COUNCIL OF EUROPE
CONSEIL DE L'EUROPE
Language Policy
Politiques linguistiques

The CEFR Grid for Speaking

www.alte.org

The CEFR Grid for Speaking Tests is designed to assist language test developers in describing, creating and reviewing tests of oral production. It is intended to stimulate critical reflection amongst those involved in the preparation and evaluation of speaking tests, and to facilitate precise reporting by testing bodies participating in audits carried out by organisations such as the Association of Language Testers in Europe (ALTE).

In the case of ALTE audits, the Grid will be useful in building arguments related to Minimum Standards 1 (theoretical construct), 4 (comparability of test versions), and 5 (alignment to the CEFR framework). Responses to the Grid may be used as supporting evidence for claims made during the auditing process.

Part 1 of this document contains 34 questions about different aspects of the speaking test as a whole and its individual speaking tasks. Some of these questions are preceded by an asterisk. These questions will be accompanied by explanatory notes, which are meant to indicate why the questions are important to the completion of the Grid. The explanatory notes will be found in Part 2 of the document.

The Grid was developed by the ALTE CEFR Special Interest Group (SIG). It contains contributions by Bart Deygers, Beate Zeidler (editors), Dianne Wall (final editing), Lyan Bekkers, Inmaculada Borrego, Michael Corrigan, Henna Tossavainen and other SIG members.

ALTE CEFR SIG, February 2014

PART 1 – THE SPEAKING TEST AS A WHOLE AND COMPONENT SPEAKING TASKS

A THE SPEAKING TEST AS A WHOLE

1. GENERAL INFORMATION

1	Name of test provider	
2	Name of test	
3	Target language	
4	*Date of last test revision	
5	Number of tasks in the speaking component	If there is more than one speaking task, complete Section B for each task.
6	Duration of the speaking test as a whole	Speaking test duration: approximately ____ minutes This includes ____ minutes of preparation time.
7	*Target CEFR level of the speaking test	o A1 o B1 o C1 o A2 o B2 o C2
8	*Channel for test delivery	o Face to face, and recording o Face to face only, in real time o Audio only, in real time (e.g. telephone conversation) o Audio recording o Video only, in real time (e.g. Skype video) o Video recording (e.g. in web-based test)
9	Test content	o General language proficiency test o Language for Specific Purposes test
10	*Test construct	o It is possible to specify the construct(s) that underlie the test. o It is not possible to specify the construct(s) that underlie the test.
11	*Intended use (CEFR, p. 183)	o Achievement (or progress) test o Diagnostic test o Placement test o Proficiency test o Other use - please specify:
12	*Target population characteristics	o Known o Unknown

2. RATING

13	*Rating method	o Holistic o Analytic: band descriptors o Analytic: checklist o Other (please specify):
14	Rating criteria (Tick all that apply)	o Argumentation o Cohesion and coherence o Content o Interactive communication o Grammatical accuracy o Grammatical range o Lexical accuracy o Lexical range o Pronunciation o Other (please specify):
15	Raters	o Machine marking o Manual marking, using _____ raters o Combination
16	Is there a procedure in place in case raters disagree?	o Yes Specify: o No
17	Are the rating criteria available to the test-taker?	o The criteria are available on the test paper. o The criteria are available elsewhere. Specify: o No

3. FEEDBACK

18	Quantitative feedback for test-takers	o CEFR level o Test-specific grade o Pass/fail only o Other (please specify):	o Percentage score o Ranking (e.g. quartile) o Raw score
19	Qualitative feedback	o Yes, general feedback o Yes, specific feedback based on criteria o No qualitative feedback	

B COMPONENT TASK/S

To which speaking task does this information relate?	Please fill in this section for each component task.

1. GENERAL TASK CHARACTERISTICS

20 *Task topic

21 Language of instructions

22 Other language used

23 Task duration

Task duration: approximately ____ minutes

This includes ____ minutes of preparation time

24 Is the performance recorded?

- o Yes, audio only
- o Yes, video
- o No - face to face only

25 *Control/guidance by the task rubric

- o Rigidly controlled
- o Partially controlled
- o Open format

2. INSTRUCTIONS & PROMPT

26	Task instructions (Tick at least one.)	o Via pictures o Spoken (recorded) o Spoken (real time) o Written
27	*Language level of task instructions	o Below target level o Same as target level o Above target level
28	Type of prompt (Tick at least one.)	o Audio o Oral only (real time by examiner) o Picture/drawing/icon o Text o Video
29	*Interaction required	o Interaction with examiner o Interaction with other test-taker(s) o Interaction with recorded prompts o Monologue
30	Discourse type required	o Discussion/conversation o Interview o Speech, presentation o Story telling / narration o Question and answer o Other (please specify.):

