starts with two or three written performances that the co-ordinator uses to illustrate the levels. For each sample, at a certain point the co-ordinator should bring the group together, and elicit from the group the way in which the performance illustrates the level, and why it is not the level above or below;
- Phase 2: practice – in this second phase, the role of the co-ordinator is to help individuals see if they are still tending to be too strict or too lenient. If voting is on paper, the co-ordinator will use a collation form to record the ratings onto a transparency. Throughout this phase, the co-ordinator 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 into 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.

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

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 (for example, 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 sub-scales of descriptors for specific areas.
2. Identifying the content relevance of the tasks analysed in terms of construct coverage vis-à-vis CEFR levels and scales.
3. Estimating the level each task and item represents in terms of the relevant CEFR descriptors.
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 co-ordinator 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 (for example, "Reception", "Overall reading comprehension" and "Overall listening comprehension"), and also specific scales that describe different receptive language activities (for example, "Listening as a member of an audience") and strategies ("Identifying cues and inferring").

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;
- Phase 2: controlled practice – once the illustration phase and the initial discussion have taken place, 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;
- 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.

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 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 the Manual (or from other literature on standard setting) will influence the procedures to follow. Users of the 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 co-ordinator 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.

5.7.1. Samples required

It is worth investing time and energy in collecting a representative set of local samples of high quality.

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 features of the samples for benchmarking using a tool like the CEFR grids for writing and speaking tasks.

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.

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;
- 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.

If agreement is not reached, the co-ordinator should discuss with participants why they are having such a problem in contrast to their success with the illustrative samples. The co-ordinator will need to make a judgment on the reason for the problem, and take appropriate action.

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 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.

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 – on 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 co-ordinator 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 benchmarking;
- 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 and 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 standard CEFR-based rating grids or develop other more sector-specific CEFR-based instruments;
- how to publish and/or disseminate the results of the standardisation process in the field;
- how to ensure good "local" dissemination and follow-up.

Chapter 6: Standard setting procedures

6.1. Introduction

The basic output from taking a test is a numerical score. In the case of tests used for the receptive skills of reading and listening, this score is usually the number of correct responses to each of the items in the test. In the case of productive skills (writing and speaking), 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 (for example, 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" (that is, B1 or lower).

It is possible that more than one standard has 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. It must also be noted that for the test to be able to distinguish between these three levels, the test must contain a sufficient number of items or tasks related to these (three) levels.

To avoid confusion the levels (the "adjacent categories") and the cut scores (the borders 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 one of the most complicated procedures in language assessment. It has been said that it is a mix of artistic, political and cultural ingredients. Policy makers and examiners should make sure that they carefully select and train those involved in standard setting and that the procedures that are to be followed are well thought out and closely adhered to.

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 these Highlights 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.

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 or tasks 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 intended primarily to detect and eliminate misunderstanding of the instructions. It is good practice to let panel members discuss this information in small

groups. After the second round, a different kind of information is usually given: the consequences of the panel's judgments. This is done 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.

Certainly in the case of high stakes examinations it is wholesome 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 of finding a rational and reasonable compromise between two highly different group decisions, and this may be sufficient to organise a fourth round of judgments.

For almost all standard setting procedures described in the testing 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. To make a judgment on the validity and efficiency of any procedure that is applied in any project, 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. This is all the more pressing now that increasingly students need to have proof of their language proficiency in a foreign language if they wish to study abroad. Certificates of language proficiency that have not been sufficiently validated may not be accepted elsewhere. Standard setting procedures will need to be documented for others to be able to judge the quality of the assessment.

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. As standard setting contexts vary, this chapter does not advocate the use of any single one of them.

Sometimes standard setting methods are divided into test-centred and examinee-centred methods. Some methods of the latter type use direct judgment of test takers by a rater who knows them well. Other methods ask holistic judgments on the "work" from a sample of students: their score on a test or examination, an essay or 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 test-centred methods, panel members may be asked to make a judgment on each item in a test. Such 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. However, methods have been developed where the distinction between examinee-centred and test-centred methods is less clear. In these methods, statistical 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.

