5 Assessing CEFR level

Section 1: Essentials

5.1.1 The CEFR and assessment

CEFR Section 9.3 discusses assessment through a series of contrasts, for example assessment of proficiency (= assessment of level) as opposed to assessment of achievement (= assessment of learning); criterion-referenced assessment (= in relation to a standard) as opposed to norm-referenced assessment (= in relation to one's peers); self-assessment as opposed to assessment by others – etc. However, in practice when people talk about CEFR assessment, they are generally talking about the assessment of CEFR level, so this chapter confines itself to that topic.

Before the advent of the CEFR communication about assessment results across even the smallest barriers was difficult. A teacher, school or examination body would carry out a test and report a result like '19', '4.5', '516', 'B', 'Good', etc. Even when the assessment was genuinely criterion-referenced rather than norm-referenced, each test reported a result in its own way. In most cases there was little or no definition at all of what this grade or score meant in terms of ability to use the language; communities of test users had to develop an interpretation. From a practical point of view it was literally a Tower of Babel. From a theoretical point of view there was, in general, a reluctance to engage with the fundamental problem in language test validity: demonstrating that a specific result has a particular meaning in terms of real world language use.

Helping to address this issue is one of the aims of the CEFR. The CEFR suggests (Council of Europe 2001:178) that the CEFR descriptors can be of help for:

– Specification of the content of tests and examinations:	*What is assessed*
– Stating the criteria to determine the attainment of a learning objective:	*How performance is interpreted*
– Describing the levels of proficiency in existing tests and examinations, thus enabling comparisons to be made across different systems of qualifications:	*How comparisons can be made*

A claim to operate 'CEFR assessment' suggests the existence of transparent and coherent assessment procedures which, in addition to being valid in relation to the context and curriculum concerned, report results in terms of CEFR Common Reference Levels. This entails well-targeted, well-constructed assessments that use task types familiar to the students in order to assess their success at meeting the objectives in the CEFR-based curriculum. As language teachers we do not need to pretend to be examination institutes. Examination institutes have considerations, particularly the standardisation of item types for machine marking, that are not relevant to classroom assessment. We can also take confidence from the fact that the same features that make a valid classroom task will also help to make a valid assessment task. Alderson, Clapham and Wall (1995) point the way when they say:

> We believe, then, that there is no important difference between writing a test item and writing a learning task or exercise. Thus whatever qualities are needed by the designer of an exercise are also needed by test writers. Perhaps more importantly, the sources of inspiration for exercises can and should also be used for test writing; test writers, in other words, can and should be as imaginative as possible when thinking about their item types and one very useful source of ideas is textbooks and other learning materials (Alderson et al 1995:41–42).

5.1.2 Validity

However, we do need to seriously consider the validity of our approach. A good assessment approach must be valid for the context and learners concerned. As the CEFR puts it:

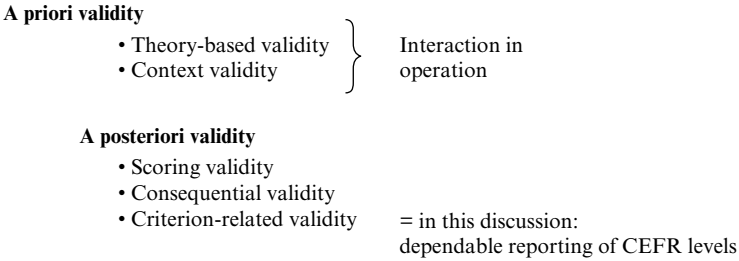
> *Validity* is the concept with which the Framework is concerned. A test or assessment procedure can be said to have validity to the degree that it can be demonstrated that what is actually assessed (the construct) is what, in the context concerned, *should* be assessed, and that the information gained is an accurate representation of the proficiency of the candidates(s) concerned (Council of Europe 2001:177).

Fortunately, newer concepts of validity for language assessment have more in common with good teaching than used to be the case. Traditionally validity, like reliability, was seen as something to be investigated through data collection and statistics *after* using the test. This is something that we as teachers do not usually have time for. Weir (2005a) has developed a new way of looking at the concept of validity that makes it far more relevant, attractive and practical for those of us working in schools. Validity is primarily a

question of getting it right in the first place, and that is achieved by taking account of *theory* (what we know about the way things happen, the processes people go through when, for example, listening), *context* (what sort of people are doing this, where and under what conditions) and the *interaction* involved between them – because the processes involved and strategies adopted always depend on context. But how do you satisfy yourself – and stakeholders – that you did get it right? By investigating the various aspects of *evidence for validity* that have been developed over the years and have been used to claim construct validity. The latter is considered in Section 5.2.1.

Weir provides several comprehensive diagrams that explain the way all aspects of the assessment process are covered by his scheme, but the essence can be summarised very simply, as in Figure 5.1.

Figure 5.1 Weir's validity model



5.1.2.1 A priori validity

Weir's concept of *a priori validation* was developed as follows. Firstly, Bachman and Palmer (1996:25–29) supplemented the familiar concept of *authenticity* (the closeness of the task to what the candidates would do in the real world) with that of *interactivity* (the extent to which the processes involved are relevant to the processes that would be needed in the real world). Then Weir (2005a:137) renamed these two aspects *situational authenticity* and *interactional authenticity* respectively, relating the former to contextual validity and the latter to theory-based validity. An important point to bear in mind here is that the closer that the assessment tasks are to the learning tasks, the tasks described in the descriptors, and the real world tasks referred to by descriptors, the better. Such directness provides *situational authenticity* and *contextual validity* as well as transparency (*we know why we are doing this*) and coherence (*it's like in the course*). The CEFR's action-oriented approach encourages this validity. On the other hand, the more indirect the tasks are, e.g. gapped dialogues for speaking, sentence completion for writing, cloze tests for reading, the more difficult it is to generalise from scores on them to the learner's real world ability. Hence it is also more difficult to relate scores

on such indirect tasks to CEFR levels, which are defined in terms of real world ability.

Theory-based validity suggests that speaking tasks should generate co-constructed discourse with claiming, maintaining and yielding the turn, back-channelling, self-repair, collaborative strategies to ensure mutual understanding, compensatory strategies, monitoring and repair. Assessment of writing should allow the redrafting that reflects the natural writing process; a portfolio approach is clearly superior in a school context. In relation to listening, tasks should require processing in real time of authentic texts with the linguistic characteristics of typical spoken language. In relation to reading, tasks should reflect the kind of discourse processing that the theory of reading suggests happens in reality. All these points apply to teaching and testing equally.

Another major language tester to make a similar point is Bachman (2002:464). He states that since teaching/learning tasks tend to be closer to 'real world' language use, the fact that many test tasks are often very different from both teaching/learning tasks and real world tasks raises serious questions about their validity. This doesn't of course mean that teaching tasks can be automatically used for testing. But the fundamental issue is that all the features that make good *a priori validity* for teaching are equally valid for testing.

5.1.2.2 A posteriori validity

All three categories of a posteriori validity shown in Figure 5.1 are also just as relevant in relation to teacher assessment as they are to examinations. If the scoring (even if this is just a comment) is based on an arbitrary personal system rather than transparent criteria related to the construct in question, this cannot be valid (*scoring validity*). If the form of an assessment causes learners to invest time in non-productive preparatory activity, demotivates the class, biases against some individuals or has unfair results, that cannot be valid (*consequential validity*). Finally if the relationship to the CEFR is just wrong, if a level from a previous system has just been relabelled without investigation, or if the salient features of B1+ are used to assess B2, then the assessment might be valid in its own terms – but any claimed relationship to the CEFR as the chosen external criterion cannot be valid (*criterion-related validity*). Schools and teachers should use the same principles as testing agencies to build an argument for the validity of their practices. No one would expect a school to invest in collecting the same degree of *a posteriori validity* evidence as an examining board. This is recognised in, for example, the Manual for relating examinations to the CEFR (Council of Europe 2009). It suggests that schools can also exploit the recommended CEFR-linking procedures *specification, standardisation* and *external validation*, but accepts that they would do so to a lower degree of rigour. After all, an examination

institute knows nothing about a student who takes a test; the only information is what is collected on the spot in the test. By contrast, in a school we have lots of evidence of a learner's development over time. We have lots of collateral information to take into account in giving grades in addition to the result on a single test. This can of course be done well or badly, but a school can introduce *moderation procedures* to ensure that it is done well, as discussed later in Section 5.2.2.

5.1.3 Specification

To ensure comparability and fairness, it is important that the form of assessment is specified in what is sometimes called a blueprint. One good school-based example of a blueprint for CEFR-based tests was developed by Ángeles Ortega for the Spanish state language schools for adults (EOI) in Ceuta and Melilla. The blueprint for reading tests at B1 is given in Figure 5.2 and the one for writing tests at the same level is shown in Figure 5.3. Both define the format, what the learner has to do, the time allowed and the marks available. Reading tests have four complementary tasks: a matching task (5 points) followed by a text with multiple choice (5 marks), an information transfer exercise (10 marks), and finally a True/False task (10 marks). Writing tests involve two complementary types of text: personal correspondence or report and a note or announcement. The candidates have 50 minutes for the reading test and 55 minutes for the writing test.

There should then be a more detailed specification of the type of texts and tasks involved and the conditions (time, support) under which the assessment is to be carried out. For CEFR-based assessment, the CEFR itself is a logical place to start. However, the CEFR is only a point of orientation. Firstly, a detailed specification is by definition context-bound since the test and texts concerned should consider the interests, habits and cognitive abilities of the learners in the context (*situational validity*). For those who would like to read more about the specification process, Davidson and Lynch (2002) provide a very readable account. Secondly, in relation to defining the construct (*theory-based validity*) for different skills we must take account of the processes involved. Here the following books can be highly recommended: for listening Buck (2001) and Geranpayeh and Taylor (Eds) (2013), for reading Alderson (2000) and Khalifa and Weir (2009), for speaking Luoma (2004) and Taylor (Ed) (2011), and Shaw and Weir (2007 and for writing Weigle (2002)).