3. EXPECTED RESPONSE

31	Response type	o Short monologue (i.e. words and phrases) o Extended monologue (i.e. formal speech) o Short interaction (i.e. words and phrases) o Extended interaction (i.e. presentation with questions and answers)
32	*Integration of skills	o None o Reading Rated? yes/no o Writing Rated? yes/no o Listening Rated? yes/no
33	*Communicative purpose	o Referential (telling) o Emotive (reacting) o Conative (argumentation, persuasion) o Phatic (social interaction)
34	Expected rhetorical function(s)	o Argumentation o Complaint o Description o Explanation o Instruction o Persuasion o Report o Summary o Other (please specify.):
35	Expected register	o Informal o Neutral o Formal
36	Expected level of response	o A1 o B1 o C1 o A2 o B2 o C2

PART 2 – EXPLANATORY NOTES

What follows are explanatory notes for some of the questions found in Part 1. The notes are intended to indicate why these questions form part of the Grid and how they may be helpful to test developers.

Question 4. Date of the last test revision

Changes may have been made in the speaking test since it was originally launched. These changes may seem minor, but even small changes can alter the nature of what is being assessed. This question encourages test developers to think about whether all the changes have been well grounded or whether there are parts of the test where the reasoning behind the changes is not clear.

The following questions may be useful in this process:

- Does the revised test present a different definition or operationalization of the speaking construct?
- Has the purpose of the test changed over time?
- Have there been any changes in the size or nature of the test-taking population?
- Have the nature, definition or weighting of the assessment criteria been altered?

If any revisions have been made to the test, the test developers should be aware of the nature and background of these changes. They should also be able to determine whether the changes have had the desired effect.

7. Target CEFR level of the speaking test

This speaking grid was developed by the ALTE CEFR Special Interest Group (SIG) and consequently refers to the Common European Framework of Reference for Languages quite often.

It is important to know what level a test is supposed to be at when deciding on the input material, rhetorical functions, discourse types and so on. In order to judge whether the test is really at the intended level, it is important to carry out an alignment procedure. The following publications will provide useful guidance:

Manual for Relating Language Examinations to the Common European Framework of Reference for Languages (CEFR)
Download: http://www.coe.int/t/dg4/linguistic/manuel1_en.asp#Manual

Martyniuk, W. (2010). *Aligning Tests with the CEFR: Reflections on Using the Council of Europe's Draft Manual*. Cambridge: Cambridge University Press.

It is of course possible to develop a high-quality language test without linking it to the CEFR. Test developers who follow this route may still find the above publications useful.

8. Channel

A speaking test can be administered through different channels (means). The choice of channel is determined by the construct, context and purpose of a test. Some channels may seem more realistic or authentic than others, but the final choice is often determined by practical considerations such as costs and technical possibilities.

The channel influences how the test is administered (e.g. should longer responses by recorded?) and marked (how many examiners are required to ensure reliability). It may also influence the test-taker's performance and motivation.

Each test is different, so it is impossible to offer one-size-fits-all solution when it comes to choosing a testing channel.

10. Construct

The term 'construct' refers to the theory underlying the design of a test – that is, the way the test developers define language ability or the particular aspects of language they are assessing.

It is important for test developers to be explicit about their test construct, as the choice of construct will affect the decisions they make about the content of the test, the types of tasks they give their test-takers, the weighting of different components of the test, the marking criteria, and the boundaries between different levels of ability.

There are different ways of defining language ability. Some experts see language in abstract terms, describing, for example, the competences that test-takers need in order to produce the right kind of language: linguistic competence, sociolinguistic competence, pragmatic competence etc. Others see language in terms of the skills that test-takers need to display. These experts might, for example, look at the skill of speaking and break it down into different sub-skills. Another way of viewing language is in terms of 'can do' statements – e.g. the test-takers can express simple opinions or requirements in a familiar context.

Test developers may decide to base their tests on any of these constructs or on others that they find in the relevant literature. They may also wish to use a combination of constructs, depending on the purpose of their test.

11. Intended use

Sometimes the purpose of a test changes over the years, and the purposes for which it is now used do not match the originally intended purpose. This change of purpose may cause methodological, ethical or operational problems. It is important to monitor whether the current use corresponds to the intended use, and what the effect of such a shift may be.

Possible test purposes include:

- Achievement tests, sometimes called progress tests, which measure what students have learned. 'The content … is generally based on the course syllabus or the course textbook'.
- Diagnostic tests, which 'seek to identify those areas in which a student needs further help'.
- Placement tests, which are 'designed to assess students' levels of language ability so that they can be placed in an appropriate group or class'.
- Proficiency tests, which 'are not based on a particular language programme. They are designed to test the ability of students with different language training backgrounds.' (Alderson, Clapham & Wall 1995, pp. 11-12)

12. Target population characteristics

It is not possible to determine whether a test works as it was intended to unless there is a match between the people who actually take the test and the people for whom it was designed. This is especially important when we talk about test content. It would not be fair, for example, to give a test that was designed for the world of work to a group of schoolchildren. The schoolchildren might have the linguistic ability to answer the questions but not the necessary subject knowledge.

