Performance Standards for the CEFR in Dutch secondary education

An international standard setting study

now you know

Performance Standards for the CEFR in Dutch secondary education

An international standard setting study

Remco Feskens
Jos Keuning
Alma van Til
Rob Verheyen

Cito | Arnhem

Publication information

- Commissioned by: Ministry of Education, Culture and Science
- Project management: Rob Verheyen
- Process coordination: Sandra van Dijk
- Conference preparation and facilitation of expert panels: Dorrie Goossens and Rob Verheyen (German), Margreet van Aken and Evelyn Reichard (English), Dennis van den Broek and Alma van Til (French)
- Authors: Remco Feskens, Jos Keuning, Alma van Til and Rob Verheyen
- Statistical analysis: Remco Feskens, Jos Keuning
- Recommendations: Examination Board, Ruud Alers, Alex van de Kerkhof, Wim Mulder
- Assistance: Soleia Garton
- Logistics: Service Unit
- Desk editing: Petra Winkes
- Graphic design and layout: Service Unit, MMS
- Cover photo: Ron Steemers

Table of Contents

1 Introduction

1 Introduction

In the first quarter of 2012, the Netherlands Institute for Curriculum Development (SLO), the Examination Board (CvE), the Association of Teachers of Living Languages (VLLT) and the Netherlands Institute for Educational Measurement (Cito), at the request of the Ministry of Education, Culture and Science (OCW), jointly prepared a memorandum of recommendation on the incorporation of the Common European Framework of Reference for Languages (CEFR) in the examination programmes for the modern foreign languages German, English and French. One of these recommendations was to conduct an international standard setting study involving subject-area experts to determine '... how the performance of students, as measured in the central examinations and the listening comprehension tests of Cito, can be interpreted in terms of the CEFR' (SLO, 2012 – p. 2). Previous standard setting studies had involved only Dutch subject-area experts (Noijons & Kuijper, 2006; Cito, 2007). These studies were in urgent need of international validation, due to the prevailing impression that the performance standards applied by testing institutions in other countries differed from those applied in the Netherlands. In late 2012, the Ministry of OCW commissioned Cito to organise an international standard setting study for the basic, middle management, combined theoretical and vocational tracks of pre-vocational secondary education (VMBO), senior general secondary education (HAVO) and university preparatory education (VWO). In this report, we provide an account of the study and present the results. In Chapter 2, we briefly describe the CEFR and explain its status in the Netherlands. In Chapter 3, we describe the content, structure and reliability of the tests that were used in the study. In Chapter 4, we address the research method and the composition of the expert panels. In Chapter 5, we discuss the outcomes of the study. We complete this report with conclusions and a summary (Chapter 6).

2 Common European Framework of Reference for Languages (CEFR)

2 Common European Framework of Reference for Languages (CEFR)

The Common European Framework of Reference for Languages (CEFR) is a framework of level descriptions for learning, teaching and assessing modern foreign languages (Council of Europe, 2001). In Section 2.1, we provide a brief description of the content of the CEFR. We then discuss the status of the CEFR in the Netherlands (2.2).

2.1 Objectives and content

The CEFR distinguishes six levels of language mastery, ranging from *breakthrough* (A1) to *mastery* (C2). In the CEFR, assessments are made concerning language scope (i.e. what the language learner should be able to do, in which contexts and to which ends), language complexity and the extent of correctness (i.e. how correct should the language expressions be). The six levels of language mastery are defined according to scales, with a definition for each level. These definitions are interpreted in the form of *can-do statements*. The following is an example of a *can-do statement*: 'Can follow short, simply written directions (e.g. go from X to Y)'. In all, the CEFR contains 54 scales. The objective of the CEFR is to make language-mastery levels comparable internationally. It can also help to improve transitions between and within educational sectors in the Netherlands by interpreting levels of language mastery in the same manner. This makes it possible to establish continuous curriculums. Moreover, the straight-forward approach to levels of language mastery provides additional clarity for employers. Finally, the CEFR offers learners the opportunity to gain insight into the progress of their own learning processes (Van Til, Beeker, Fasoglio& Trimbos, 2011).
The CEFR is based on an action-oriented approach. As stated in the CEFR (see Council of Europe, 2011): 'Language use, embracing language learning, comprises the actions performed by persons who as individuals and as social agents develop a range of competences, both general and in particular communicative language competences. They draw on the competences at their disposal in various contexts under various conditions and under various constraints to engage in language activities involving language processes to produce and/or receive texts in relation to themes in specific domains, activating those strategies which seem most appropriate for carrying out the tasks to be accomplished. The monitoring of these actions by the participants leads to the reinforcement or modification of their competences'. The following list provides descriptions of three types of language users (basic, independent and proficient) and six levels of language proficiency (A1, A2, B1, B2, C1 and C2):

- *A1 (basic user)*: 'Can understand and use familiar everyday expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help'.
- *A2 (basic user)*: 'Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine items requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need'.

- *B1 (independent user)*: 'Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can deal with most situations likely to arise whilst travelling in an area where the language is spoken. Can produce simple connected text on topics which are familiar or of personal interest. Can describe experiences and events, dreams, hopes and ambitions and briefly give reasons and explanations for opinions and plans'.
- *B2 (independent user)*: 'Can understand the main ideas of complex text on both concrete and abstract topics, including technical discussions in his/her field of specialisation. Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party. Can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue giving the advantages and disadvantages of various options'.
- *C1 (proficient user)*: 'Can understand a wide range of demanding, longer texts, and recognise implicit meaning. Can express him/herself fluently and spontaneously without much obvious searching for expressions. Can use language flexibly and effectively for social, academic and professional purposes. Can produce clear, well-structured, detailed text on complex subjects, showing controlled use of organisational patterns, connectors and cohesive devices'.
- *C2 (proficient user)*: 'Can understand with ease virtually everything heard or read. Can summarise information from different spoken and written sources, reconstructing arguments and accounts in a coherent presentation. Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations'.