In addition, in their manual for test development for use with the CEFR, ALTE (2011) gives very good advice on the organisation and sequencing of test development. The text is short and clear in order to make it accessible to 'novice language testers'. The main points made are that tasks and tests need to be developed in an iterative cycle of feedback and improvement, not

Figure 5.2 EOI blueprint for a B1 reading test

PART ONE

FORMAT	5 texts of approximately 60 words each (300 words in total) and 8 titles.
PROCEDURE	Choose from the 8 titles presented one title for each text and copy it to the corresponding place on the answer sheet.
MARKS	● 1 mark for each correctly titled text ● 5 marks total

PART TWO

FORMAT	A text of approximately 450 words and 5 multiple choice items about it.
PROCEDURE	Choose the correct option (a, b, . . .) in order to answer or complete the questions.
MARKS	● 2 marks for each correctly answered completed item ● 10 marks total

PART THREE

FORMAT	A text of approximately 350 words in which 20 discourse words have been replaced with gaps. These words are listed for the candidate together with 5 distractors.
PROCEDURE	Complete the text by transferring to the numbered answer sheet 20 of the 25 words provided.
MARKS	● 0.5 marks for each correctly placed word ● 10 marks total

PART FOUR

FORMAT	One or more texts and 10 True/False items.
PROCEDURE	Complete the text by transferring to the numbered answer sheet 20 of the 25 words provided.
MARKS	● 1 mark for each correctly answered item ● 10 marks total

TOTAL TIME: 50 minutes

Source: Reproduced with kind permission of El Ministerio de Educación, Cultura y Deporte.

just written. This is because, even if a task is designed to reflect the features referred to in relevant CEFR descriptors, it is simply impossible for even experienced item writers to predict how learners will react to specific test tasks and therefore how difficult they will turn out to be. Piloting may have to be very small scale, with a single class, but is an essential step in order to have some data – rather than just guesswork – about how difficult the tasks really are.

Figure 5.3 EOI blueprint for a B1 writing test

PART ONE

FORMAT	Personal correspondence / report (describing experiences, impressions, feelings, events).
PROCEDURE	Write a text of approximately 165 words on one of the two subjects given as options.
MARKS	15 marks total

PART TWO

FORMAT	Note / announcement (simple information of immediate character).
PROCEDURE	Write a text of approximately 60 words.
MARKS	10 marks total

TOTAL TIME: 55 minutes

Source: Reproduced with kind permission of El Ministerio de Educación, Cultura y Deporte.

5.1.4 Eliciting a sample of speaking and writing

The CEFR has made a significant contribution to a realisation that it is important to assess both *Interaction* (short turns) and *Production* (long turns) in a speaking test. Table 5.1 shows the relevant scales for spoken language activities and interaction strategies, and Table 5.2 collates and summarises the descriptors on those scales. The three columns in Table 5.2 list actions, the settings and types of topics, and specific topics.

Table 5.1 CEFR scales for spoken interaction and production

Communicative language activities	Communicative language strategies
Overall Spoken Interaction	
Understanding a native speaker interlocutor	Taking the floor (Turn taking)
Conversation	Cooperating
Informal discussion	Asking for clarification
Formal discussion (Meetings)	
Goal-oriented cooperation	
Obtaining goods and services	
Information exchange	
Interviewing and being interviewed	
Overall Spoken Production	
Sustained monologue: describing experience	Planning
Sustained monologue: putting a case	Compensating
Public announcements	Monitoring and repair
Addressing audiences	

Table 5.2 CEFR collated scale content for interaction and production

Level	Actions	Settings and topic types	Specific topics
C2	Convey finer shades of meaning precisely Eliminate ambiguity		
C1	Express him/herself fluently and spontaneously Get the floor, gain time and keep the floor while thinking Argue a formal position convincingly Answer complex counterarguments Summarise complex issues and texts Produce clear, smoothly flowing, well-structured speech Relate contribution skilfully to those of others	Social, academic and professional purposes Complex interactions between third parties Complex subjects	
B2+	Follow up statements and inferences by others Summarise Develop an argument systematically Mark clearly the relationships between ideas Highlight significant points appropriately and give relevant supporting detail	General, academic, vocational topics	Solution to dispute/case for compensation
B2	Interact with a degree of fluency and spontaneity Intervene appropriately in discussion Initiate, maintain and close discourse appropriately with effective turn taking Adjust to the changes of direction, style and emphasis Explain a viewpoint Account for and sustain opinions in discussion Gain time and keep the floor while formulating what to say Ask follow-up questions to check understanding and get clarification of ambiguous points	Discussion with native speakers Most general topics and topical issues A wide range of subjects related to his/her field of interest	Degrees of emotion Personal significance of events Relevant explanations, arguments and comments Advantages and disadvantages of various options

Table 5.2 (continued)

Level	Actions	Settings and topic types	Specific topics
B2	Give feedback on and follow up statements and inferences and so help the development of the discussion Help the discussion along, confirming comprehension, inviting others in Summarise		
B1+	Explain Take down and pass on Exchange information Interview Intervene in discussion, using a suitable phrase to do so Check and confirm Summarise, give opinion and help focus the talk Answer further questions of detail Help to keep a conversation or discussion going	Less routine situations Interviews/consultations Concrete information required Accumulated factual information Familiar routine and non-routine matters within his/her field	Abstract subjects, e.g. music, films Unpredictable occurrences e.g. accident Details of problems, messages, instructions How to do something A short story, article, talk, or discussion
B1	Initiate, maintain and close simple conversation Enter unprepared into conversations Keep going comprehensibly Give or seek personal views and opinions Express the main point he/she wants to make Ask someone to clarify or elaborate what they have just said Repeat back part of what someone has said to confirm understanding and help keep the development of ideas Invite others into the discussion	Informal discussion with friends Most topics pertinent to his/her everyday life Variety of familiar subjects within his/her field of interest; topics of personal interest	Experiences, reactions to them Plot of book/film, a story, reaction to it Dreams, hopes, ambitions Family, hobbies and interests, work Current events Travel, situations while travelling Detailed directions

A2+	Initiate, maintain and close simple conversation Exchange ideas and information Give an extended description Ask for repetition when he/she does not understand Ask for clarification about key words or phrases not understood If other person helps if necessary can ask for help to express what he wants to	Restricted face-to-face conversation Structured situations Simple, routine exchanges Familiar situations Predictable everyday situations Routine, everyday transactions Common aspects of everyday living Familiar topics Straightforward information Survival and routine travel needs Basic themes	Personal experience How he/she feels Pastimes, habits, routines Likes and dislikes People, places, a job or study Experience Past activities Simple directions and instructions Objects, pets possessions Events and activities Plans/arrangements
A2	Greet people and ask how people are React to news Ask and answer questions Express and respond to social functions (greetings, offers, invitations, etc.) Discuss what to do, where to go Make arrangements to meet Say whether you are following Ask for attention	Very short social exchanges Simple, routine direct information exchange Simple transactions and purchases Simple enquiries and requests for information Basic everyday needs of a concrete type Simple, predictable survival needs	Personal details, background, job Wants, needs and how he/she feels People, appearance, Daily routines; work and free time Places and living conditions Information about travel, public transport Everyday goods and services Quantities, numbers, prices
A1	Interact in a simple way Ask and answer simple questions	Basic everyday needs of a concrete type Areas of immediate need Very familiar topics	Themselves and other people Where they live Time, quantities, numbers, prices

Probably the single most important point in relation to the assessment of spoken language is that the task generates *discourse* and not just single sentences, in question and answer style. This may seem an obvious point, but it sets limits to what can be achieved with a simple interview or with paired role play. Traditional question and answer interviews or pair activities cannot take account of the dynamic relationship between cognitive, contextual and linguistic variables in performance, or of the way that skills, competences and strategies are integrated in language use.

To elicit a representative sample, we also need to generate *different types* of discourse. Essentially we have a choice between the following two options:

- A series of short, separate tasks each related to a CEFR descriptor, with each rated separately with a simple scale (e.g. 0–3 or 1–5). Alternatively there might be two to three such scales (e.g. for fluency, accuracy and task completion). This approach has been adopted in many schools. It is particularly suitable for Levels A1 and A2.
- A single longer activity like an interview or group task that relates to several CEFR descriptors, in which there are different phases. Learners will be assessed either separately for each of the phases or, more usually, once for the overall conclusion that the rater has come to on the basis of the varied evidence provided by the different kinds of language generated in the different phases.

Some examples of tests of spoken language that provide balanced phases generating different kinds of discourse are given below. Expressions in italics are the titles of relevant CEFR illustrative scales.

A test developed by the International Certificate Conference for ERASMUS students is cited in the CEFR (Council of Europe 2001:179) because it is such a good example. First there was a *Conversation* as a warm up, plus an *Informal Discussion* of topical issues in which the candidate had declared an interest. This was followed by a simulated telephone *Information Exchange* and then a *Spoken Production* phase, based upon a written *Report* in which the candidate gave a *Description* of his/her academic field and plans. Finally there was a *Goal-oriented Cooperation*, a consensus task involving two candidates.

The focus in the CIEP's DALF examination for French is on the use of one or more completely authentic texts on the same subject as a springboard for discussion. The candidates read the text(s). Then in the oral test they (a) summarise the main points made, glossing that report with their opinion in a *Sustained monologue: making a case*, (b) answer follow-up questions from an examiner in an *Information exchange* and then (c) engage in an *Informal discussion* with the examiner on the subject. There is no interaction between candidates in the exam, but the approach can be easily adapted to group work

in the classroom. Each person can have different text input and make their short presentation in turn.

Cambridge English Language Assessment in their core examinations use tasks that elicit different types of discourse too. All Cambridge English exams include a *Conversation* and *Information exchange* as warm up. There is often a *Sustained monologue* after a very short preparation, an *Information exchange* task between candidates and then a consensus task between candidates (*Goal-oriented cooperation*). The latter lend themselves particularly well to adaptation for classroom group work.

The exams of the Spanish EOI state language schools, whose test blueprints were mentioned in Section 5.1.3, offer another example. At B1, the first phase is an informal *Conversation* which is followed by a picture story *Spoken production sustained monologue* for which the candidates are given 1–2 minutes to prepare. The final resolution task between two or three candidates is *Goal-oriented cooperation*. At B2 the monologue phase is longer and the third phase is an *Informal discussion* involving elements of *Sustained argument: Putting a case*.

Eurocentres uses classroom assessment in which tasks are carried out in groups of 3–5 students, as mentioned in Section 3.3.2. The tasks have a structure that ensures that everyone speaks. They provide three distinct phases, each generating a different kind of discourse, that are used for three stages of assessment: initial impression, detailed analysis with criteria, and considered judgement, as summarised in Table 5.3. First there is a collaborative phase in which the group prepares something (*Goal-oriented cooperation*). The groups are then remixed so that each student has unique information and takes their turn to tell the others what their first group suggested or decided (*Information exchange, Sustained monologue, putting a case*). This then inevitably leads into a discussion phase as the group compares the proposals of each of the first groups (*Informal discussion*). A senior teacher acts as second assessor with less experienced teachers, with grades negotiated between the assessors after the lesson.