In this chapter, we will now briefly discuss four examples of methods of standard setting: (1) one examinee centred, (2) one test centred that can be applied (with or) without any empirical test taking data and (3) a fourth example where panel members use a summary of empirical data.

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 the final chapter of these Highlights.

6.3. The Body of Work method: examinee centred

The Body of Work method is perhaps the most suitable one for handling holistic judgments, although it can be used with any mixture of item types and tasks. Here is a brief list of what is needed to apply the method:

- 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);

- 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 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 scores, one has to apply 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 such as taking the midpoint between averages may lead to serious biases.

6.4. The Tucker-Angoff method: test centred

This is one of the most widely used standard setting methods and many variations of it have been proposed.

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". Where a standard has to be set, for example, for CEFR Level B1, a borderline 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 borderline person would give a correct answer. As a next step in the procedure, the probabilities are summed across items for every rater. One method, often applied in practice, is just to take the average of the sums for each rater, and to consider the averages of all raters as the standard. This is illustrated in the following table.

	Rater 1	Rater 2	…	Rater 15	
Item 1	25%	32%	…	35%	
Item 2	48%	55%	…	45%	
Item 3	33%	38%	…	28%	
…	…	…	…	…	
Item 49	21%	30%	…	35%	
Item 50	72%	80%	…	90%	
Average	**65%**	**72%**	**…**	**78%**	**Standard 75%**

Table 6.1: Basic data using the Tucker-Angoff method: percentage chance of correct answer by a borderline person

To summarise: three components are essential in the procedure: the concept of the borderline 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.

6.5. The Basket method: test centred

This method requires a comparison of the demands of an item in terms of the "can do" descriptors of the CEFR. The basic question asked of the panel members 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? In other words: what is the minimum CEFR level that is required to give a correct response to each task or question in a test? The panel members are asked to put each item in a "basket" (this may be a basket, tray or simply a pile) corresponding to one of the CEFR levels relevant to the test in question. 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.

The method to convert judgments to cut scores is 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 anyone who is at Level B1 or higher. This number, the minimum requirement, is interpreted as the cut 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 even 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 show that the Basket method tends to produce lower (more lenient) standards than other methods.

6.6. The Bookmark method: test centred

This method is applicable for binary items (such as multiple-choice items, yielding either a right or a wrong answer) and for constructed responses (CR) or tasks (yielding a partial credit in the range 0 to 2 or 0 to 3, for example), which are more likely to occur with productive skills. Panel members use the concept of borderline person. For multiple standards (as, for example, A1/A2, A2/B1 and B1/B2 for the same test), the procedure has to be repeated for each standard.

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. Panel members have to decide for an item whether a borderline person (at the given standard) masters the item or not. If a person masters an item, one can expect that he or 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 two thirds, although some authors prefer to set it at 50% and others at 80%. In the standard setting literature this mastery criterion is referred to as the Response Probability (RP). For RP = two thirds, 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.

The panel members are instructed to start with the lowest standard (for example, 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 or 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 booklet with the items ordered in difficulty.

In each round, each panel member indicates his or her provisional standard in a table like the one displayed in Figure 6.2 for the case in which 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 6.2. The page numbers 11/12 for the standard A1/A2 mean that (in the view of the participant) a borderline person at 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 6.2: Panel member recording form for Bookmark method

6.7. 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 sub-tests addressing a different skill. 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 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 or she 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.

6.8. 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 (for example, for the following year) by applying a technique called test equating. 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 score for this year is 37.

To apply equating techniques, there are two issues that call for attention. 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.

Another 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.

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.

6.9. 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 the 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. Also test-centred methods, such as the Tucker-Angkoff method, can be used.

6.10. Conclusion

This chapter has given an overview of a number of standard setting procedures, but it pretends in no way to be exhaustive. A more comprehensive overview can be found in the Manual and in section B of the Reference Supplement. The emphasis 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 focused 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? Similar questions are also 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 or herself across the various tasks he or she 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 answers 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 the next chapter.