In addition to language users and language proficiency levels, the CEFR describes five skills: listening, reading, spoken interaction, spoken production and writing. Several sub-skills (i.e. global descriptors) are distinguished within each skill. Reading comprehension includes the following: reading correspondence; reading for orientation; reading for information and argument; and reading instructions. Listening comprehension includes understanding interaction between native speakers; listening as a member of a live audience; listening to announcements and instructions; listening to audio media and recordings. *Can-do statements* have been elaborated for each sub-skill. In some cases, a distinction is made between the language skills of a learner who has just reached a given CEFR level and those of a learner who is approaching the next CEFR level. As an illustration, the explanatory scale for the sub-skill 'listening to audio media and recordings' is included below:

- *A1 (basic user)*: No descriptor available.
- *A2 (basic user)*: 'Can understand and extract the essential information from short, recorded passages dealing with predictable everyday matters which are delivered slowly and clearly'.
- *B1 (independent user)*: 'Can (a) understand the information content of the majority of recorded or broadcast audio material on topics of personal interest delivered in clear standard speech, and (b) can understand the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly'.
- *B2 (independent user)*: 'Can (a) understand recordings in standard speech likely to be encountered in social, professional or academic life and identify speaker viewpoints and attitudes as well as the information content, and (b) can understand most radio documentaries and most other recorded or broadcast audio material delivered in standard speech and can identify the speaker's mood, tone etc.'
- *C1 (proficient user)*: 'Can understand a wide range of recorded and broadcast audio material, including some non-standard usage, and identify finer points of detail including implicit attitudes and relationships between speakers'.
- *C2 (proficient user)*: As C1.

As shown in this example, a learner who has just attained the CEFR level B1 is capable of understanding '...the main points of radio news bulletins and simpler recorded material about familiar subjects delivered relatively slowly and clearly'. A B1 learner who is approaching the CEFR level B2 has mastered this and can also '...understand recordings in standard speech likely to be encountered in social, professional or academic life and identify speaker viewpoints and attitudes as well as the information content' (Council of Europe, 2011).
In addition to skills and sub-skills, the CEFR distinguishes several domains: '...work or focus areas in social life in which communicative situations occur' (Council of Europe, 2011; p. 10). The CEFR distinguishes four domains. The *personal domain* refers to situations that one encounters as a private individual. Examples include hobbies, contact with relatives and friends or reading for pleasure. The *public domain* involves situations in which one is acting as a member of society (e.g. in a restaurant, at an information counter, in contact with business or other organisations). The *professional* domain comprises all work-related situations, including part-time jobs. Finally, the *educational* domain refers to all situations having to do with school and training.

2.2 The status of the CEFR in the Netherlands

In the Netherlands, secondary schools are not obliged to work with the CEFR. However, the examination syllabuses for senior general secondary education (HAVO) and university preparatory education (VWO) do indicate that examination materials should be specified in terms of the CEFR. The examination syllabuses for pre-vocational secondary education (VMBO) do not contain such a clause. Teachers' syllabi and the construction assignment specify the CEFR levels at which the items in the examination should be set (see also Section 3). These specifications are based on a study from 2006, in which subject-area experts in the Netherlands related the central examinations to the CEFR (Noijons & Kuijper, 2006). In addition, a wide range of activities concerning the CEFR have been organised in the context of secondary education. For example, a special website on the CEFR has been developed (www.erk.nl; in Dutch only), containing important information about the framework. In addition, many teaching methods state the CEFR levels on which they are based, special CEFR tests have been published and a web-based version of the European Language Portfolio has been developed. Finally, many CEFR publications have appeared and many CEFR projects have been initiated. This does not take away the fact that there is considerable freedom of choice in secondary education. Each secondary school may determine for itself whether to translate levels of mastery into CEFR levels.
Although there are currently no central examinations for the modern foreign languages in senior secondary vocational education (MBO), this situation will soon change. Beginning in the 2017/2018 school year, a central examination in English will be introduced for Level-4 MBO programmes. In addition, since the 2012/2013 school year, all MBO programmes of this level have been subject to the same generic requirements for English, regardless of whether the language is necessary for the vocation. The requirements are formulated in terms of the CEFR (see Driessen, Van Kleve & Van Kleunen, 2012). In the MBO programmes at Levels 1, 2 and 3, examinations for the languages German and French are held only at the institutional level. The examination requirements are formulated in terms of the CEFR, and they are established in qualification files for each institution. Within all teacher-training programmes at the level of higher professional education (HBO), agreements have been made regarding the CEFR levels that must be achieved in order to qualify for teacher certification in the subjects German, English, French and Spanish. This also applies to several other HBO programmes in which modern foreign languages are taught (see www.erk.nl). Unlike secondary schools, MBO and HBO institutions work according to CEFR guidelines relatively often.

3 Materials

3 Materials

The central examinations have been used to determine performance standards for reading comprehension. These examinations are compiled under the auspices of the Examination Board (CvE), and their administration is compulsory in all secondary schools. The performance standards for listening comprehension have been determined using the listening comprehension tests developed by Cito. The listening comprehension tests are part of the school-based examinations. This means that schools are not obliged to administer them. They are thus free to evaluate listening comprehension in other ways. In light of the number of orders, however, the listening comprehension tests appear to be widely administered within the system of secondary education. In the following sections, we provide a description of the central examinations and the listening comprehension tests that were used in the standard setting study. We begin in Section 3.1 by discussing the exit qualifications that the Ministry of OCW has formulated for reading and listening comprehension. We then address the content, structure and reliability of the central examinations and the listening comprehension tests (Section 3.2).