Target population characteristics are also relevant for test statistics, because any sample of the population that you use (e.g. for pretesting) should be representative of the whole of the population.

Target population characteristics that are often analysed include age, gender, level of education, occupation, or type of motivation.

13. Rating method

In a holistic approach, the test-taker's performance is judged as a whole. The rater does not give separate scores for different features of the performance, such as grammatical control, vocabulary etc.

In an analytic approach the rater gives separate scores for several different language features. This approach recognises that a test-taker's grammar, for example, may be very good, but his/her vocabulary may be weaker.

Analytic raters may use a scale for each language feature, or they may use checklists. Scales may take many forms – e.g. 1 to 9, A to E, A1 to C2. Checklists are often binary – does the test-taker master a particular feature or not?

It has been claimed that the holistic approach more closely resembles how language production is judged in real life, and can be quicker than using an analytic approach. However, analytic marking can offer richer diagnostic information for L2 learners.

There are mixed results from research into the reliability of using the two approaches.

20. Task topic

Different topic choices are possible, depending on the target language use domain, the target population characteristics, and the target language level.

Topics can be classified in different ways. The CEFR (p. 52) presents one influential scheme, which lists fourteen general categories. These categories can be further sub-divided to suit the purpose of the test.

Topics can become more abstract and more complex as the target language level grows more demanding.

By including broad topic categories in the task specifications, the thematic focus of the test can be maintained from one test administration to the next.

25. Control/guidance by the task rubric

In rigidly controlled tasks the task determines the structure of the test-taker performance, leaving no room for spontaneous interaction. Partially controlled tasks may present a scenario in which the main conversational path is outlined, leaving some room for spontaneous interaction. Tasks with an open format may depend entirely on the interaction between the examiner and the test-taker or may require the test-taker to produce a monologue.

Rigidly controlled tasks may seem inauthentic at times, but they make it easier to compare test-taker performances. Open tasks may seem more authentic, but it can be more difficult to assess the resulting interaction.

27. Language level of task instructions

Understanding instructions is a prerequisite for adequate task performance. It is paramount that the instructions be clear and easy to follow. Vagueness should be avoided at all cost and the lay-out should be clear.

If possible, the language in the instructions should be simpler than the language the test-takers are expected to produce. In CEFR terms, instructions should preferably be one CEFR level below the desired level of performance. In some cases the instructions may be written in the test-takers' first language.

29. Interaction type required

Once test designers have decided on the features of speaking they wish to assess (their construct), they need to think about the types of tasks that will elicit those features. One test may need several interaction types to cover one construct.

If, for example, the construct requires tasks that assess a test-taker's ability to use formal language during a long turn, a monologue might be a suitable interaction type. If, on the other

hand, a construct includes tasks that assess whether a test-taker can respond quickly and spontaneously, a dialogue could be the best alternative.

32. Integration of skills

Test-takers' speaking scores may depend not only on their speaking skills but also on their other skills. These other skills may include reading (e.g. skimming a text to comment on it), writing (taking notes while conducting a telephone call), or listening (understanding an audio prompt).

Test developers may consciously chose to integrate other skills with speaking or they may chose to assess speaking alone. The choice depends on the construct underlying the test. If the speaking required in the target language use situation involves other skills, then it may make sense to design test tasks that involve these skills. The test developer should be aware of the problems of 'construct-irrelevant variance' however, where the test-takers' ability in the other skills may affect their speaking performance unintentionally.

33. Communicative purpose

Specifying the communicative purpose of a task is important, both for the test developer and for the test-taker. The communicative purpose should be in line with the test specifications, since it helps to control a task's difficulty and allows for rating criteria that focus on the most valid aspects of a task. For the test-takers, being aware of the main communicative purpose is vital, since different communicative purposes require very different skills.

A task with a referential communicative purpose, for example, might require a test-taker to summarise a lecture by rephrasing the main and supporting ideas in a structured way. Alternatively, the test-taker could be asked to agree or disagree (emotive), add a convincing personal assessment of the input material's content (conative), or engage in meaningful conversation about the lecture (combination of referential, conative, emotive and phatic).

34. Expected Rhetorical function

By keeping track of the expected rhetorical functions, the test developer will be able to compare each new test version to previous versions and to the original speaking construct of the test. This may decrease the risk of construct irrelevant tasks and will increase the comparability across test versions.