These examples show that there is no single way to assess speaking in relation to the CEFR. However, Eurocentres experience over 20 years suggests that classroom assessment with small group tasks as described above can be a viable alternative to examination-style interviews. They:

- can be very motivating, incorporating real world materials and issues that learners need to take a view on, report to a third party and support their conclusions in discussion
- offer a natural monologue that is embedded in interaction as is the case in real life, since all learners have a right to a long turn, and that turn provides an extended speech sample for each learner that the rater can focus on

Table 5.3 Eurocentres oral assessment procedure

	Assessment procedure	Instructions for a training activity
1. **Collaborative phase** Group works out what to do (short, slow turns with high use of communication strategies)	**Impression:** Write down the overall impression of the global level of the candidates that you have after about 5 minutes.	While viewing the video, after 4–5 minutes, write a single level – your overall, initial impression – in the space at the top of the rating form.
2. **Exchange phase** in which each student has a chance to take the floor (long, coherent turns which are semiprepared)	**Analysis:** Consciously read the descriptors for that level across the assessment grid. If you confirm that the candidate does meet the criterion description for a category at that level, look at the level above in that same category to see if they are even better than that. Write a result for each assessment category (Range, Accuracy, Fluency, Interaction, Coherence if using CEF Table 3).	While viewing, after marking that initial judgement, consciously read the descriptors for that level across the assessment grid, for the level above and the level below. After viewing, read the criteria closely and mark your decision for each category on the form in the space provided
3. **Discussion phase** in which some members of the group take things further (spontaneous, short turns).	**Judgement:** Compare your analysis result to your original impression and make a considered judgement.	Consult the CEFR scales for 'Overall Spoken Interaction' and 'Overall Spoken Production'. Write your final decision at the bottom of the form in the space provided.

- generate spontaneous discussion in which learners may well use interaction strategies like indicating when they are following, checking understanding, asking for clarification, checking common ground, summarising, and correcting misinterpretations (see Council of Europe (2001:86–87) for descriptor scales on interaction strategies).

Naturally there will be circumstances, for example when a speaking test serves a formal gatekeeping function, in which an examiner/interlocutor may be deemed unavoidable. But even in such cases it is possible to have the best of both worlds by adding such examiner 'probing' at the end of a small group task.

In eliciting a performance for the assessment of writing, the issues are not entirely different from those with speaking: there need to be different types of writing which involve different types of discourse. The relevant CEFR scales are shown in Table 5.4.

Written interaction is essentially writing in the same way that one would speak, as for example in personal letters. In today's world of email, texting and internet chatting, is has become even clearer that this is fundamentally

Table 5.4 CEFR scales for written interaction and production

Communicative language activities	Communicative language strategies
Overall Written Interaction	
Correspondence	Cooperating
Notes, Messages and Forms	Asking for clarification
Overall Written Production	
Creative Writing	Planning
Writing Reports and Essays	Compensating
	Monitoring and repair

different from written production, in which completely different conventions and standards of correctness apply. There are practical constraints that limit how many writing samples can be collected, but two contrasting samples would seem to be a sensible minimum.

Whatever writing tasks we choose, one fundamental question is the attitude we take to task completion. Is the task rubric a 'coat hanger' on which to hang a sample of written language, or is it a detailed instruction to produce a piece of genre writing? There is no simple answer to this question; it depends on the learning context. Even in apparently very similar contexts, adjacent pedagogic cultures may have opposing traditions, as with Norwegian and Swedish secondary schools.

Portfolio assessment is a good way of covering different types of genre, avoiding the restrictions of timing, and varying how one deals with the 'coat hanger/genre' issue. Above all a portfolio approach takes account of the drafting and editing processes inseparable from serious writing in the real world. Over a period of time, learners can be set a variety of relevant tasks and also encouraged to redraft and correct them after feedback. The main value of a portfolio approach comes from this additional learner training. The ideal, of course, would be to combine a portfolio approach (continuous assessment) with assessment of a timed writing task done in class.

5.1.5 Criteria for judging speaking and writing

Developing a criteria grid involves specifying the categories that will serve as criteria to be rated, deciding whether to use the same categories for all tasks and for all levels, drafting the descriptors, and then, if scores are given, developing a scheme to translate those scores into CEFR levels. The same grid might be used with all tasks because the tasks might always be similar, designed to a tight specification. On the other hand, we may select a set of qualitative criteria particularly appropriate for the specific task concerned. The former is the approach taken by many examination institutes and schools; the latter is the one suggested by the CEFR-based scenario approach introduced in Section 4.2.2. There are arguments on both sides.

One crucial issue is the *number* of categories and grades (or levels). Three grades for three categories ($3 \times 3 = 9$ rater judgements) are very easy to work with; nine grades for nine categories ($9 \times 9 = 81$ rater judgements) are definitely not. It is no coincidence that criteria grids tend to stick to four or five categories and four or five grades/levels. Any more than that and raters may start to suffer from cognitive overload and assessments may become less reliable.

CEFR Chapter 5 offers illustrative scales for a range of qualitative aspects of language use that can be exploited in the process of defining the criteria grid. In addition, CEFR Table 3 defines qualitative aspects of spoken language using published CEFR descriptors, regrouped into the five categories: Range, Accuracy, Fluency, Interaction and Coherence. The 'plus levels' (see Section 3.2.2.) are indicated on CEFR Table 3, but they are not defined. A grid supplementing CEFR Table 3 with descriptors for 'plus levels' was developed for the international benchmarking seminars that produced the DVDs of CEFR illustrative samples. Table 5.5 combines the two grids into one single criteria grid with the five categories defined for the resultant nine levels. All but two of the descriptors in Table 5.5 are published, validated CEFR descriptors. The two exceptions (A2+ Accuracy; B1+ Coherence) are given in italics.

We must bear in mind that CEFR Table 3 and its extension in Table 5.5 are *reference tools* not operational tools. Using Table 5.5 for live assessment would mean $9 \times 5 = 45$ rater decisions. That is enough to get many people's heads spinning. Also, unless one is likely to encounter learners at any CEFR level, it makes little sense to include all the levels. What CEFR Table 3 and Table 5.5 illustrate is that when drafting a criteria grid, we need descriptors for aspects of quality (CEFR Chapter 5), not descriptors of communicative activities (CEFR Chapter 4). Precisely which qualitative aspects should be selected as criteria is a question of context and assessment purpose. The content of Table 5.5 is used in this section for illustration, as CEFR Table 3 is used for illustration.

As discussed in the CEFR (Council of Europe 2001:179–182) there are different ways in which descriptors can be presented as assessment criteria:

- *Grid for all levels*, like Table 5.5 – or for just the range of relevant levels.
- *Grid for one level*, with criteria defined for different grades, norm-referenced around the standard for the target level. An example is given in Table 5.6.
- *Short checklist for one level*, with one descriptor per category at that level only. Examples are given in Tables 5.7 and 5.8.

Rather than showing all levels as in Table 5.5 or a range of levels, the grid can be focused just on the level that has been set as the standard, adding the

Table 5.5 CEFR Table 3: Qualitative aspects of spoken language use, expanded with 'plus levels'

	Range	Accuracy	Fluency	Interaction	Coherence
C2	Shows great flexibility reformulating ideas in differing linguistic forms to convey finer shades of meaning precisely, to give emphasis, to differentiate and to eliminate ambiguity. Also has a good command of idiomatic expressions and colloquialisms.	Maintains consistent grammatical control of complex language, even while attention is otherwise engaged (e.g. in forward planning, in monitoring others' reactions).	Can express him/ herself spontaneously at length with a natural colloquial flow, avoiding or backtracking around any difficulty so smoothly that the interlocutor is hardly aware of it.	Can interact with ease and skill, picking up and using nonverbal 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.

Table 5.5 (continued)

	Range	Accuracy	Fluency	Interaction	Coherence
B2+	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	Shows good grammatical control; occasional 'slips' or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.	Can intervene appropriately in discussion, exploiting a variety of suitable language to do so, and relating his/her own contribution to those of other speakers.	Can use a variety of linking words efficiently to mark clearly the relationships between ideas.
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.

B1+	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influences.	Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and 'cul de sacs', he/she is able to keep going effectively without help.	Can exploit a basic repertoire of strategies to keep a conversation or discussion going. Can give brief comments on others' views during discussion. Can intervene to check and confirm detailed information.	Can use connecting words to link sentences into a coherent sequence, though there may be some 'jumps'.
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can initiate, maintain and close simple face-to-face conversation on topics that are familiar or of personal interest. Can repeat back part of what someone has said to confirm mutual understanding.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.
A2+	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics, though he/she will generally have to compromise the message and search for words.	Can use some simple structures correctly in common everyday situations.	Can adapt rehearsed memorised simple phrases to particular situations with sufficient ease to handle short routine exchanges without undue effort, despite very noticeable hesitation and false starts.	Can initiate, maintain and close simple, restricted face-to-face conversation, asking and answering questions on topics of interest, pastimes and past activities. Can interact with reasonable ease in structured situations, given some help, but participation in open discussion is fairly restricted.	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points.

Table 5.5 (continued)

	Range	Accuracy	Fluency	Interaction	Coherence
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can 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 structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can ask and answer questions about personal details. Can interact in a simple way but communication is totally dependent on repetition, rephrasing and repair.	Can link words or groups of words with very basic linear connectors like 'and' or 'then'.

Table 5.6 Assessment grid focused on one level, including levels above and below

	Range and precision	Accuracy	Fluency
5 B1	Can talk about family, hobbies and interests, work, travel, news and current events. Can make the other person understand the most important points.	Can express self reasonably accurately in familiar, predictable situations.	Can keep a conversation going, but sometimes has to pause to plan and correct.
4			
3 A2+	Can talk about familiar everyday situations and topics, with searching for the words; sometimes has to simplify.	Can use some simple structures correctly in common everyday situations.	Can participate in a longer conversation about familiar topics, but often needs to stop and think or start again in a different way.
2			
1 A2	Can communicate in a simple and direct exchange of limited information in everyday situations; otherwise has to compromise the message.	Can use correctly simple phrases learned for specific situations, but often makes basic mistakes – for example mixing up tenses and forgetting to use the right endings.	Can make self understood with short, simple phrases, but often needs to stop, try with different words – or repeat more clearly what was said.

levels above and below as points of reference. Table 5.6 gives an example of a grid of this type with Level A2+ set as the standard, but with A2 and B1 as reference points. A rating scale from 1 to 5 allows for each learner to be given a finer score for each criterion. A score of 15 (= 5 on all criteria) would mean B1. Notice that it is not necessary to define the 'in between' grades '2' and '4' on such grids. It is clear that 4 is a really good A2+ but not yet B1.

On the other hand we may prefer to rate learners' performance only in terms of their success at achieving the targeted CEFR level, ignoring all other levels. A simple and quite common approach is shown in Table 5.7.

Table 5.7 Assessment at one level (A2+)

	Candidate A				
Range and precision: Can talk about familiar everyday situations and topics, with searching for the words; sometimes has to simplify.	1	2	**3**	4	5
Accuracy: Can use some simple structures correctly in common everyday situations.	1	2	**3**	4	5
Fluency: Can participate in a longer conversation about familiar topics, but often needs to stop and think or start again in a different way	1	2	**3**	4	5

Here only the descriptor for a '3', the target level, on the 15-scale, is used. To ensure coherence between grids for different levels, the simplest approach

is to say that '5' is a performance meeting the criteria for the level above and "1" is one meeting those for the previous level as was the case with Table 5.6. In this way the criteria grids for all levels are locked together through the descriptors acting as the criteria for each level, at point '3'. This is the kind of approach taken by both Cambridge and Pearson. If this is not done, the transparency promoted by the CEFR is undermined. What would '5' mean if it is not the next level? Of course there is no guarantee that a learner achieving a '5' on a B1 task would definitely achieve a '3' on a B2 task; one would have to administer a B2 task to be sure. Nevertheless, this approach has the advantage of directness and simplicity. It is easier to inform learners what the qualitative objectives are, and to include them on the list of aims for the level concerned.