3.1 Reading and listening comprehension in the final year of secondary education

The Ministry of Education has established an examination programme with exit qualifications for reading and listening in modern foreign languages. Each year, the Examination Board supplements the examination programme with a syllabus for each type of education. These syllabi provided further explanation for the exit qualifications included in the central examination. The construction assignments that the Examination Board provides to Cito each year describe the guidelines to which the central examinations must conform. The listening comprehension tests are constructed according to the suggestions published each year by the Netherlands Institute for Curriculum Development (SLO). These suggestions do not involve any compulsory implications for schools. The exit qualifications formulated by the Ministry of Education are not stated in CEFR terms. In the syllabi and the construction assignment developed by the Examination Board, as well as in the suggestions published by the SLO, the exit qualifications are, however, related to the CEFR. These documents identify the CEFR levels that the items should involve, in addition to indicating the CEFR *can-do statements* that should be measured. The following exit qualifications apply for reading comprehension in pre-vocational secondary education (VMBO): A student in pre-vocational secondary education (VMBO) can:

- identify the relevant information contained in a text, given a certain need for information;
- identify the main ideas of a text (or portion of a text);
- identify the meaning of major elements of a text;
- compare information from one or more texts with each other and draw conclusions from this comparison;
- identify relationships between parts of a text;
- draw conclusions with regard to the writing objective, the views, the feelings of the author and the intended audience (only for the combined theoretical and vocational track);
- recognise special stylistic language characteristics (only for the combined theoretical and vocational track).

Other exit qualifications for reading comprehension have been established for students in senior general secondary education (HAVO) and university preparatory education (VWO). Students in senior general secondary education (HAVO) and university preparatory education (VWO) should be able to:

- indicate which information is relevant, given a predetermined need;
- identify the main ideas of a text (or portion of a text);
- identify the meaning of major elements of a text;
- identify relationships between parts of a text;
- draw conclusions with regard to the intentions, views and feelings of the author.

In addition to reading comprehension, exit qualifications have been formulated for listening comprehension. A student in the final year of pre-vocational secondary education (VMBO) should be able to:

- identify the relevant information contained in a text, given a certain need for information;
- identify the main ideas of a text (or portion of a text);
- identify the meaning of major elements of a text;
- anticipate what is most likely to be said next in a conversation.

The following is expected of students in the final year of secondary education of senior general secondary education (HAVO) and university preparatory education (VWO). A student can:

- indicate which information is relevant, given a predetermined need;
- identify the main ideas of a text;
- identify the meaning of major elements of a text;
- anticipate what is most likely to be said next in a conversation;
- take notes as a strategy for approaching a text;
- draw conclusions with regard to the intentions, views and feelings of the speaker (or speakers).

Although the sub-skills and domains that are mentioned in the CEFR were not used as guidelines in the construction of the tests used in the standard setting study, we can establish that the tests do correspond to the thinking on which the CEFR is based. All of the tests reflected a communicative approach. The tests can be used to determine whether students have understood the messages that the authors and speakers in question intended to send.

3.2 Content, structure and reliability of the tests

For German and English, the standard setting study distinguished performance standards for five different types of education: pre-vocational secondary education (VMBO) – basic, advanced, combined theoretical and vocational track – senior general secondary education (HAVO) and university preparatory education (VWO). For French, three types of education were included in the standard setting study. No performance standards have been determined for the basic and middle-management vocational track of VMBO, since only a few students in these tracks take examinations in French. The performance standards for reading comprehension were determined using the central examinations. The examinations in German, English and French for the combined and theoretical vocational tracks of VMBO and for HAVO and VWO are paper-based. In principle, all of the examination items for a given school year were presented to a panel of experts as part of the standard setting study. The examinations for the basic and middle-management vocational track for VMBO are computer-based. The electronic examinations cover reading comprehension, as well as listening comprehension and writing skills. The writing tasks were not used in the standard setting study, as they differed too widely from the other tasks in terms of structure, content and development. Based on the remaining

examination tasks, a combined performance standard was determined for reading and listening comprehension. It was not possible to determine a performance standard for the basic and middle-management vocational track of VMBO, as the number of reading tasks in the electronic examinations is too small to allow the reliable assessment of students' reading levels.
The performance standards for listening comprehension were determined using the listening comprehension tests developed by Cito. The listening comprehension tests for the advanced vocational, combined theoretical and vocational track of VMBO were partly the same. For this reason, a decision was made to exclude the tests for the middle-management vocational track of VMBO from the standard setting study. The overlap between the tests makes it possible to use an equating procedure to extend the performance standard for the combined theoretical and vocational track of VMBO to the middle-management vocational track of VMBO.
The central examinations and the listening comprehension tests are revised each year. We selected the most recent version for the standard setting study. For the central examinations, we used the versions from 2012. We used an older version for certain subjects, because (a) the 2012 answer key had been revised, (b) the 2012 examination contained a defective item, or (c) the 2012 examination was relatively easy or difficult in comparison with the examinations from other school years. Because we did not want to go further back in time than 2009, however, we did use examinations in a few cases that contained problems with items or the answer key. The items in question were omitted during the standard-setting procedure. The 2012 versions of the listening comprehension tests were usually selected as well. There were two exceptions. The 2013 test was used for German in the basic vocational track of VMBO, as listening comprehension tests in German for this type of education are published only in odd-numbered years. A test from 2013 was also selected for HAVO, due to a substantial change that was introduced that year: since 2013, test questions are in German, and no longer in Dutch. In the following sections, we provide detailed descriptions of the examinations and the listening comprehension tests that were used in the standard setting study. We address the texts, the question formats, test timing, length and duration, the tools that may be used, the available test data and reliability.

