

[www.alte.org](http://www.alte.org)

**The CEFR grid for speaking**

**Language Policy Division**

This document is designed to assist language test developers in creating, reviewing and describing speaking tests. Its purpose is to stimulate critical reflection and to facilitate precise reporting in case of an ALTE audit. As such, this document may be used as supporting evidence in the ALTE auditing process.

The first section of the speaking grid contains 34 questions about different aspects of the test and the test tasks. The second section contains explanatory notes belonging to specific questions.

This speaking grid was developed by the ALTE’s CEFR Special Interest Group and contains contributions by Bart Deygers (Ed.), Beate Zeidler (Ed.), Dianne Wall (Ed.), Henna Tossavainen, Inmaculada Borrego, Lyan Bekkers, Michael Corrigan and the SIG members.

**A TEST**

*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, consider duplicating section 2 | | |
| 6 | Total duration of the whole speaking test | Speaking test duration: approximately \_\_\_\_ minutes  Of which \_\_\_\_ minutes preparation time. | | |
| 7 | \*Target CEFR level of the speaking test | * A1 * A2 | * B1 * B2 | * C1 * C2 |
| 8 | \*Channel | * Face to face, recording * Face to face only: real time * Audio only: real time (e.g. telephone * Audio recording (e.g. audio recording) * Video only: real time (e.g. Skype video) * Video recording (e.g. in web-based test) | | |
| 9 | Test content | * General proficiency test * Specific purpose (Language for Specific Purposes): | | |
| 10 | \*Test construct | * It is possible to specify the construct(s) that are fundamental to the test * It is not possible to specify the construct(s) that are fundamental to the test | | |
| 11 | \*Intended use  (CEFR p. 183) | o Achievement test  o Diagnostic test  o Placement test  o Proficiency test  o Progress test  o Other use, please specify: | | |
| 12 | \*Target population characteristics | o Known  o Unknown | | |

*2. RATING*

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

*3. FEEDBACK*

|  |  |  |  |
| --- | --- | --- | --- |
| 18 | Quantitative feedback for candidates | * CEFR level * Exam-specific grade * Pass/fail only * Other: | * Percent score * Ranking (e.g. Quartile) * Raw score |
| 19 | Qualitative feedback | * Yes: specific feedback based on criteria * Yes: general feedback * No qualitative feedback | |

**B TASK**

|  |  |
| --- | --- |
| To which speaking task does the information relate? |  |

*1. GENERAL TASK CHARACTERISTICS*

|  |  |  |
| --- | --- | --- |
| 20 | \*Task topic: | See CEFR p52 |
| 21 | Language of instructions: |  |
| 22 | Other language used: |  |
| 23 | Task duration | Task duration: approximately \_\_\_\_ minutes  of which \_\_\_\_ minutes preparation time |
| 24 | Is the performance recorded? | * Yes: Audio only * Yes: Video * No: face to face only |
| 25 | \*Control/guidance  by the task | * Rigidly controlled * Partially controlled * Open format |

*2. INSTRUCTION & PROMPT*

|  |  |  |  |
| --- | --- | --- | --- |
| 26 | Task instruction (Tick at least one) | * Pictorial * Spoken (recorded) * Spoken (real time) * Written | |
| 27 | \*Language level of task instruction | * Below target level * Equal to target level * Above target level | |
| 28 | Type of prompt  (tick at least one) | * Audio * Oral only (real time by examiner) * Picture/drawing/icon * Text * Video | |
| 29 | \*Integration of skills | * None * Reading * Writing * Listening | Rated? yes/no  Rated? yes/no  Rated? yes/no |
| 30 | \*Interaction type | * Interaction with examiner * Interaction with other candidate(s) * Interaction with recorded prompts * Monologue | |
| 31 | Discourse type | * Discussion/conversation * Interview * Speech, presentation * Story telling / narration * Q&A * Others (please specify): | |

*3. EXPECTED RESPONSE*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 32 | Response type | * Short monologue (i.e. words & phrases) * Extended monologue (i.e. extended speech) * Short interaction (i.e. words & phrases) * Extended interaction (i.e. extended speech) | | |
| 33 | \*Communicative purpose | * Referential (telling) * Emotive (reacting) * Conative (argumentation, persuasion) * Phatic (social interaction) | | |
| 34 | Expected rhetorical function(s) | * Argumentation * Complaint * Description * Explanation * Instruction * Persuasion * Report * Summary * Other (please specify): | | |
| 35 | Expected register | * Informal * Neutral * Formal | | |
| 36 | Expected level of response | * A1 * A2 | * B1 * B2 | * C1 * C2 |