The examples above have, purely for the purpose of illustration, maintained the categories and the wording of our reference tool Table 5.5. However, both the selection of categories and the formulation of the criteria for an operational tool should be a local development, with the descriptor scales in CEFR Chapter 5 as a reference. Table 5.8 gives an example for Writing at A2 from

Table 5.8 Writing assessment grid: Level A2, Avo-Bell, Sofia

Writing A2	**Candidate A**					**Candidate B**				
Text management Can make her/himself understood in short sentences Can produce a short but logically connected text which is relevant to the task Can link groups of words and sentences with simple connectors like 'and', 'but', 'when' and 'because'	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Communication strategies/Effect on the target reader Can convey more complex meaning using strategies like: reporting events in chronological order; describing aspects of everyday life; filling in questionnaires	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Layout and organisation Can use more confidently opening/closing expressions in a limited number of written tasks, e.g. simple letters, postcards, descriptions Can link ideas in clear paragraphs	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Grammar and vocabulary (accuracy and appropriacy) Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple texts on everyday topics Uses some simple structures correctly, but still systematically makes basic mistakes	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
Global achievement Overall impression mark/Task achievement (all points covered)	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**

an EAQUALS member school in Bulgaria (Avo-Bell, Sofia) that is focused on one level like Table 5.7. Here the more linguistic content comes from the CEFR, whereas the aspects more specifically concerned with writing have been developed locally. In the first category *Text management* the first and third descriptors are CEFR descriptors, but the middle one has been formulated locally. In the last analytic category *Grammar and vocabulary (accuracy and appropriacy)*, CEFR content has been slightly reworded. The two middle categories, on *Communication strategies*/*Effect on the reader* and *Layout and organisation*, have been developed entirely locally.

The Eurocentres criteria grid gives an example of the opposite approach. It is used in an intensive teaching situation with classes at all levels. It defines 10 levels for four categories and the current entries for Levels B1 and B1+ are given in Table 5.9. Descriptor elements shared with the CEFR are in bold. As can be seen, in the sections on Range and Accuracy, comments are made about the use of particular language. This sort of comment is not included in the CEFR firstly because the CEFR applies to a range of languages and secondly because it may well vary according to the context.

Table 5.9 Eurocentres spoken assessment grid: RADIO (Note: bold = exact CEFR content)

	Range	Accuracy	Delivery	Interaction
B1+	Able to use a range of simple language flexibly, and **explain a point with reasonable precision**, but can't always say what they would like to. Familiarity with main tenses, modals and major sentence patterns.	**Reasonable accuracy with basic tenses etc. in everyday contexts.** Frequent errors and inappropriate expressions occur, **partly due to mother tongue influence, but it is clear what he/she is trying to express**.	**Gives extended descriptions, able to keep going effectively without help, despite some problems with formulation resulting in pauses and 'cul de sacs'.** Stress and intonation may be very foreign, but can generally be followed okay.	Handles structured **discussion on familiar topics easily, inviting others in, commenting on views, comparing and contrasting alternatives**. Participation more restricted in freer or unfamiliar contexts.
B1	Relatively wide repertoire of simple language, for **familiar subjects, but limited alternatives**. Normally requires simplification of intended message. Good level of familiarity with basic tenses and sentence patterns.	**Reasonable accuracy with a repertoire of frequently used 'routines' and patterns.** Tendency otherwise to mix up tenses and pick the wrong word or expression; may be conscious of this and try to self-correct.	Can keep going clearly and comprehensibly, though perhaps slowly, pausing especially in longer stretches, Frequent reformulations and hesitations and/or heavy interference from L1 may make comprehension difficult.	**Initiates, maintains and closes simple interaction** with some cooperation from the interlocutor. **Can exchange information and repeat back part of what someone has said to confirm mutual understanding.**

In both these examples, Avo-Bell and Eurocentres, the same categories are used as the criteria for all tasks. But there is of course no reason why the categories for the criteria should always be the same. If a portfolio approach is taken with writing, or if a series of radically different spoken tasks are undertaken over a period of time, why not have a couple of 'core' criteria for all tasks (like Range/complexity and Accuracy) and vary two or three other criteria according to the pragmatic and sociolinguistic demands of the task?

Even when criteria grids have been developed it is still necessary to carry out training to standardise the interpretation of the CEFR levels in the school by showing what type of performance is typical at different CEFR levels. Such training should be carried out with CEFR illustrative samples (see Section 3.1.4). All assessment is essentially a comparison, either a comparison between performances or a comparison against an internalised standard. That standard can only be accurately internalised in relation to the standards for other levels. In other words raters need perspective and they can get this from standardisation training. Discussing concrete examples of performances in relation to common criteria, supported by detailed documentation that explains why a performance is one particular level, is an effective way of counteracting problems like the following:.

- Raters often think they know the CEFR levels without having looked at either the CEFR descriptors that define the levels or the samples that illustrate them.
- Raters' impression of CEFR levels may be formed by an (incorrect) association of a local textbook, course level or examination with a CEFR level, even though the book, course organisation or exam predates the CEFR and has been merely relabelled without building any validity argument to support the claim.
- Raters can interpret the written word (the descriptors) in different ways; some people are just stricter than others. People do not realise this; they naturally think that they and their colleagues share the same professional interpretation until shown that this belief is an illusion.
- Raters often make judgements based on private criteria that they are unaware of. In particular, teachers often focus too much on linguistic accuracy, with unrealistic expectations as to the level of accuracy that it is reasonable to expect at any given level.
- Whatever scale or grid is used, raters will tend to refuse to give the top grade, a classic rater error noted since the 1950s.
- Teachers rating at a level that they do not teach may be excessively strict.

Standardisation training should be conducted in a clear, logical order to avoid any procedure which causes a participant who is an 'outlier' (someone with an extreme view) to be forced to 'out' themselves as such at the beginning

of the process (e.g. 'Hands up who thinks A1?'). Exposing participants to ridicule in this way at a moment when their opinion is in fact malleable – since outliers normally move to the consensus in a second round – is counter-productive as well as thoughtless. There is a possibility that the 'outed' outlier may in reaction dig in rather than backing down. This may in turn sabotage the chances of reaching a consensus that coincides with the common interpretation. In training, small group sessions with volunteer rapporteurs and anonymous data collection are far more effective than public shows of hands.

The Council of Europe's Manual for relating tests and examinations to the CEFR gives detailed advice on running standardisation sessions. They are best conducted in the following order: familiarisation exercises with CEFR descriptors; illustration using the samples and documentation provided, small group collective rating, and finally individual rating.

5.1.6 Developing tests for the receptive skills

The CEFR scales give a typology of different kinds of listening and reading, which whilst not exhaustive, suggests sampling different genres and user purposes in a test. Scales are provided for the following areas:

Reception	**Spoken**	**Overall Listening Comprehension** Understanding Interaction between Native Speakers Listening as a Member of a Live Audience Listening to Announcements and Instructions Listening to Radio and Audio Recordings
	Audio/Visual	Watching TV and Film
	Written	**Overall Reading Comprehension** Reading for Orientation Reading for Information and Argument Reading Instructions Reading Correspondence
	Working with Text	Note-taking in Seminars and Lectures Processing Text

In almost all conceivable contexts, the distinction will be relevant between the following two types of reading:

- *Reading for orientation:* skimming through a text quickly to decide whether to read it; searching or scanning through a text quickly to find specific information; sometimes called *expeditious reading*, and
- *Reading for information and argument:* to understand main ideas and important details; sometimes called *careful reading*.

Scanning a brochure to find the right product or service is a skill needed at even an elementary level; skimming through a long article or book to see whether it is relevant to the current field of enquiry is a skill very necessary at higher levels for academic study and the world of work. This suggests that a reading test at any level might contain tasks for the different types of reading: expeditious/careful. There might first be a section with a couple of short texts or artefacts (like adverts, extracts from catalogues etc.) accompanied by single item tasks focused on identification and matching. These might be followed by a section with one or two longer, prose texts – perhaps of different genres – for detailed comprehension. This may sound like obvious good practice, yet it is remarkable how many reading and listening tests have only one prose text. Apart from the fact that candidates familiar with that particular genre and topic will have a huge advantage, how can an individual's result in regard to one text be generalised to their ability at reading or listening as a whole?

The CEFR descriptors for Reception were summarised schematically in Table 2.2. It provides a starting point for a selection of texts and helps to define the assessment conditions for tasks. Table 5.10 extracts the micro-skills from that summary and adds information processing such as that described in the CEFR scales for *Working with Text* and that which occurs in academically oriented language examinations. Bold text indicates wording from a CEFR descriptor; italic indicates descriptors from the prototype Portfolio and normal print indicates new points added during discussion in the EAQUALS Special Interest Project whose work is reported later in this section.

Making a list of micro-skills/actions that should be tested at a particular level is a first key step in developing a specification. A second question is to specify the text genres to which people should apply those skills. Tables 5.11 and 5.12 provide summaries of genre types relevant at different CEFR levels for listening and reading, developed by the author from the wording of CEFR descriptors for the British Council/EAQUALS Core Inventory project. Following the convention used in the Core Inventory, darker shading represents the level(s) at which the source seems 'core'.

CEF-ESTIM (cefestim.ecml.at) offers a mechanism that claims to estimate the CEFR level of texts and tasks between A2 to B2 for classroom teachers. One can enter information about the text, the communicative language activities, the communicative language competences, and the communication strategies involved. Then the site calculates a score for the estimated level of a learner who could cope with the resultant task.