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 their context;
- whether to adopt an examinee-centred method or a test-centred method;
- whether to adopt a method judging the difficulty of individual items (Basket method) or a method judging the cut score on the reporting scale for the trial test (for example, 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[i] feasible?
- whether difficulty estimates will be available to inform the standard setting process;
- 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.

Chapter 7: Validation

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 these Highlights focus 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 to be discussed will also be focused on the linking process proper.

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.

7.2. Prerequisites: 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 (for example, the Body of Work method), whilst others involve global judgments about where to place the cut-off between levels on a test, informed by a lot of psychometric information (for example, the Bookmark method). Please refer to Chapter 6 for more information about this.

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. It is therefore 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 of 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 (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) 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.

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, 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.

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, 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, that is, in Year zero.

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 (for example, 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. In estimating the reliability, the following should 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. The Manual gives suggestions on how to best estimate the reliability of (heterogeneous) tests.

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 (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.

We will briefly discuss here 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. The Manual gives detailed suggestions on what one possibly could do to avoid or control the pretest effect.

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: 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.

Other aspects for demonstrating procedural validity 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.

Practicability: although some procedures are quite complicated, the preparation must be practical, such that:

- the standard setting method can be implemented without great difficulty;
- data analysis can be addressed without laborious computations. The preparatory work must be completed well before the sessions;
- the procedures are seen as credible and interpretable by non-technicians.

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.

- Removing items or tasks is an even more delicate problem. If linking to the CEFR is the main purpose of the examination, for example, 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.

Some other issues relate to consistency and accuracy in standard setting procedures:

- 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.

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 these Highlights to the 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.

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, that is, 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 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 the Manual some examples of external validation procedures are discussed, and arguments as to their limits and persuasiveness (or lack of it) are put forward. A general remark is in order here. 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, that is, part of their variance is unwanted or irrelevant for showing the validity of a test procedure, such as the results of a standard setting.

7.6. 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.

Thus, it is to be hoped that many standard setting endeavours, under way or planned in the future, drawing on the information provided in the 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 score;
- how they will make their detailed findings available to professional colleagues.

References and further reading

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 referential.* 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. DGIV/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.

Council of Europe (2009): Relating Language Examinations to the Common European Framework of Reference for Languages: learning, teaching, assessment (CEFR): a Manual. 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*. Jyvaskyla, Finland: University of Jyvaskyla, 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 (4th 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 (2002) Developing descriptor scales of language proficiency for the CEF common reference levels. In: Council of Europe (2002): *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Case Studies.* Strasbourg: Council of Europe Publishing.

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: Ruegger 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.

Glossary related to linking processes

Accreditation: The granting of recognition of a test or an examination, usually by an official body such as a government department, examinations board, etc.

Aggregate: To combine two or more related scores into one total score.

Alignment: The process of linking content and performance standards to assessment, instruction, and learning in classrooms. One typical alignment strategy is the step-by-step development of (a) content standards, (b) performance standards, (c) assessments, and (d) instruction for classroom learning.

Assessment grid: A set of assessment criteria presented in a tabular format.

Benchmark: A detailed, validated description of a specific level of student performance expected of students at particular ages, grades, or levels in their development. Benchmarks are often represented by samples of student work.

Bias: A test or item can be considered to be biased if one particular section of the candidate population is advantaged or disadvantaged by some feature of the test or item which is not relevant to what is being measured. Sources of bias may be connected with gender, age, culture, etc.

Borderline performance: A level of knowledge and skills that is just barely acceptable for entry into a performance level (e.g., B2-level).

Classical test theory (CTT): CTT refers to a body of statistical models for test data. The basic notion of CTT is that the observed score *X* obtained when a person *p* is administered form *f* of test X, is the sum of a true-score component and an error component. See also Item Response Theory (IRT).