**Speaking Grid - Explanatory Texts**

*The purpose of the Speaking Grid is to raise awareness and to provide a starting point for reflection on various aspects of a speaking test. When preparing an audit, this will help to build an argument in relation to Minimum Standard 1 (theoretical construct), 4 (comparability of versions) and 5 (alignment to framework). When developing or reviewing a speaking test this speaking grid will help the test developer to ask the right question and to think about the test in a critical and objective manner.*

*The following texts will explain in what way some of the questions may be helpful. They are arranged in the order of the questions in the grid. An asterisk with the question indicates that a comment on the question is included here.*

**4. Date of the last test revision**

Perhaps the speaking test has changed since its original launch. Sometimes impromptu seemingly small changes can alter the nature of a test however. This question encourages the test developer to take on a diachronic perspective so as to find out if all changes have been well grounded or if there are parts of the speaking test that are remnants from some previous test with no explicit grounds.

The following questions may be useful when determining whether some further validation is required:

* Does the revised test reflect a changed perspective on the definition or operationalization of the speaking construct?
* Has the purpose of the test changed in time?
* Have there been any changes in the size or nature of the test-taking population?
* Has the nature, the definition or the weighing of the assessment criteria changed?

If any revisions have been made to the test, the test developer should be aware of the nature and background of these changes. It should also be known whether the changes have had the desired effect.

**7. Target CEFR level of the speaking test**

Even though the CEFR level of different speaking tasks may fluctuate within one and the same test because of the input material or the level of processing it requires, many tests are aligned with a CEFR level.

It is important to know what level this is when deciding on the input material, rhetorical functions, discourse types and so on. If the CEFR level of the test is unknown, it may be desirable to undertake an alignment procedure. The following publications may prove useful in this respect:

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

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>

**8. Channel**

A speaking test can be administered through different means or channels. The choice of channel is determined by the construct, context and purpose of a test. In turn, the channel influences the exam conditions and may influence a test taker’s performance and motivation.

The use of a specific channel depends on the situation. Some channels are more realistic or authentic than others, but the context of the exam and practical aspects such as costs and technical possibilities often determine the final choice.

Since each test has its idiosyncrasies, it is impossible to offer one-size-fits-all solutions when it comes to the testing channel. A test developer may want to consider questions arising from the choice of channel, such as whether longer answers should be recorded, or how many examiners are required in order to ensure reliability.

**10. Construct**

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

It is important for test designers 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 candidates, 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 learners 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 candidates 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 candidates can express simple opinions or requirements in a familiar context.

Test designers may decide to base their tests on any or a combination of these constructs, or any others that can be found in the relevant literature.

**11. Intended use**

Sometimes, over the years, the actual purpose of a test starts to differ from the original intended use. If a test is used for a purpose it was not originally created for, this may cause methodological, ethical problems or operational issues. For that reason it is important to monitor whether the actual and intended use of a test still correspond and what the effect of a shift may be.

Possible test purposes (CEFR, 183) are:

* Achievement tests, which “are similar to proficiency tests, but they are given at the end of a course. The content […] is generally based on the course syllabus or the course textbook”. (Alderson 1995: 286)
* Diagnostic tests, which “seek to discover those areas in which a student needs further help”. (Alderson 1995: 287)
* Placement tests, which are “designed to assess students’ levels of language ability so that they can be placed in an appropriate group or class”. (Alderson 1995: 292)
* 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 1995: 293)
* Progress tests, which are given at various stages throughout a language course to see what the students have learned”. (Alderson 1995: 293)

**12. Target population characteristics**

It is not possible to determine whether a test works unless there is a match between the actual people who take it and the people for whom it was designed. This is important in relation to test content because the test population should have a fair chance to be able to deal with it, so questions related to the world of work might contain an additional non-linguistic difficulty if used in a test for schoolchildren.

The target population characteristics are also relevant for test statistics, because any sample of the population that you use eg. for pretesting should be representative of the whole of the population.