Table 5.10 CEFR micro-skills for reception

	A1	A2	B1	B2	C1	C2
Recognise	• Familiar names, words and very basic phrases	• Specific information in lists, reference works • **Changes in topic**	• **Useful information** • *Relevant facts and information*	• **Which part(s) of the text(s) is relevant to the purpose** • **A change of direction, style or emphasis** • **Different formulation of the same ideas**	• **Which part(s) of the text(s) is relevant to the purpose** • **Highlighting of the most important points** • **Variation of style for effect** • **Register shifts** • What will come next • *The social, political or historical background of a literary text* • **Irony**	• **Subtle distinctions of style** • *Hidden value judgements* • **Understatement** • **Irony and sarcasm**
Process with accompanying text	• Figures/numbers • An image • A sign/symbol • Timetable • Calendar • Contact details • Town map	• Figures/numbers • Diagrams (equipment) • Organigrams • **Map or plan**	• Diagrams (object, machine, organism) • Tables • Pie charts etc.	• Charts • Graphs • Diagrams (flow charts, classifications, contrasts)	• Complex, specialised diagrams, charts and graphs	
Distinguish			• **Main point/** *relevant facts and information* **from specific details**	• **Main points from relevant supporting detail/** arguments**/ examples**	• **Main point from supporting** – **arguments** – **subthemes**	

Table 5.10 (continued)

	A1	A2	B1	B2	C1	C2
Distinguish			• **Main conclusion** *from preceding detail*	• Such supporting arguments and more precise information from a digression • *Aspects reported as facts from those reported as opinion*	– **details** – **example**s	
Understand	• *Info about people* • *Times, locations* • *Simple messages* • *Basic information*	• **An idea of the overall meaning** • Well-signalled **main points** • **Essential/** *most important* **information** • **Specifically required information**	• **Main points/** *most important information* • **Relevant factual information** • An explicitly signalled **line of argument** • **Main conclusions** • **Specific details**	• **Main ideas** • **Essentials/ essential meaning** • Explicitly stated **points of view, opinions and complex lines of argument** • *Expressed, specific attitudes* • **Specific details** • **Mood and tone** explicitly expressed in the text	• **Attitude, mood, intentions** • **Implicit meanings, ideas and connections** • **Implied as well as stated opinions** • **Implied**, *indirect and ambiguous* **relationships** • **Use of humour** • **Implicit cultural references**	• **Implied attitudes** • *Nuances and* **finer shades of meaning** *and differentiation* • **Implied opinions** • **Implied**, *indirect and ambiguous* **relationships** • *Metaphors, symbols, connotations, and their function within the text* • *Very colloquial style* • **Use of humour** • *Plays on words, puns* • *Satire and its function*

Table 5.11 Written text sources relevant to CEFR levels

	A1	A2	B1	B2	C1
Written sources					
Signs and notices	simple	everyday			
Directions	A to B		detailed		
Technical instructions		telephone		complex	outside area
Warnings on hazards				detailed	
Regulations		safety		detailed	complex
Conditions				details	
Menus	simple				
Maps, tourist leaflets and posters					
Advertisements	simple				
Timetables	simple				
Reference lists (*Yellow Pages* etc.)					
Web pages, catalogues					
Brochures and leaflets					
Guides			short		
Forms, invoices					
Correspondence: formal letters		basic			
Official documents			short		
Technical texts (e.g. contracts)					
Factual descriptions	visual too		events		
Newspaper and magazine articles		events	main point		
Factual texts, articles and reports					
Lengthy complex texts, reports					
Highly specialised sources					
Argumentative texts			conclusion		
Reviews, editorials, commentaries					
Narratives					
Anecdotes, jokes			simple		

Table 5.11 (continued)

	A1	A2	B1	B2	C1
Fiction			simplified		
Literature				standard	
Messages on postcards					
Short text message/ Twitter					
Correspondence: informal letters		short, simple			
Personal descriptions			feel, wish	feel, wish	literary

Although the approach taken by CEF-ESTIM is perhaps a little ambitious, the CEFR descriptors can be taken as the starting point for the development of listening and reading tests, despite the criticism that they received from some language testers, as discussed in Section 2.3.3. The issue is to identify the key terms in relevant descriptors, and to elaborate how one should interpret those key terms in relation to the item concerned. This approach influenced an EAQUALS Special Interest Project that has produced classroom assessment tasks for listening and reading. The group followed the CEFR-based scenario approach introduced in Section 4.2.2 (objectives and implementation) in order to develop scenarios for assessing listening and reading and to provide illustrative materials for those scenarios.

Table 5.13 shows the first page of a B1 scenario for listening to a tour of a historic site. At the top one sees the domain, context, real world tasks, type of language activity and type of text. On the left is the descriptor-based information, first 'Can Dos' from CEFR Chapter 4 or Portfolio and then the micro-skills and text features stated or implied in relevant descriptors. Formulation from published B1 CEFR descriptors is in bold. The competence section on the right hand side has linguistic content from the British Council/EAQUALS Core Inventory for General English. The entries for strategies do not come from a specific source since reception strategies are not well developed in the CEFR.

Table 5.14 shows the completed specification template for an illustrative implementation of this scenario. Three tasks accompany a YouTube

Table 5.12 Spoken text sources relevant to CEFR levels

	A1	A2	B1	B2	C1
Spoken sources					
Interlocutor					
3rd party interaction		topic	main points	modified	complex
Discussions and debates			everyday	modified	complex
Technical discussions					complex
Directions		A to B	detailed		
Messages		main point			
Announcements		main point			distorted
Instructions				everyday	complex
Conditions, warnings					
Lectures, talks, presentations		outline	essentials		nonstandard
Film			visual/ action		idiomatic
Shows, drama					idiomatic
TV news reports		events			
TV interviews					
TV documentaries			visual/ action		
TV current affairs					
TV talk shows					
Radio news			main points		
Radio documentaries					
Wide range of radio broadcasts					
Narratives					
Recorded passages		short, slow			
Recorded audio materials			simple	standard	

Table 5.13 Assessment task scenario for B1 listening (Site tour): Overview

LISTENING ASSESSMENT: SITE TOUR OVERVIEW B1				
DOMAIN	**CONTEXT**	**TASKS**	**ACTIVITIES**	**TEXTS**
Personal/Educational	Home School	Find out if you are interested in a historical site	Listening as a member of an audience Listening to audio (visual) media	A guided TV/Web commentary on a place (historic building, resort, town etc.)

LEVEL	**B1**
CANDOS*	Can catch the main points in simple, factual TV/web programmes or documentaries on familiar topics when the delivery is relatively slow and clear.
MICROACTIVITIES*	
RECOGNISE	• Useful information • Relevant facts and information
DISTINGUISH	• Main points/relevant facts from specific details
UNDERSTAND	• Main points/essential information • An explicitly signalled line of argument • Main conclusions • Specific details
TEXT FEATURES*	• Long but straightforward text • Clear, standard, relatively slow • Clearly signposted/signalled with explicit markers

COMPETENCES		
STRATEGIC	**Recognise** the beginning of a significantly new and different part of the text **Recognise** where difficulty lies (subject/assumed knowledge, linguistic) **Use** context to deduce probable meaning of unknown words (repetition, visuals, gesture, what comes next) **Use** the beginning of a significantly new and different part of the text to intensify effort **Use** visual support	
PRAGMATIC	**Functional**	Describing places Describing events Describing feelings, emotions, attitude
	Discourse	Linkers: sequential – past time Connecting words expressing cause and effect, contrast etc. (e.g. on the other hand; however; despite) Summarising

* *Content from CEFR scales and/or Swiss EAQUALS/ALTE Portfolio in normal print.*

Overview and implementation: Brian North

	LINGUISTIC	**Grammatical**	Past time: Simple past, past continuous, used to, past perfect Passive (past) Reporting structures 3rd conditional/ mixed conditionals Must/can't/might have Intensifiers Comparatives/Superlatives
		Lexical	Adjectives for places and people Time phrases (e.g. In the last century; 50 years ago) Verbs describing construction, development (e.g. plan, construct, rebuild, renovate, demolish)
		Phonological	Emphasis in sentence stress

Table 5.14 Assessment task scenario for B1 listening (Site tour): Implementation

LISTENING ASSESSMENT: SITE TOUR: IMPLEMENTATION (HAMPTON COURT) B1

TASK 1		LISTENING
GENERAL DESCRIPTION		Students watch twice the introduction to a TV guided commentary on Hampton Court, answering five True/False/Not Stated questions while listening.
SOURCES		'The Tudors: Behind Hampton Court': http://www.youtube.com/watch?v=mX7ABmAlcAE *Natalie Dormer of The Tudors celebrates the 500th anniversary of Henry VIII's coronation by touring the Hampton Court Palace*
TEXT FEATURES	AUTHENTICITY	Authentic
	LENGTH	7–10 minutes
	VISUAL SUPPORT	Yes – commentary should match unfolding film
ITEM TYPE / NUMBER		Introduction: True/False/Not stated: (5 questions)
TASK RUBRIC		Listen twice to the introduction to a TV guided commentary on a historic place of interest 'Hampton Court Palace', near London and answer the 5 questions. Mark 'T' if the statement is True, 'F' if it is false and 'NS' if the information is Not Stated in the commentary.
CONDITIONS		While playing
MARK SCHEME		5 x 1 mark = 5 marks
TASK 2		LISTENING
GENERAL DESCRIPTION		Students watch the main body of the video twice and match 10 things that are mentioned (avoiding 3 distractors) to the parts of the palace they relate to.
ITEM TYPE / NUMBER		Matching: Match things mentioned in the commentary to the part of Hampton Court concerned (3 extra distractors).

TASK RUBRIC	Listen twice to the main body of the guided commentary. While you are listening match the 10 things in the list to the parts of the palace they relate to. Put a cross in the correct box for each point on the list. Note: 3 of the points on the list are not mentioned.
CONDITIONS	While playing
MARK SCHEME	10 x 0.5 mark = 5 marks
TASK 3	LISTENING
GENERAL DESCRIPTION	Students read the 5 multiple-choice questions before listening a third and final time to the whole video in order to find the answers.
ITEM TYPE / NUMBER	Multiple choice: 5
TASK RUBRIC	While listening one last time, answer the following 5 questions with information from the commentary. Choose the best answer.
CONDITIONS	While playing
MARK SCHEME	5 x 1 mark = 5 mark.

video extract from a BBC programme showing a tour of Hampton Court. The video is a trailer for the series 'The Tudors'. It has a short, voiced over introduction that is used for five True/False questions. The main body of the recording is then a walkabout interview between one of the actresses in the series and a historian, with cuts to short scenes from the series, as the pair go around the palace. This walkabout is used for a 10-item matching exercise. Finally there is a post-viewing task – originally five open questions. In the completed template, in addition to a general description, information is given about the source text and its degree of authenticity, about item types, task rubric, conditions (time allowed, number of times played – here twice) and the marking scheme.

At the first piloting of this particular assessment task, with a Swiss-German 14 year old secondary school class, a difficulty was encountered. One text feature from the descriptor states 'Familiar topics regularly encountered in a school, work or leisure context'. Yet this is not really the case with the Tudors. The class managed the True/False questions okay (mean 60%) and the matching task was a little easy as intended (mean 85%), but the open questions on an unfamiliar subject were too hard (mean 40%). The teacher wrote in her feedback: 'It was difficult for them to write about it instantly in English as they are not used to doing so without a dictionary and they do not have much general knowledge about Tudor times anyway.' However, the learners had become interested, so she went on to do more material on the topic. The experience throws up the fact that, even when following this kind of systematic approach, an element of interpretation inevitably remains. The text-delivery features of this video were ideal for B1. The guided tour of a tourist site was fine as a 'familiar topic encountered in a school or leisure environment'. But that did not apply to finer points of the historical background. However, it was not so much the comprehension of the video that caused the problem; it was writing answers to open questions. Swiss-German secondary students are not accustomed to writing spontaneously in English. As a result of this experience the open questions were replaced with multiple-choice questions. This underlines the importance of piloting even low-stakes assessment items in order to check (a) that the desired construct is being tested (here listening rather than writing), and (b) that item types appropriate to context have been selected.