Compensatory strategy: A strategy that allows a high level of competence in one of the components of the assessment to compensate for a low level of the other components.

Conjunctive strategy: A strategy that requires attaining some predefined minimum level of competence for each one of the separate components to allow the final, summarized result to be judged as acceptable (sufficient).

Construct: A hypothesized ability or mental trait which cannot necessarily be directly observed or measured; for example, in language testing, listening ability.

Content standards: Broadly stated expectations of what students should know and be able to do in particular subjects and grade levels.

Content validity: A test is said to have content validity if the items or tasks of which it is made up constitute a representative sample of items or tasks for the area of knowledge or ability to be tested.

Constructed response (CR): A form of written response to a test item that involves active production, rather than just choosing from a number of options.

Cross-language standard setting: A method intended to verify that examinations in different languages are linked in a comparable way to the common standards.

Cross validation: The application of a scoring system derived in one sample to a different sample drawn from the same population.

Cut score (cut-off score): The minimum score a candidate has to achieve in order to be assigned to a given level or grade in a test or an examination.

Decision validity: The degree to which classification decisions will be identical in repeated testing with the same examinees.

Direct test: A test which measures the productive skills of speaking or writing, in which performance of the skills itself is directly measured.

Examinee-centred method: A standard setting method in which someone who knows examinees well provides a holistic assessment of the level of their language proficiency, for example a CEFR level.

External validation: Collecting evidence from independent sources which corroborate the results and conclusions of procedures used.

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.

High stakes testing: A form of testing with important consequences for test takers.

Holistic judgment: Evaluating student work in which the score is based on an overall judgment of student performance rather than on specific separate criteria.

Indirect test: A test or task which attempts to measure the abilities underlying a language skill, rather than testing performance of the skill itself. An example is testing writing ability by requiring the candidate to mark structures used incorrectly in a text.

Internal validation: The process of finding out the accuracy and the consistency of an assessment based on the judgments in the test.

Illustrative samples (benchmarked samples): Examples of student performance that have been validated to represent a certain level of performance.

Inter-rater reliability: The degree to which different raters agree in their assessment of candidates' performance.

Intra-rater reliability: The degree to which the same rater judges the same performance similarly on different occasions.

Item Response Theory (IRT): It is a theoretical approach to relating student ability to test data; it focuses on the items as opposed to classical test theory (CTT) focusing on the test scores.

Item difficulty: In classical test theory, the difficulty of an item is the proportion of candidates responding to it correctly. In IRT it is an estimate of a difficulty of an item calculated independently of the population.

Judge: Someone who assigns a score to a candidate's performance in a test, using judgment to do so.

KR20: A measure of internal consistency developed by Kuder and Richardson and used to estimate test reliability.

Logistic regression: A statistical technique that yields a formula for translating one of more pieces of information (e.g., a person's test score) into the estimated probability of a specified event (e.g., a sample of the student's work being judged as proficient).

Low stakes testing: A form of testing with less important consequences for test takers.

The Manual: The document produced by the Council of Europe to provide guidance to link tests and examinations to the CEFR.

Mastery: The indication that the student has met a set of criteria, defined in terms of well-defined domains of skills or knowledge.

Panel: A group of judges.

Panellist: A member of a group of judges.

Performance standards: Explicit definitions of what students must do to demonstrate proficiency at a specific level on the content standards.

Performance level descriptors (PLD): Descriptions of standards the students should have reached. The level descriptions in the CEFR are examples of PLDs.

Piloting: A preliminary study through which test developers try out tasks on a limited number of subjects in order to locate problems before launching a full-scale trial.

Pretesting: A stage in the development of test materials at which items are tried out with representative samples from the target population in order to determine their difficulty. Following statistical analysis, those items that are considered satisfactory can be used in live tests.

Procedural validation: Collecting evidence that proper procedures have been used during the different stages of standard setting.

Rater: A person who evaluates or judges student performance on an assessment against specific criteria.

Rating: The process of assigning a score to performance in a test through the exercise of judgement.