- *Text materials*: Both the central examinations and the listening comprehension tests, it holds that one test contains several texts. The central examinations consisted of 8–16 texts. Each examination included both shorter and longer texts. In the electronic examinations for the basic and middle-management vocational track of VMBO, the central examinations and the listening comprehension tests assessed listening comprehension in different ways. The texts were presented in fragments of 30–60 seconds. In the listening comprehension tests, each fragment was presented once, while students were able to listen to fragments multiple times in the central examinations. The standard listening comprehension tests consist of an audio portion and a video portion.
- *Question formats*: For HAVO and VWO, the central examination consists of approximately 60% multiple-choice questions and 40% pre-structured and open items. Multiple-choice questions account for a slightly greater share (65%) of the examinations for the combined theoretical and vocational track of VMBO. In general, multiple-choice items were presented in the target language, and the pre-structured and open items were presented in Dutch. Students were also required to answer the open items in Dutch, in order to prevent the student's writing style in the target language from affecting the assessment. The quotation items were an exception to this rule. The electronic examinations for the basic and middle-management vocational track of VMBO consisted primarily of multiple-choice items. All of the questions were presented in Dutch. The listening comprehension tests developed by Cito consisted entirely of multiple-choice items. The choice of the language for the items was dependent upon the type of education. In general, items for the lower types of education were presented in Dutch, while those for the higher educational types were presented in the target language.

- *Timing of tests*: All of the paper-based tests for which a performance standard was specified were administered in May, during the first period. The greatest majority of students take the tests during this period. The testing period for the electronic examinations, which covers several weeks, is in the spring. In order to prevent fraud, different versions of each examination are used. Version 1a was used in the standard setting study. The listening comprehension tests developed by Cito were administered in January by the majority of schools.
- *Length and duration of the tests*: The paper-based examinations consisted of 40–45 items, for a combined total of 45–55 points. The scoring scale for the electronic examinations comprised 36 points, including the writing assignment. The listening comprehension tests consisted of approximately 35–45 items. The testing time for the paper-based examinations varied from 120–150 minutes. Students were allowed 60–90 minutes for the electronic examinations. The testing time for the listening comprehension tests was 55–60 minutes.
- *Tools*: For both paper-based and electronic tests, students are allowed to use a bidirectional dictionary. Until 2011, students with dyslexia were allowed to have enlarged copies of the examination and additional testing time. From 2012, the font for all examinations was adjusted, thus eliminating the need for enlarged copies. Dictionaries were not allowed for the listening comprehension tests. Tests with extended answer periods are available for students with dyslexia.
- *Test data*: We had access to test data from both the central examinations and the listening comprehension tests. Test data for the central examinations were collected using the WOLF computer application. Each school is required to submit the data for the first five (or, in some cases, 10) students, based on the alphabetical order of their names. Nevertheless, many schools choose to submit the test data from all of their students through WOLF. The number of students whose scores are submitted is somewhere between 151 and 32000. Although fewer data were available for the electronic examinations, the response rates for these examinations were 100%, given that the scores of all students are submitted. The test data for the listening comprehension tests were collected using a scoring service offered by Cito. The scoring service is available for all languages and types of education, with the exception of German in the basic and middle-management vocational track of VMBO and French in the middle-management vocational track of VMBO. The numbers ranged from 642 to 9 003 students per test.
- *Reliability*: Based on the test data, it was possible to determine how students performed on the examinations and the listening comprehension tests. In Tables 3.1 to 3.4, we present the most important psychometric data. The tables also provide information on which examinations and tests we selected and which skills were measured. As shown in the table, the *p*-value ranged from .56 to .72 . This means that, on average, students earned 56%–72% of the maximum number of points. Average scores ranged from 6.1 to 7.0. Reliability scores were between .64 and .86. For the central examinations, reliability is reported in terms of the *greatest lower bound* (GLB), while the reliability of the listening comprehension tests is reported in terms of Cronbach's alpha. Both values reflect an underestimation of the actual level of reliability, with the GLB providing a weaker underestimation of reliability than do Cronbach's alpha scores (Ten Berge & Sočan, 2004). It is necessary to impose high demands with regard to the reliability of the tests, given that the allocation of CEFR levels involves major decisions at the individual level. Reliability scores lower than .80 are usually regarded as insufficient (see Evers, Lucassen, Meijer & Sijtsma, 2010).

Table 3.1 Characteristics of examinations for pre-vocational secondary education (VMBO)

	German			English			French
Reference	**VMBO [Bas.Voc.]**	**VMBO [Adv.Voc.]**	**VMBO [Comb./Th.]**	**VMBO [Bas.Voc.]**	**VMBO [Adv.Voc.]**	**VMBO [Comb./Th.]**	**VMBO [Comb./Th.]**
Year	2012	2012	2012	2012	2012	2012	2011
Testing method	Electronic	Electronic	Paper	Electronic	Electronic	Paper	Paper
Skill	*	*	Reading	*	*	Reading	Reading
Sample survey scope	151	472	19 875	3 912	3 777	32 000	5 518
Number of items	31	41	43	29	33	33	41
Scoring scale	0-36	0-45	0-48	0-36	0-45	0-48	0-47
p' value	.56	.65	.58	.72	.64	.66	.60
Average score	6.3	6.1	6.1	6.7	6.3	6.3	6.2
Reliability	.86	.83	.73	.84	.84	.81	.80

* = Reading, listening and writing

Table 3.2 Characteristics of examinations for senior general secondary education (HAVO) and university preparatory education (VWO)

	German		English		French	
Reference	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**
Year	2012	2010	2012	2011	2011	2009
Testing method	Paper	Paper	Paper	Paper	Paper	Paper
Skill	Reading	Reading	Reading	Reading	Reading	Reading
Sample survey scope	14948	14411	32000	29141	9342	2257
Number of items	42	46	45	43	40	45
Scoring scale	0-50	0-51	0-55	0-53	0-49	0-49
p' value	.63	.57	.64	.64	.61	.66
Average score	5.9	5.8	6.2	6.1	5.8	6.2
Reliability	.76	.77	.81	.79	.79	.84

Table 3.3 Characteristics of listening comprehension tests for pre-vocational secondary education (VMBO)