The scenario concept generated a lot of interest. However, early staffroom consultation produced the reaction: 'Oh no. We can't all do that every time.' As a result of this feedback, we realised that whilst the scenario is a nice heuristic concept to promote motivated choice of texts and tasks, it was an awareness-raising tool, rather than a help for busy teachers. We therefore complemented the illustrative scenarios with wider sets of ideas for assessment tasks. This produced a 'task collection' for listening and for reading at

each level. These were produced with a systematic methodology adapted from the approach that Glyn Jones had used in the development of Pearson PTE-General. In the EAQUALS group we used Word tables. For each descriptor in column one, the micro-skills stated or implied are put in the second column with the text features stated or implied put in the third column. At this point a fifth column is completed with an example task, including the type and number of items. It sometimes happened that the creation of a concrete example for a task at this juncture led to the addition of a second or even third CEFR descriptor to the first column, requiring the addition of further micro-skills and text features to the second and third columns. Then, working in both directions from the micro-skills that should be assessed on the one hand, and from the example test task in the fifth column, the fourth column, for task features, was completed. Considering the micro-skills that should be assessed in this task and the item types that have been chosen, what precisely should the learner be expected to do in the task, under which conditions and with what support, if any?

An example of this process for one entry in the 'task collection' for listening at B1 is given in Table 5.15. It is from this specification that the Hampton Court Site Tour listening scenario (Tables 5.13–5.14) was produced. The first three columns in Table 5.15 relate to the CEFR/Portfolio descriptor-based objectives on the left-hand side of the scenario model shown in Table 5.13.

Table 5.15 Example entry in the Task Collection: Tour of a historic site

	CEFR descriptor	Micro-skills	Text features	Text features	Example with item types
B1	I can catch the main points in TV programmes on familiar topics when the delivery is relatively slow and clear. I can follow a lecture or talk within my own field, provided the subject matter is familiar and the presentation straight-forward and clearly structured.	• Recognise new sections • Distinguish main points from specific details • Understand an explicitly signalled line of narrative/ argument • Understand specific details	• TV programme with short report/ guide and interview(s) • Topics; Familiar and regularly encountered in a school, work or leisure context • Presentation: straight-forward and clearly structured • Delivery: Clear, standard, relatively slow	• Recognising new sections (relevant to topics of questions) • Identifying main points • Understanding essential information • Could hear twice/three times • Dictionaries allowed for open questions	Follow a TV guided commentary on a place (e.g. Tour of Hampton Court, Versailles; extract from travel programme/ tourism promotion) • T/F/NS • Matching • Information transfer (table or diagram) • Open-ended questions

The first column shows the CEFR/Portfolio descriptors themselves, followed by micro-skills and text features in the second and third columns respectively. As can be seen from the lighter highlighting, most aspects of the text features are taken directly from these two CEFR descriptors. Text features include genre, topic, spoken delivery features, length, organisation and functional discourse types. On the other hand, only one aspect of the micro-skills comes from these two descriptors; the others are micro-skills for recognising, distinguishing and understanding taken from the summary given in Table 5.10. These are shown in darker highlighting.

The three right-hand columns in Table 5.15 concern the realisation of the task: task features include receptive micro-skills, actions required, item aspects and conditions. Item aspects include issues like the order of the questions and extent to which information required is formulated in a similar way in question and text. Conditions define time constraints and support allowed. Table 5.16 shows more entries for the B1 task collection for listening.

In the process of defining micro-skills and text features, we expanded the list considerably from the original set taken directly from CEFR descriptors that was shown in Table 5.10. The entries for B1 micro-skills have doubled from the original CEFR-based 12 to 24, as shown in Tables 5.17 and 5.18.

The elaboration of this richer description of micro-skills and text features at CEFR levels by a process of logical deduction is a good example of the way in which the CEFR levels come alive when we work with them. This deeper specification is a necessary part of any CEFR-based test development. The difficulty in this process is not so much teasing out the implications of CEFR descriptors, which is actually straightforward. The main problem concerns the level of difficulty of the *language* in the source text itself. In the development of tasks for Pearson PTE General following a methodology of the type described above, Pearson has reported that the most frequent reason for the rejection of tasks at the review stage, affecting approximately half of the 6.5% that got rejected at that stage, is the fact that the language of the source text is not appropriate for the level. And in three-quarters of those cases, the texts were the wrong level because they contained vocabulary and structures that were too complex for the level concerned.

As a (former) teacher one tends to have a good instinct regarding what level of text learners at a certain level can handle – hence the rejection rate of only 6.5% cited above. Nevertheless it is a problem, and demonstrates that all assessment tasks should be reviewed by someone who did not produce them – and then piloted.

The EAQUALS project has a number of aims:

- to provide a simple, systematic methodology to help identify the significant features in the CEFR/Portfolio descriptors to guide the sourcing and development of good classroom assessment tasks

Table 5.16 Further examples from the B1 task collection for listening

	CEFR descriptor	Micro-skills	Text features	Task features	Example with item types
B1	• I can generally follow the main points of extended discussion around me, provided speech is clearly articulated in standard dialect.	• Understand main/most important information • Understand main conclusions	• Series of 3–4 extracts • Perhaps from same TV talk show • Clear, standard, straightforward, relatively slow • Extremely short (c 1 min)	• Hear once only (in test) • One item per extract • No tricky distractors	• Understand main point of 3 short conversation extracts from a TV talk show • True/False/NS • Matching (really a repeated Multiple Choice Questions (MCQ) with 5–6 alternatives) • New MCQ each extract
B1	• I can listen to a short narrative and form hypotheses about what will happen next.	• Follow, though not necessarily in detail • Understand an explicitly signalled line of narrative/argument • Distinguish main point/relevant facts and information from specific details – not central to story line	• Narrative in linear order • Chain of events – consequences • Clear, standard, straightforward, relatively slow	• Hear once only • Guess what comes next (recording stops at question point) • Answer to each question on separate paper, hand in each time	• Follow a narrative and answer 5 MCQs to predict what comes next when the audio text stops • Series of MCQs – only one alternative makes sense
B1	• I can understand the main points of radio news bulletins and simpler recorded material on topics of personal interest delivered relatively slowly and clearly.	• Understand an explicitly signalled line of narrative/argument • Understand main/most important information	• Topics in field of general personal interest • Clear, standard, straightforward, relatively slow	• Hear once only • Straightforward transfer of info, in order of text	• Follow radio news and answer questions/complete the table about 4 or 5 of the stories (Total c 10 questions) • True/False/NS or Matching • Information transfer to table or open-ended questions

Table 5.16 (continued)

	CEFR descriptor	Micro-skills	Text features	Task features	Example with item types
B1	• I can catch the main points in TV programmes on familiar topics when the delivery is relatively slow and clear.	• Follow, though not necessarily in detail • Understand an explicitly signalled line of narrative/argument • Understand main/most important information • Distinguish conclusion from preceding detail	• Familiar topics regularly encountered in a school, work or leisure context • TV programmes: (interviews) short lectures, news reports • Clear, standard, straightforward, relatively slow	• Could hear twice/three times • Understanding main information • Identifying when conclusion is starting • Understanding main conclusion	• Follow a simple, factual TV/web documentary (c 5–10 mins), understand the main points and complete the table/answer open questions • Information transfer to table • Open-ended question
B1+	• I can understand a large part of many TV programmes on topics of personal interest such as interviews, short lectures, and news reports when the delivery is relatively slow and clear.	• Understand main/most important information • Understand main conclusions	• Daytime TV/local TV news – perhaps with phone-ins/studio audience • Factual interview • Straightforward factual interview questions (probably notified beforehand) • Descriptions of events/ plans • Descriptions of feelings, wishes	• Understanding advantages/ disadvantages of a plan • Understanding consequences • Understanding the opinions of the main speakers (for/against: why?)	• Listen to a straightforward, factual interview from a current affairs TV magazine programme (or local news) and understand both main points and specific details • T/F/NS • Information transfer (table) • Open-ended questions

Table 5.17 Micro-skills – collated from task collections

	A1	A2	B1	B2	C1/2
Recognise	• Idea of the content • Numbers • Familiar names, words and very basic phrases • Questions asking for something • Relevance	• (Changes in) topic • If the information is important for carrying out my tasks • The relevant section of the instructions • The main characters in a story • Dates, places and main people involved • Common abbreviations used • The relationship between places, dates and prices presented	• Useful information • New sections • The gist or general topic of article • The general impression conveyed by the text • The different steps in a description • The main stages in a process • The links between the different parts of the text	• Relevance • Changes of topic • Change of mood, emphasis • Changes to new section • Keep track of different characters	• Social/political/ historical background of text • Variation of style for effect • Complex structural patterns • Register shifts • Irony
Process with text	• Visual prompt • Numbers • Map	• Information in a picture • Map • Figures, numbers • Diagrammatic information • Images • An action to be carried out	• Diagrams (object, machine, organism) • Tables	• Diagram • Table	

Table 5.17 (continued)

A1	A2	B1	B2	C1/2
Distinguish	• Necessary contact details	• Conclusion from preceding detail • Main points from specific details • Main point/relevant facts and information from specific details – not central to story line	• Relevant information from background noise (listening) • Between factual reporting and personal point of view • Between text elements, e.g. fact and opinion, different writers' opinions • When emotional and objective • Between positions of speakers • Main points from relevant supporting arguments/details • Such supporting arguments from digressions	• Main points from supporting arguments/ sub-themes/details/ examples

Understand	• Simple message • Main point • Basic factual information • Specific details such as times, locations, dates, people • Official instructions • What a person wants	• An idea of the overall meaning • The main events • The main point • Essential information • Specifically required information • Time references (past? future?) • The chronological sequence in the events • The main point(s) being related • What the person expects of me • The procedure to follow to get tickets etc. • What action is necessary next	• Main/most important information • Significant points: *where/when/who/ what* and essential information • Specific details • The evolution of an explicitly signalled line of narrative/argument • A point of view • Main arguments • Main conclusions • Wishes/hopes/ ambitions • Writer's attitude, intention • The aim of a letter	• Main ideas/points • Specific details • Points of view, opinions • Complex lines of argument • Evidence • Speaker's feelings and mood • In detail why speaker feels this way • Mood and tone explicitly conveyed • Implicit relationship • Speakers' implicit feelings towards each other	• Main ideas/points • Relevant details • Technical phraseology • Finer points of detail • Attitude/mood/ intentions • Implicit meanings, ideas and connections • Implied, indirect and ambiguous relationships • Use of humour and irony • Implicit meanings, ideas and connections • Implied as well as stated opinions • Implicit as well as stated cultural references • Implied attitudes and relationships