Response probability (RP): In standard setting it is a mastery criterion. In many tests, it is set at two-thirds of the maximum score, although some authors prefer to set it at 50%, while others at 80%.

Specification: A stage in the linking process that 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.

Test-centred methods: A set of methods where judges estimate, for example at what level a test taker can be expected to respond correctly to a set of items.

Test equating: The process of comparing the difficulty of two or more forms of a test, in order to establish their equivalence.

Test specifications: A description of the characteristics of an examination, including what is tested, how it is tested, details such as number and length of papers, item types used, etc.

Transparency: Implies openness, communication and accountability. It is an extension of the meaning used in the physical sense (cf. a transparent object can be seen through).

Sales agents for publications of the Council of Europe
Agents de vente des publications du Conseil de l'Europe

BELGIUM/BELGIQUE
La Librairie Européenne -
The European Bookshop
Rue de l'Orme, 1
BE-1040 BRUXELLES
Tel.: +32 (0)2 231 04 35
Fax: +32 (0)2 735 08 60
E-mail: info@libeurop.eu
http://www.libeurop.be

Jean De Lannoy/DL Services
Avenue du Roi 202 Koningslaan
BE-1190 BRUXELLES
Tel.: +32 (0)2 538 43 08
Fax: +32 (0)2 538 08 41
E-mail: jean.de.lannoy@dl-servi.com
http://www.jean-de-lannoy.be

BOSNIA AND HERZEGOVINA/
BOSNIE-HERZÉGOVINE
Robert's Plus d.o.o.
Marka Maruliça 2/V
BA-71000, SARAJEVO
Tel.: + 387 33 640 818
Fax: + 387 33 640 818
E-mail: robertsplus@bih.net.ba

CANADA
Renouf Publishing Co. Ltd.
1-5369 Canotek Road
CA-OTTAWA, Ontario K1J 9J3
Tel.: +1 613 745 2665
Fax: +1 613 745 7660
Toll-Free Tel.: (866) 767-6766
E-mail: order.dept@renoufbooks.com
http://www.renoufbooks.com

CROATIA/CROATIE
Robert's Plus d.o.o.
Marasovièeva 67
HR-21000, SPLIT
Tel.: + 385 21 315 800, 801, 802, 803
Fax: + 385 21 315 804
E-mail: robertsplus@robertsplus.hr

CZECH REPUBLIC/
RÉPUBLIQUE TCHÈQUE
Suweco CZ, s.r.o.
Klecakova 347
CZ-180 21 PRAHA 9
Tel.: +420 2 424 59 204
Fax: +420 2 848 21 646
E-mail: import@suweco.cz
http://www.suweco.cz

DENMARK/DANEMARK
GAD
Vimmelskaftet 32
DK-1161 KØBENHAVN K
Tel.: +45 77 66 60 00
Fax: +45 77 66 60 01
E-mail: gad@gad.dk
http://www.gad.dk

FINLAND/FINLANDE
Akateeminen Kirjakauppa
PO Box 128
Keskuskatu 1
FI-00100 HELSINKI
Tel.: +358 (0)9 121 4430
Fax: +358 (0)9 121 4242
E-mail: akatilaus@akateeminen.com
http://www.akateeminen.com

FRANCE
La Documentation française
(diffusion/distribution France entière)
124, rue Henri Barbusse
FR-93308 AUBERVILLIERS CEDEX
Tél.: +33 (0)1 40 15 70 00
Fax: +33 (0)1 40 15 68 00
E-mail: commande@ladocumentationfrancaise.fr
http://www.ladocumentationfrancaise.fr

Librairie Kléber
1 rue des Francs Bourgeois
FR-67000 STRASBOURG
Tel.: +33 (0)3 88 15 78 88
Fax: +33 (0)3 88 15 78 80
E-mail: librairie-kleber@coe.int
http://www.librairie-kleber.com