	German		English		French
Reference	**VMBO [Bas.Voc.]**	**VMBO [Comb./Th.]**	**VMBO [Bas.Voc.]**	**VMBO [Comb./Th.]**	**VMBO [Comb./Th.]**
Year	2013	2012	2012	2012	2012
Testing method	Paper	Paper	Paper	Paper	Paper
Skill	Listening	Listening	Listening	Listening	Listening
Sample survey scope	---	1801	670	3969	642
Number of items	36	37	34	36	35
Scoring scale	0-36	0-37	0-34	0-36	0-35
p' value	---	.65	.63	.69	.67
Average score	---	6.2	6.3	6.2	6.2
Reliability	---	.67	.82	.79	.72

Table 3.4 Characteristics of listening comprehension tests for senior general secondary education (HAVO) and university preparatory education (VWO)

	German		English		French	
Reference	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**
Year	2013	2012	2012	2012	2012	2012
Testing method	Paper	Paper	Paper	Paper	Paper	Paper
Skill	Listening	Listening	Listening	Listening	Listening	Listening
Sample survey scope	2878	3399	9003	6960	2038	3166
Number of items	40	38	39	36	35	36
Scoring scale	0-40	0-38	0-39	0-36	0-35	0-36
p' value	.68	.71	.72	.71	.65	.65
Average score	6.3	6.3	6.4	6.3	6.2	6.3
Reliability	.64	.71	.78	.72	.64	.64

4 Methods

4 Methods

In order to interpret the test results from a CEFR perspective, it is necessary to have the performance standards for language levels A1–C2. The performance standards indicate which test results are required in order to demonstrate a given language level. It is important to be able to legitimise the performance standards that are selected. This means that a performance standard should preferably be developed methodically, in such a way that subject-area experts, teachers and students can clearly see the manner in which the performance standard was developed. Various methods for establishing standards have been proposed in the literature. A distinction exists between test-centered and examinee-centered standard setting procedures (Jaeger, 1989; Kaftandjieva, 2004; Berk, 1986; Hambleton, Jeager & Plake, 2000). In test-centered methods, subject-area experts base the performance standard on the content of the test and on the learning materials. The performance standard is independent of the testing results that students actually achieve. In examinee-centered methods, the performance standard is established based on the test scores of a group of students. The performance standard thus depends upon the scores of the group of students tested. Subject-area experts play an important role in both types of methods. They are the ones who determine which behaviour may be expected of students located exactly on the borderline of a given CEFR level. In this chapter, we discuss the procedure followed in establishing the performance standards for the central examinations and the Cito listening comprehension tests. In Section 4.1, we provide a description of the conference that was held. This is followed by Sections 4.2 and 4.3, in which we devote extensive attention to the composition of the expert panels and the standard-setting procedure used during the conference. Finally, in Section 4.4, we report which data were collected during the conference and how we analysed them.

4.1 Structure of the conference

The performance standards for the central examinations and the listening comprehension tests were established during a five-day conference, which took place in September 2013 in Scheveningen, the Netherlands. Three different panels, each consisting of 17–20 subject-area experts, participated in the conference. In Section 4.2, we describe the composition of the panels. Prior to the conference, the tests to be evaluated were sent to the subject-area experts. Each expert was asked to make a preliminary estimate of the CEFR level that was measured by each test. The conference started with a plenary opening, in which the purpose of the conference was explained and the function of the central examinations and the listening comprehension tests were examined in the context of the educational system in the Netherlands. After the plenary opening, the three panels of subject-area experts continued in separate sessions for the languages German, English and French. The standard setting procedure that was to be followed was explained in the target language, and there was an opportunity to ask questions. The process of establishing the first performance standard was then started.

A conference day officially consisted of two blocks of approximately three hours each. In the morning, a listening comprehension test was submitted for consideration. In the afternoon, we asked the panels to evaluate a central examination. In each assessment, the performance standard to be established was discussed in advance. This meant that subject-area experts jointly established the CEFR level that was being measured with each test based on their individual preliminary estimates. The CEFR levels the subject-area experts selected for each test are presented in Table 4.1. After the CEFR level had been established, the subject-area experts were asked to participate in two rounds of assessment in order to determine the test score at

which a student could be seen as having demonstrated the selected CEFR level. In both rounds of assessment, the subject-area experts worked individually, applying the work method prescribed in the standard setting procedure (see Section 4.3). Prior to the second round, the assessments from the first round were presented to the subject-area experts, and the results were discussed. The objective of the interim discussions was to clarify the arguments that the subject-area experts had used in assigning their assessments. The interim discussions were also used as an attempt to reduce the initial differences between the assessments of the individual subject-area experts.

Table 4.1 Selected performance standards, broken down by language, skill and education

		Educational level				
Language	**Skill**	**VMBO [Bas.Voc.]**	**VMBO [Adv.Voc.]**	**VMBO [Comb./Th.]**	**Senior general secondary education (HAVO)**	**Pre-university education (VWO)**
German	Reading	B1	B1	B1	B2	C1
	Listening	B1	---	B2	C1	C2
English	Reading	B1	B1	B2	C1	C1
	Listening	B1	---	B2	C1	C1
French	Reading	---	---	B1	B1	B2
	Listening	---	---	A2	B1	B2