Table 5.18 Text features – collated from task collections

	A1	A2	B1	B2	C1
Length	• Text length: 30–100 words • Max A5 for scanning, including pictures – target test items short, c50 words	• Text length: 150–300 words • Target text item for scanning: c100–200 words	• Long but not complex • Normally 500–600 words • Target text item for scanning 150–300 words	• Long enough extract to contain a number of steps • Long and complex enough to contain expressions of opinions supported by arguments • 700–800 words	• Lengthy, up to 1,200 words • Short/Succinct
Authenticity	• Semi-authentic/ Adapted authentic • Authentic forms • Authentic poster, magazine page or public notice • Authentic simple notices both text based and using symbols	• Edited for length but maintaining authentic language and genre structure • Authentic	• Edited for length but maintaining authentic language and genre structure • Simplified short stories • Authentic	• Authentic	• Authentic
Organisation	• Very short, simple texts and dialogues • Narrative or descriptive with news reporting style • Simple and direct headings often not in sentence form	• Explicit information • Limited number of information points • Information points are easy to spot • Direct, factual text • Very factual – describing incidents – perhaps local rather than national TV	• Presentation: straightforward and clearly structured • Explicitly signalled line of narrative • Narrative in linear order • Chain of events – consequences	• Information rich • Presentation straightforward and clearly structured • Clearly identifiable purpose • Repetition of critical information • Structured text, clearly signposted	• Clearly structured/ signposted • Multi-layered structure (opinion, commentary, quotation, example etc.) • Not clearly structured • Not explicit/implied relationships

		• Content: at least 4 different actions/ events related • Narration, explicit information, direct • Straightforward, linear text • Short, concise in a list of points, preferably numbered, not more than 6 steps/ information units • Very clear use of simpler connectors • Clear signposting and overt signalling with explicit use of linkers/ markers • Overt, clear signalling with explicit use of markers/linkers	• Clearly structured description • Well signposted • Helpful layout and headings • Not complex	• Clearly structured argument (opinion, supporting arguments, examples) • Layout and structure appropriate to genre	
Linguistic aspects	• Separate, short, basic sentences and phrases • Many proper nouns and words related to cultural events • Dates and time expressions • Familiar names, words and basic phrases • Unfamiliar names, verbs and adjectives and formulaic structures	• Short, simple texts containing the highest frequency vocabulary • High-frequency, everyday vocabulary • Use of past, present and future time • Informal (even if professional)	• Description of places, offers • Descriptions of events/plans • Descriptions of feelings, wishes	• Neutral register • Transactional • Narrative and description • Clearly expressed opinions • Argument • Justification • Variety of viewpoints • Abstract and technical vocabulary • Specialised vocabulary defined	• Formal • Colloquial/idiomatic • Propositionally complex • Linguistically complex • Technical vocabulary (but obscure items not essential or deductible from context) • Descriptive/narrative • Descriptive/ exhortative • Argumentative

Table 5.18 (continued)

	A1	A2	B1	B2	C1
	• Standardised conventions relating to machine/computer operation			in text or not essential to global comprehension	• Variety of viewpoints • Implied as well as stated opinions
Topics	• Basic personal information • Familiar topics referring to common everyday situations relevant to the home and place of work or study • People, places, weather and common holiday events • Simple, direct message involving time and/or place and set phrases	• Personal/familiar topics • Concrete, predictable, everyday topics • Guided tour of tourist site • Being shown around a new workplace • Places • Events • Plans	• Topics in field of general personal interest • Familiar topics regularly encountered in a school, work or leisure context • Personal and familiar topics • Current affairs • Professional topics • Leisure • Places • Events • Plans • Feelings, wishes	• General topics of personal or professional interest • Nonspecialist topic • Topical and/or controversial issue • Field of interest of test takers/general interest • Familiar subject matter • Technical but for general reader • Personal	• Academic content • Professional topic • Current affairs

Support	• Visual support (showing context)	• Perhaps tables of figures, or diagrams explaining figures • Visual support – illustrations, headings • Explanation, elaboration of what is being said (e.g. with video or speech)			
Spoken delivery	• Very slow, carefully articulated • Long pauses to allow assimilation of meaning • Well articulated • No background noise • Repeated	• Two speakers only • Clear, straightforward, slowly articulated • Slow, clearly articulated speech • Repetition/ clarification of directions provided through dialogue • Series of 3–4 extracts • Series of 4–5 public announcements • Circa 1 minute per extract or announcement	• Clear, standard, straightforward, relatively slow • Delivery: clear, standard, relatively slow	• Standard • Delivery adapted to environment • Clear and well articulated • Some animated discussion – speakers overlapping • Some background noise • Speech sometimes hard to catch (whispering, distance from mic)	• Non-adapted versions • Standard and non-standard varieties • Considerable degree of slang and idiomatic usage • High-speed delivery • Colloquial/Idiomatic • Turn overlap • Poor sound quality

- to identify text features, micro-skills, task features and related item types that are particularly relevant at different CEFR levels
- to provide illustrative examples of scenarios at each level for English and French, with sample assessment materials for them
- to create a sufficiently large bank of piloted tasks illustrating those scenarios to promote regular use in class in a continuous assessment 'portfolio' approach for listening and reading
- to identify a small set of the most archetypical, appropriate, effective scenarios for each level/skill and create simpler templates for them in order to assist in the rapid duplication of such assessment tasks
- to thus systematise the efforts for the assessment of listening and reading in schools that implement the EAQUALS Certificate of CEFR Achievement
- to create a network of teachers in participating schools who develop, pilot and share assessment materials, whether or not they are members of the above scheme
- to exploit the examples and templates in continuing professional development (CPD) in order to stimulate the regular use in class teaching of appropriate authentic materials with suitable tasks.

During 2013 the assessment materials for English and French were piloted and revised through a process of peer review and trialling. A guide containing content specifications like those aspects shown in Tables 5.16 and 5.17, together with illustrative scenarios, task collections and reports on trialling are available on the EAQUALS website (www.eaquals.org).

The approach being taken in the EAQUALS project is continuous assessment: the collection of scores for different types of tasks over a period of time that is used to inform teacher judgements at the end of the course. The alternative for the assessment of reading or listening is to develop a blueprint like those used in the Spanish official language schools (Figure 5.2) in order to group tasks into balanced, one-off tests which report results in terms of the CEFR level(s) concerned. However, if we take such an approach the tests that we use should go through a more formal validation process. There are two main issues here:

- how do I know my test really works properly? and
- how can I convert scores on the test properly to CEFR levels?

The former question is addressed in the next section and the issue of converting test scores to CEFR levels is treated in Section 5.3.2, with some of the implications involved in a high-stakes context discussed further in 5.3.3.

Section 2: More details

5.2.1 Quality control and statistical validation of tests

Many of us feel that we do not have the resources or expertise to check through statistical validation that a test is functioning as planned. However, anyone can do basic quality control that will ensure that the test is more valid and more reliable. As mentioned when discussing validity: the crucial issue is to get things right in the first place. The following list of suggestions may help in doing that:

- Check with colleagues that the source text appears to be the right level.
- Check that the tasks are realistic and appropriate for the level.
- Check that the learners will be familiar with the type of text, the tasks and the item types.
- Check that learners are informed of what is expected; the criteria or marking key has been explained to them as well as the type of tasks.
- Check that each item type has clear, unambiguous instructions.
- Check that the questions accompanying a text are in considerably simpler language than the text itself and do not themselves constitute a reading test item!
- Check that questions cannot be answered from general knowledge without reading or listening to the text.
- Ensure that questions come in the same order as in the text and that candidates have time to read them before tackling the task if they wish.
- Ensure that the candidates will have time to write/mark the answer in a listening test.
- Ensure that the marks awarded for each task relate to the proportion of time and effort spent on it.
- Try and have a total score of 30–40 marks for your test. Tests with more marks are more reliable. If you want to have a shorter test for practical reasons, then have some information transfer items scored 012 or even 0123, rather than just 01 (right/wrong).
- Define procedures very clearly and prominently. How long can people take? Is the tape played once or twice? Are dictionaries allowed?
- Ensure that everybody uses strictly the same marking key.
- Ensure that you have a defined way of determining whether other answers that markers may suggest are acceptable or not.
- Try out the test yourself. With a listening test, ensure you then use the recording and do not just read a script.
- Pilot the test with a couple of classes under full test conditions.
- In piloting, tell markers to note any borderline answers they think may be acceptable and to record the name or number of the respective answer sheets. Then make a final decision at the end of the marking, and afterwards alter the individual scores and the future key accordingly.

In addition to procedural quality controls such as those suggested above, one can do a lot with very simple statistical techniques. The main problem is the time needed to enter the data, because one needs the score of each learner on each item. Thereafter, with today's analysis programs it is really very easy. The ALTE test development manual explains basic statistical techniques in very simple language and recommends programs and further reading (Association of Language Testers in Europe 2011:75–78). The most important results, which will be reported by any simple analysis program, are the following:

- reliability coefficient (Cronbach's alpha)
- facility values
- discrimination index
- correlation coefficient.

5.2.1.1 Reliability coefficient

The important thing to know about reliability coefficients is that a wide spread of levels, a single item type and a lot of marks (e.g. 100) will give a higher reliability coefficient. With a placement test one should thus be aiming for a coefficient above 0.95; with a task-based listening or reading test of 25 marks aimed at one CEFR level one would be satisfied with 0.80. The best way to influence reliability is to ensure appropriate design and content, solid instructions and consistent marking. Afterwards, reliability can only be improved by getting rid of badly performing items (see below) or by making a longer test.

5.2.1.2 Facility values

Test items collect most accurate information when the learner is getting 50% right and 50% wrong. This is the reason why tests in the 1960s used to be not only very long but also very difficult. Any items with a value below 0.2 (= fewer than 20% answering correctly) or above 0.8 (= over 80% answering correctly) could be eliminated. However, one is more loath to remove the latter and so might keep items with a value of up to 0.95.

5.2.1.3 Discrimination index

The discrimination index reports how well the items separate the people who are strong in the construct being tested (the top third) from those who are a lower level (the bottom third). A low result means that a lot of the lower group are getting the item correct. That means that the item is not actually testing what you want it to; something else is getting in the way. That something could be a different skill, the effect of an item type, or simply an item that the candidates have difficulty understanding. The discrimination index helps us to identify such items in order to remove them. If we cannot pretest in order to identify these items before administering the test for real, then they

should be excluded from the data during the analysis, *before* reporting results. If the test is a very easy one for the group – that is if you kept items with facility values of above 0.90 and have a high mean score of 80% or more as a result, then the point-biserial correlation is more appropriate.