Typical population characteristics are level of education, gender, age, occupation, motivation to take the test etc.

**13. Rating method**

In a holistic approach, a performance is judged as a whole. Holistic scales assume that skills develop evenly and are ideally suited for judging one all-encompassing ability.

Analytic rating scales on the other hand take into account separate features of language, such as grammar, vocabulary and structure and do not judge the performance as a whole.

Analytic scales can be further subdivided according to the judgment a rater is to make. If the rating scale is an analytic *checklist*, the rater is asked to judge a performance according to a series of criteria which are scored in a binary way. Band descriptor scales on the other hand are not binary and share their architecture with band descriptors similar to those found in the CEFR scales.

To date, the effects of employing a holistic or an analytic rating scale on the reliability of the rating process have been researched with mixed results. Analytic scales do appear to offer richer diagnostic information for L2 learners and holistic scales have been shown to be more authentic and are quicker to use than analytic rating scales.

**20. Task topic**

This question applies to the broad topic of the speaking task at hand. Depending on the target language level, the target language use domain and the target population characteristics, different topic choices are possible.

Topics tend to become more abstract, more complex and less immediately relevant as the target language level grows more demanding. The CEFR (p. 52) lists fourteen general themes that can be tapped into when designing communicative test tasks at the threshold level.

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**

In rigidly controlled tasks the task determines the structure of the candidate performance without any room for spontaneous interaction. Partially controlled tasks have a scenario in which the main conversational path is outlined, leaving some room for spontaneous interaction. Tasks with an open format on the other hand depend entirely on the interaction between the examiner and the examinee.

Rigidly controlled tasks may be inauthentic, but they will make it easier to compare candidate performances. Open tasks may be more authentic than rigidly controlled ones, but complicate the rating process.

**27. Language level of task instruction**

Since the successful completion of any speaking task depends on understanding the instruction, any task will be integrated to some extent. Consequently, as an item writer, it is important to be aware of the different possible ways of giving task instructions. According to the target group of the test (language level, literacy, age, etc), the most appropriate form of instruction should be chosen in order to minimize the effect of other skills.

Understanding the instruction is a prerequisite for adequate task performance. Consequently it is paramount that the instructions are clear and understandable. Vagueness should be avoided at all cost and the lay-out should be clear. If possible, the language should be simpler than that in the items themselves. In CEFR terms, the instruction should preferably be one CEFR level below the expected level of performance. In some cases the instructions may be written in the candidates’ first language.

**29. Integration of skills**

Other skills apart from speaking may play a role in determining a candidates’ score. Most tasks require a candidate to read or listen to a prompt and this additional skill may influence the speaking performance. A test developer may wish to isolate speaking ability from other skills or may wish to entwine speaking with other skills.

These other skills may include reading (e.g. reading a longer text to comment on), writing (e.g. taking notes while conducting a telephone call), listening (e.g. understanding an audio prompt).

The integration of other skills may interact with the candidates’ result in various ways. A test may be designed as an integrated skills test, and a deficiency in eg. reading has a controlled and indeed sought-after impact on the test result. On the other hand, a test may also wish to assess speaking alone but fail to do so because of the interference of another skill. It is also possible that a test construct is underdefined, which may lead to vagueness about its integrative nature.

Since any language test should minimise the effect of construct-irrelevant variance, it is vital to specify and control the role and impact of any skill the language test appeals to.

**30. Interaction type**

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

If, for example, the construct requires task types that assess a candidate’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 task types that assess whether a candidate can respond quickly and spontaneously, a dialogue could be the best alternative.

**33. Communicative purpose**

The communicative purpose of a task is related to the level of processing it requires and the sociolinguistic abilities it calls upon. As such, the purpose of a task is entwined with its target level and the target level of the test as a whole.

Specifying the communicative purpose of a task matters, both for the test developer and for the candidate. 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 candidate, being aware of the main communicative purpose is vital, since different communicative purposes may fundamentally change what is asked of a candidate.

A task with a referential communicative purpose might ask of a candidate to summarize a lecture by rephrasing the main and supporting ideas in a structured way. Based on the same input material, candidates could also be asked to agree or disagree (emotive), to add a convincing personal assessment of the input material’s content (conative) or to 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 is 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 test tasks and will increase the comparability across test versions.