4.2 Composition of the expert panels

Subject-area experts play an important role in any standard-setting procedure. It is therefore important to proceed carefully when recruiting and selecting subject-area experts. Potential subject-area experts should at least have knowledge of the domain addressed by the test, and they should ideally have experience in the assessment of test content and the work of students (see Evers, Lucassen, Meijer & Sijtsma, 2010). The subject-area experts who participated in the conference were recruited according to four criteria. The first criterion related to knowledge of and experience with the CEFR. We exclusively invited people who, based on their background and/or experience, were known to have a comprehensive knowledge of the CEFR. The second criterion related to the professions of the subject-area experts. This means that, in the selection process, an attempt was made to ensure that the following professions were represented in each panel: (a) researchers, (b) testing experts, (c) teachers and (d) other occupations (e.g. policy officials and publishers). This process allowed us to combine the insights of people who were involved with the CEFR in a variety of ways. The third criterion concerned the native languages of the subject-area experts. We distinguished between people who speak the target language as their native language and those who speak the target language as a second or foreign language. The goal was to primarily involve native speakers in the conference. The final criterion concerned the countries in which potential subject/area experts were employed. We invited subject-area experts in the target-language countries, in the Netherlands and in other European countries to participate. By involving subject-area experts from the target-language countries in the conference, we were able to ensure that the results of the standard setting procedure would correspond to the prevailing views about the CEFR in these countries. The inclusion of subject-area experts from the Netherlands allowed us to create a base of support amongst teachers and

other CEFR professionals in the Netherlands. By selecting subject-area experts from other European countries to participate in the panels as well, we were able to prevent the influence of particular countries or regions from reaching undesirable proportions.

In all, 56 subject-area experts from 19 different European countries participated in the conference. The German tests were assessed by 17 subject-area experts, while 19 subject-area experts addressed the English tests and twenty experts discussed the French tests. A description of the panels according to several relevant background characteristics is presented in Table 4.2. The panels for German and French included eight nationalities, while the panel for English included 15 nationalities. With few exceptions, all of the subject-area experts for German and French were from Western Europe. The distribution of the subject-area experts for English across the European regions was relatively even. The majority of members in the German panel were from the target-language countries of Germany, Switzerland and Austria. In the French panel, the number of subject-area experts from France was equal to the combined total from other countries. All of the subject-area experts in the English panel came from countries in which English is not the official language. In each panel, the different occupations were represented by at least two different subject-area experts. The panels for German and English contained a relatively high share of subject-area experts who were employed as test developers or project leaders at testing institutes or in the testing divisions of language institutes.
A relatively large share of the subject-area experts in the French panel were employed as French teachers or as scientific researchers at universities. Native speakers accounted for approximately two third of the panels for German and French. The other third consisted of people who spoke German or French as a second language. The situation was different for the English panel: one quarter of the participants were native speakers, while three quarters had not been raised speaking English as a native language. It is interesting to note that all of the native speakers in the English panel were employed outside of the UK.

Table 4.2 Composition of expert panels for German, English and French

Variable	Definition	German	English	French
Region	Northern Europe	---	4	1
	Western Europe	15	5	18
	Eastern Europe	1	3	---
	Southern Europe	2	7	1
Country of origin	Speaks target language	11	---	10
	Does not speak target language	7	19	10
Occupation	Testing expert	8	7	3
	Scientist/researcher	5	6	5
	Teacher of German, English or French	2	4	5
	Other*	3	2	7
Native language	Target language	12	5	14
	Not target language	6	14	6

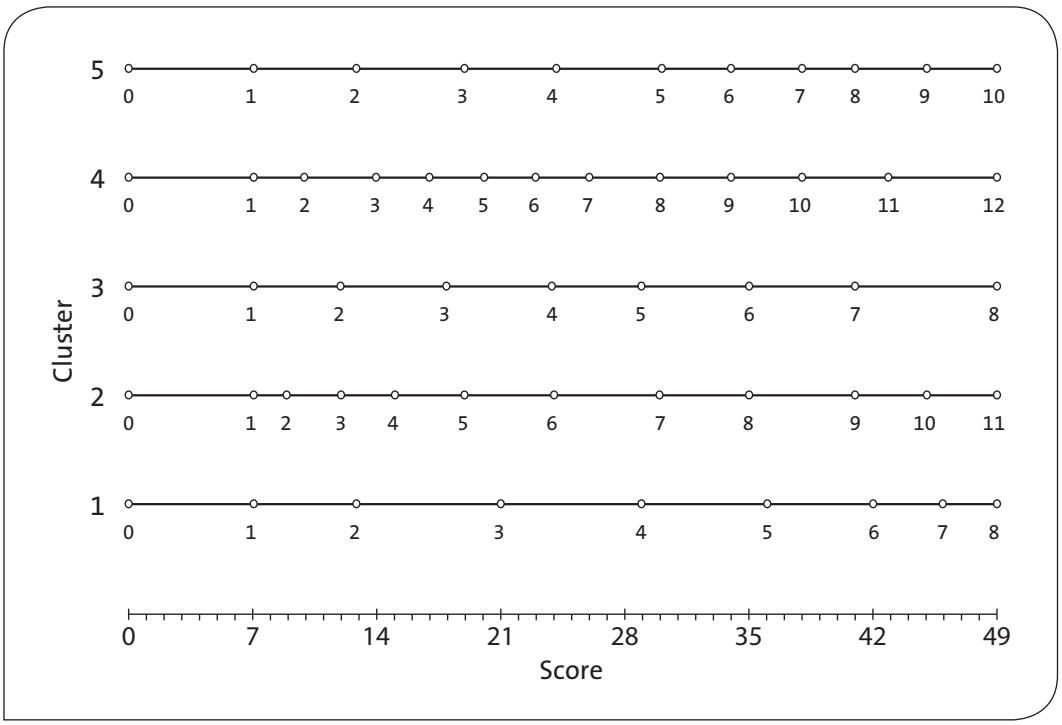
* Note: The *Other* category consisted of subject-area experts who were employed as policy officials or as curriculum developers, or who were working in embassies, trade associations, for publishers or knowledge institutes

The composition of the panels for German and French differed from that of the panel for English (see Table 4.2). This was due to the position of the various languages within Europe. Whereas English is of particular importance internationally as a second language, the primary roles of French and German are largely in the target-language countries. For English, we therefore attempted to achieve an even distribution over the various regions of Europe. In the recruitment for the French and German panels, it made more sense to concentrate on the target-language countries. Because the importance of German is concentrated in the German-speaking countries, we focused our search for experts primarily on these countries, and particularly in the leading testing institutes. We followed the same strategy for the French panel. The subject-area experts for the English panel were recruited primarily from amongst the members of international organisations in the testing field, including the Association of *Language Testers in Europe* (ALTE) and the *European Association for Language Testing and Assessment* (EALTA). Considerable coordination in the area of the CEFR takes place within these organisations, including coordination with the target-language country of the UK. Most of the subject-area experts for the German and French panels were recruited directly from the actual institutions. An overview containing the names of all participants is included in the appendix.