5.2.1.4 Correlation coefficient

With a correlation coefficient we might compare results on our test to results in a CEFR-linked examination, using the examination as an *external criterion*. The correlation coefficient is very simple to calculate by putting the scores from two measures into two columns of a Microsoft Excel table, each row being a candidate. Then you use the CORREL or PEARSON function – these are identical – stipulating the range of rows concerned in each of the two respective columns (Insert/More functions/Statistical). With very small numbers – under 30 – one should use a ranking coefficient – the Spearman coefficient, but unfortunately this is not available in Excel. However, any correlational evidence to an external criterion based on fewer than 30 cases is hardly convincing anyway.

As with the reliability coefficient, a wide spread of levels will lead to a higher correlation because the scores of the candidates are more spread out. In practice correlations between two different tests are very rarely much above 0.80, even when testing across a range of levels, unless the comparison is between scores on two forms of the same test which have been deliberately developed with identical content in order to be interchangeable versions. In such cases correlations should be well above 0.90.

5.2.2 Moderating teacher assessment

The advice given in the previous section related to listening and reading tests. This section is concerned with the quality of the assessment of spoken and written language. A staffroom of teachers working with a CEFR-based curriculum may well have a similar interpretation of the levels, since learners, classes and materials are all referred to in terms of those levels. Provided that this interpretation was based on an engagement with the CEFR guided by the illustrative descriptors and illustrative samples made available for that purpose, as opposed to being merely a convenient relabelling of pre-existing course levels, then the interpretation of the CEFR levels by the majority of the teachers can be really very accurate. However, even after a good assessment criteria grid has been developed, and even after standardisation training in both the CEFR levels and in the use of that assessment grid has been implemented, the reliability of the judgements made by the teachers can still remain a problem. Standardisation training improves assessors' consistency of judgement and it reduces extreme lenience and severity. However, some assessors can be quite resistant to training and the effects of the

standardisation also start to wear off immediately after the training. The sad fact of the matter is that training cannot change 'hard cases'.

In addition, there are particular problems when teachers implement their – possibly very accurate – internalised understanding of the CEFR level(s) in their assessments of actual individual learners. Here lots of classic rater errors come into play. These include, for example: ignorance of the criteria they are supposed to apply; use of personal criteria rather than the intended ones; unconscious excessive focus on one criterion (e.g. linguistic accuracy; pronunciation with a particular L1 influence); refusal to give the top grade, and stereotyping (assumption that because a learner comes from a particular place or social, linguistic or cultural background, that their competence is of a certain type).

The traditional measure of the reliability of assessments by raters is called inter-rater reliability (IRR). This is usually reported as a correlation, which can also be created with the CORREL function in Microsoft Excel. Weir (2005a:199–201) and the Association of Language Testers in Europe (2011:79–80), among others, discuss this in more detail. Assessment with well-trained assessors can achieve inter-rater reliability correlations over 0.9 when using the same technique and criteria grid as demonstrated with the American ACTFL scale (Dandonoli and Henning 1990). However, this is the exception rather than the rule. Alderson et al (1995:132) suggest that we could be happy with IRR correlations of 0.80. However, even a very high IRR coefficient doesn't tell us that the raters agreed whether learners were B1, B1+ or B2. It just tells us that they tended to put them in the same order. Furthermore, in order to report an IRR estimate all the raters have to rate all the candidates, which in our operational contexts is simply impossible. Therefore it makes most sense to implement a form of collective assessment to improve reliability, for example:

a) Double marking: use two assessors who rate individually and then negotiate the final grade. Here it is best if the second assessor has a wider focus (knowing all the levels) to counteract the narrower focus (knowing all the learners) that may cause a class teacher to exaggerate differences in the class.
b) If it is not feasible to have a second assessor with experience of all the levels, get the class teacher to assess all the learners first and create a list of results in rank order. Then have a second assessor rate the strongest, weakest and middle learner, plus the learner halfway between top and middle and between middle and bottom. This forms a 'structured sample', at the hundredth, seventy-fifth, fifth, twenty-fifth and first percentile, of the full supposed range of proficiency in the class. Second assessment of such a sample can show whether the teacher is exaggerating the level of those at the top and bottom of the class, as well as checking for overall strictness/leniency.

c) If that is too complex, then systematically double-rate every fifth or tenth learner.
d) Alternatively, assess the first three to five speaking candidates in a team of teachers and then, once 'tuned in', split up to rate the other candidates individually. This approach is even easier with writing. The three to five scripts that were 'benchmarked' at the beginning can be photocopied and kept available for reference and comparison during the session, and future sessions.

If one *can* conduct a statistical check, a far better technique than IRR is available. It doesn't require all raters to rate all learners, but it does require a linked data set. If all the raters independently assessed the same three to five learners at the start, then those common candidates create a linked data set. If one person is always the second assessor, then that second assessor provides the link. If in a large-scale operation, two or three second assessors first rate three classes together before splitting up to second-assess classes individually, then the first three classes provide the link. With data linked into one chain in this way, we can use the Multi-Faceted Rasch model (Linacre 1989, 2008). The great advantages of this method are that it identifies inconsistent raters and in addition it adjusts for severity and thus ensures an objective result. It was used in the Swiss research project (North 2000a:176–178; 208–230), it is recommended by Weir (2005a:199–200) and has a downloadable, user-friendly guide provided by Eckes (2009).

Finally, the most obvious way to reduce subjectivity is to support teacher assessment with tests. Even if a test is not formally referenced to the CEFR it is difficult to give a learner a low grade when they got one of the highest test scores. Another possibility is to use a standardised test that has been linked to the CEFR for what used to be called *statistical moderation*. In the 1970s and 1980s, Cambridge ESOL used to moderate the results from the subjectively marked papers (interview and writing), corrected in those days by a single marker, with the results from the more reliable Use of English paper. This approach can also be used without statistics. Eurocentres uses tests of language usage, targeted at the level, which are produced from a calibrated, validated item bank. When we are finalising grades we are aware of the grade previously reported by the test. For any individual learner the relationship between their language usage and their communicative language performance will be limited. However, there will be a relationship between the *groups of scores* for whole classes. Given a record sheet showing both the test results and teacher assessments, we can *eyeball the columns* and spot whether any particular person is awarding their class grades that are systematically above those from the test (= being too lenient) or systematically below those from the test (= being too strict). The advantage of this approach is that it can be applied in circumstances in which an accurate interpretation of CEFR

levels has not yet been fully achieved through curriculum development and standardisation training, without the implementation of the Multi-Faceted Rasch model discussed above.

5.2.3 Relating norm-referenced teacher assessment to CEFR levels

In the previous section we discussed moderating teacher assessment of CEFR levels. However, such criterion-referencing is not the only kind of assessment that we do. Tests and the grades and comments awarded from them are also a way of motivating learners to revise and learn the content that has been taught. Giving evaluative feedback (e.g. Excellent; Very Good; Poorly finished, etc.) is a natural part of this process. Such assessment in relation to what we expect is *norm-referenced* assessment. The standard of performance that we expect from the group concerned is the *norm*. Norm-referencing is *assessment relative to one's peers*. The peers may be the rest of the class/year. More usually, however, the norm is the average achievement of equivalent groups of learners at the same time in the school year over many years. Sometimes secondary school teachers in particular may find themselves in a situation in which such norm-referenced class grades need to be linked to CEFR levels that have been set as a standard. How can one do this systematically?

In Swiss schools the teacher awards for each assessment a grade between 1.0 – truly awful – and 6.0 – truly excellent. These grades are then summed for each subject and aggregated across subjects into a global grade. Half grades (e.g. 4.5) are an essential part of the system and finer gradations are sometimes used. 4.0 is the norm – the minimum standard expected of the learners. If a realistic achievement standard has been set for the school year concerned, then the process of fixing the norm to that standard is in itself simple. If B1 is an official achievement standard that the teaching body accepts as realistic, then a learner needs to be B1 to get a grade 4.0. The complications are things like:

- How good do you have to be to get top grade (6.0)?
- How does the norm progress over time towards the standard? If you have the class for two years with two semesters per year, how does the norm develop across four test points?
- How does one deal with underachievement in relation to an unrealistic standard?

These are questions that can only be resolved through discussion. If the standard set is B1, is it reasonable to expect people to be B2 in order to get top grade (6.0)? As with any rating scale for an examination, it is important to answer this question and fix the minimum standard for the top grade. The

answer depends on how reasonable or 'aspirational' the B1 achievement standard is. Was this standard set as a result of monitoring the progress students actually make? Also, how great is the spread of proficiency level among the type of learners concerned? This can be very considerable in a secondary school classroom.

Table 5.19 shows one solution to this problem based on examples used by Hanspeter Hodel and Oliver Töngi on an intercantonal modular CEFR training course held in 2008 and 2009. It assumes, for the sake of simplicity, that you do need to be the next level – here B2 – in order to get a 6.0, because the standards in most cantons in Switzerland were set following a concept informed by the research results from the Swiss research project, as mentioned in Section 3.3.2. Table 5.19 also employs grades between plus levels and criterion levels (e.g. Good B1+). This represents a really good B1+ that still does not fulfil the criterion for B2. This grade is unlikely to be defined with a descriptor, but having such non-defined intermediary grades is not a problem in practice, provided that the adjacent grades above and below it are properly defined with CEFR-based descriptors.

Table 5.19 Teacher grades and CEFR levels over time

Test point	3.0	3.5	4.0	4.5	5.0	5.5	6.0
Mid-year – Year 1	A1	A2	Good A2	A2+	Good A2+	B1	Good B1
End year – Year 1	A2	Good A2	A2+	Good A2+	B1	Good B1	B1+
Mid-year – Year 2	Good A2	A2+	Good A2+	B1	Good B1	B1+	Good B1+
Final – end Year 2	A2+	Good A2+	B1	Good B1	B1+	Good B1+	B2

If the B1 standard was in fact 'aspirational' for the group concerned and unlikely to be reached by 80% of the learners, then all cells could be shifted to the right, for example by one cell. Then to get a 4.0 (= minimal standard) in the final assessment shown in the bottom row, instead of achieving B1, a learner would only have to be good A2+. Achieving the official standard B1 would then give a more respectable 4.5. Finally, in order to get the top mark, 6.0, one would only have to demonstrate a strong B1+: a truly excellent result at the official standard set, but not the next level.

The approach described above, whilst it deploys a simple logic, is quite sophisticated in that it includes a mechanism to bend the official achievement standards that have been set to the reality of the educational situation, without compromising the integrity of the interpretation of the CEFR levels. There is of course a very real danger that, with less conscientious teachers, the local standard (4.0) that is supposed to be B1 could just be

deemed to be B1, reported as B1 and – over time 'become' B1 as far as the local pedagogic community was concerned. This underlines the fact that CEFR-based teacher assessment will only function if there is regular standardisation training and if there are moderation techniques in place, as discussed in Section 5.2.2, in order to ensure that the relationship to the CEFR is maintained over time.