GERMANY/ALLEMAGNE
AUSTRIA/AUTRICHE
UNO Verlag GmbH
August-Bebel-Allee 6
DE-53175 BONN
Tel.: +49 (0)228 94 90 20
Fax: +49 (0)228 94 90 222
E-mail: bestellung@uno-verlag.de
http://www.uno-verlag.de

GREECE/GRÈCE
Librairie Kauffmann s.a.
Stadiou 28
GR-105 64 ATHINAI
Tel.: +30 210 32 55 321
Fax.: +30 210 32 30 320
E-mail: ord@otenet.gr
http://www.kauffmann.gr

HUNGARY/HONGRIE
Euro Info Service
Pannónia u. 58.
PF. 1039
HU-1136 BUDAPEST
Tel.: +36 1 329 2170
Fax: +36 1 349 2053
E-mail: euroinfo@euroinfo.hu
http://www.euroinfo.hu

ITALY/ITALIE
Licosa SpA
Via Duca di Calabria, 1/1
IT-50125 FIRENZE
Tel.: +39 0556 483215
Fax: +39 0556 41257
E-mail: licosa@licosa.com
http://www.licosa.com

NORWAY/NORVÈGE
Akademika
Postboks 84 Blindern
NO-0314 OSLO
Tel.: +47 2 218 8100
Fax: +47 2 218 8103
E-mail: support@akademika.no
http://www.akademika.no

POLAND/POLOGNE
Ars Polona JSC
25 Obroncow Street
PL-03-933 WARSZAWA
Tel.: +48 (0)22 509 86 00
Fax: +48 (0)22 509 86 10
E-mail: arspolona@arspolona.com.pl
http://www.arspolona.com.pl

PORTUGAL
Livraria Portugal
(Dias & Andrade, Lda.)
Rua do Carmo, 70
PT-1200-094 LISBOA
Tel.: +351 21 347 42 82 / 85
Fax: +351 21 347 02 64
E-mail: info@livrariaportugal.pt
http://www.livrariaportugal.pt

RUSSIAN FEDERATION/
FÉDÉRATION DE RUSSIE
Ves Mir
17b, Butlerova ul.
RU-101000 MOSCOW
Tel.: +7 495 739 0971
Fax: +7 495 739 0971
E-mail: orders@vesmirbooks.ru
http://www.vesmirbooks.ru

SPAIN/ESPAGNE
Díaz de Santos Barcelona
C/ Balmes, 417-419
ES-08022 BARCELONA
Tel.: +34 93 212 86 47
Fax: +34 93 211 49 91
E-mail: david@diazdesantos.es
http://www.diazdesantos.es

Díaz de Santos Madrid
C/Albasanz, 2
ES-28037 MADRID
Tel.: +34 91 743 48 90
Fax: +34 91 743 40 23
E-mail: jpinilla@diazdesantos.es
http://www.diazdesantos.es

SWITZERLAND/SUISSE
Planetis Sàrl
16 chemin des Pins
CH-1273 ARZIER
Tel.: +41 22 366 51 77
Fax: +41 22 366 51 78
E-mail: info@planetis.ch

UNITED KINGDOM/ROYAUME-UNI
The Stationery Office Ltd
PO Box 29
GB-NORWICH NR3 1GN
Tel.: +44 (0)870 600 5522
Fax: +44 (0)870 600 5533
E-mail: book.enquiries@tso.co.uk
http://www.tsoshop.co.uk

UNITED STATES and CANADA/
ÉTATS-UNIS et CANADA
Manhattan Publishing Co
2036 Albany Post Road
USA-10520 CROTON ON HUDSON, NY
Tel.: +1 914 271 5194
Fax: +1 914 271 5886
E-mail: coe@manhattanpublishing.coe
http://www.manhattanpublishing.com

Council of Europe Publishing/Editions du Conseil de l'Europe
FR-67075 STRASBOURG Cedex
Tel.: +33 (0)3 88 41 25 81 – Fax: +33 (0)3 88 41 39 10 – E-mail: publishing@coe.int – Website: http://book.coe.int