4.3 Standard setting procedure

A test-centered standard setting procedure was used during the conference. The method that was applied contains elements of the *Bookmark* method (Mitzel, Lewis, Patz & Green; 2001), the *Angoff* method (Angoff, 1971) and the *Direct Consensus* method (Sireci, Hambleton & Pitoniak, 2004). This new method is called the 3DC method. The designation '3DC' stands for *Data-Driven Direct Consensus*. The 3DC method shares many similarities with the *Direct Consensus* method, and it adds the use of empirical data to this method. The 3DC method assumes that a test consists of multiple items that can be divided into a number of clusters with a comparable number of points. The central examinations and the listening comprehension tests are split into four to six clusters. In general, a cluster consists of items relating to the same text or the same speakers. The instructions in Dutch and the Dutch items in some of the central exit examinations and audio-visual tests were, for the benefit of the subject-area experts, translated into the target language. This was indicated where appropriate, so that the participants could take this into account in their assessment. The subject-area experts were asked to indicate the scores that students would be expected to achieve in each cluster if they were exactly on the borderline of the selected CEFR level. In theory, it should be sufficient to present only the clusters (i.e. the text or listening fragments and the corresponding test items) to the subject-area experts. In such a case, they would use the CEFR to estimate the level of difficulty of a cluster, using this estimate to determine the minimum score that a borderline or minimally competent candidate would achieve. One major disadvantage of this approach is that the subject-area experts would have only the descriptors of the CEFR on which to rely in the assessment. They would not know how the different clusters related to each other empirically. It could happen that a population of students, contrary to expectations, performed better in one cluster than they did in others. It is therefore advisable to provide the subject-area experts with additional information about how the students performed in the various clusters. This allows the assessment to be based on both content and the empirical information. The assessment is thus likely to be more realistic and to contain fewer inconsistencies. Figure 4.1 uses a sample test to demonstrate how the empirical information on the behaviour of students was presented to the subject-area experts during the conference.

Figure 4.1 Marking sheet in which the various clusters in the test are related to each other according to an item theory response model



As shown in the sample displayed in Figure 4.1, the subject-area experts were presented with five clusters. The number of test items in each cluster varied from 8 (for Clusters 1 and 3) to 12 (for Cluster 2). The lines in the figure represent the scoring scales associated with the different clusters. The length of the scoring scale depends upon the number of test items in a cluster and the scoring prescription. If the test items are scored dichotomously (true/false), the length of the grading scale is equal to the number of test items. This is the case in this example. The scoring scale for the full test is displayed on the horizontal axis. As shown in the figure, the scores ranged from .0 to .49. A student answering none of the test items correctly would achieve a score of 0. A student answering all of the test items correctly would achieve a score of 8 + 11 + 8 + 12 + 10 = 49.

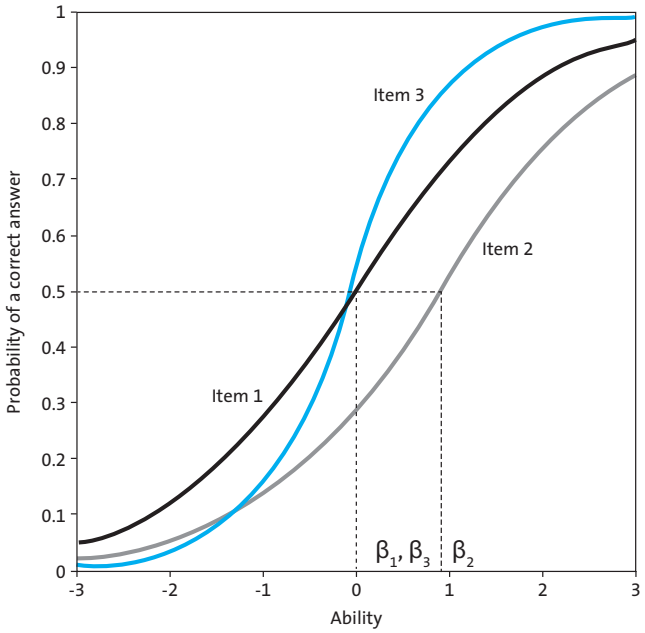
In Figure 4.1, the scoring scales for the individual clusters are related to the scoring scale for the full test. If a student's score on the first cluster is 4, the student would be expected to achieve a score of 29 on the full test. If a student's score on the sixth cluster is 7, the expected score on the full test would be 38. The prediction can also be performed in reverse. If a student's score on the full test is 34, we would expect the student to have the following scoring profile: (Cluster 1) Score 5, (Cluster 2) Score 8, (Cluster 3) Score 6, (Cluster 4) Score 9, and (Cluster 5) Score 6. Figure 4.1 thus allowed the subject-area experts to see exactly how the clusters and the complete test related to each other. They were able to consider this information in their assessments. For example, if a subject-area expert is considering setting the border for the first cluster at Score 6 (based on the CEFR) and the border for the second cluster at Score 5, subsequently comparing this assessment to the empirical data, it would become clear that such a scoring pattern would be relatively rare for the first two clusters. The two scores mentioned above reflect different skill levels. Whereas Score 6 on the first cluster represents a relatively high skill level (42), Score 5 on the second cluster represents a significantly lower skill level (19).

The clusters and the full test can be related to each other using an item response theory model. Item response theory models offer excellent possibilities for predicting how students with a particular skill level on subsets will respond to test items. Many item response theory models have been proposed in the literature (Hambleton, Swaminathan & Rogers, 1992; Embretson & Reise, 2000; Van der Linden & Hambleton, 1997). The students' answers on the listening comprehension tests and the central examinations were analysed according to the *One-Parameter Logistic Model* (OPLM) developed by Verhelst and Glas (1995). For dichotomously scored test items, the OPLM item-response function is calculated according to the following function:

$$P(x_j = 1 | \theta) = P_j(\theta) = \frac{\exp[a_j(\theta - \beta_j)]}{1 + \exp[a_j(\theta - \beta_j)]}, \qquad \text{for } j = 1, \ldots, J,$$

where θ represents the ability of a student, $a_j > 0$ is the discrimination index for test item j, β_j represents the item difficulty, and x_j is a random variable with a value of 0 or 1. As demonstrated in this example, the model presents the probability of a correct answer ($xj = 1$) to test item j with a discrimination index of aj and a difficulty parameter β_j as a function of θ. By way of illustration, three different item response curves for the OPLM are displayed in Figure 4.2. Taking the point at which a student has 50% probability of answering the test item correctly as a reference, we see that Test Item 2 requires more ability in order to reach a probability of 50% than do Test Items 1 and 3. From an empirical perspective, therefore, the second test item thus has a higher level of difficulty than do the first and third test items. Figure 4.2 also shows that the shape of the item response curves for Test Items 1 and 2 is the same, and that it differs for Test Item 3. The different shape of the curve for Test Item 3 is the result of a different value for a_j. From this information, it can be derived that Test Item 3 is better than Test Items 1 and 2 are at discriminating within a specific ability region between students with a lower and higher ability, as the probability of answering the test item correctly increases more rapidly along with an increasing ability.

Figure 4.2 Three item-response curves for the OPLM



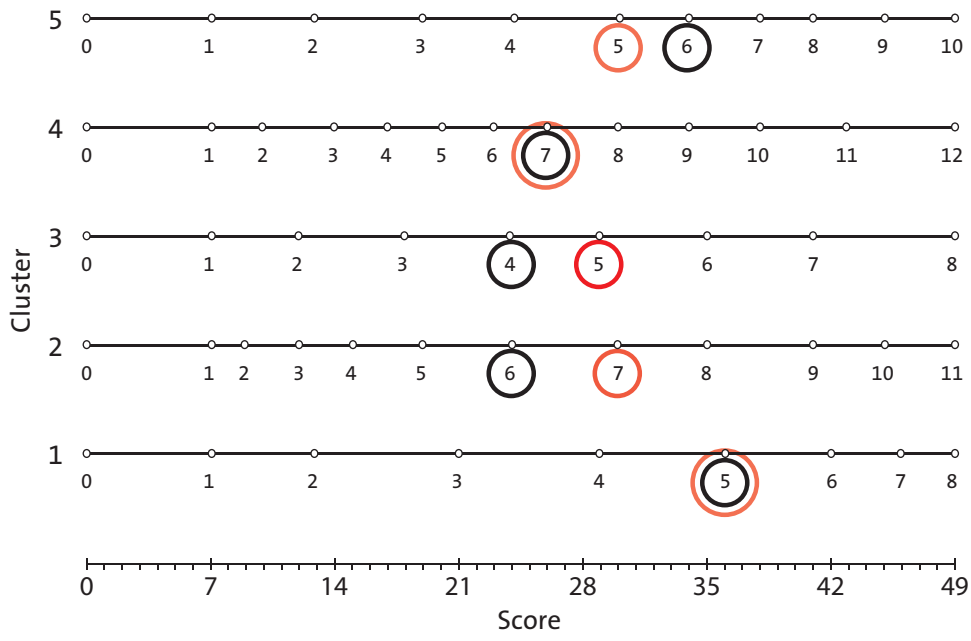
The item-response functions are estimated using the students' answers to the items in a test. The 3DC standard setting procedure, which was applied during the conference, can thus be used only if data are available. Once the item response functions have been estimated, simulation techniques can be used to calculate expected cluster scores for each possible score on the full test. We proceeded as follows. First, for *N* = 500000 students, we drew a possible value for θ from a normal distribution with an average of μ and a standard deviation of σ. In the second step, we used the values drawn for θ and the item parameters to generate answers to the items on the test. Assuming a randomly drawn number g from the interval (0, 1), Score 1 is assigned to a test item as $g \geq Pj(\theta)$, and otherwise Score 0. In the third step, we used the answer patterns for all possible scores $x+$ on the full test in order to calculate an expected scoring profile. The expected value for a cluster k, $k = 1, ..., K$, is equal to:

$$E(x_k \mid x_+) = \frac{1}{n_{\mathbf{x}|x_+}} \sum_{\mathbf{x}|x_+} \sum_{j=1}^{k_j} X_{jk},$$

where the sum continues over all simulated answer patterns $\mathbf{x} = (x_j, \ldots, x_J)$ with $\sum_{j=i}^{J} x_j = x_+$. Finally, we displayed a figure containing the expected score for the full test, given a particular score on a cluster. We thus did not present all of the results from the simulation to the subject-area experts. For the listening comprehension tests and the central examinations, the estimate of the item-response theory model was calculated using the OPLM computer application (Verhelst, Glas, & Verstralen, 1995). The simulation used to make the figure that the subject-area experts used during the standardsetting procedure was executed in R2.14.1 (R Development Core Team, 2011).

During the conference, each performance standard was established in two assessment rounds. The subject-area experts used the same figure in both assessment rounds. The following question was posed to the subject-area experts: 'Which score would a student be expected to achieve on this cluster if his/her ability is exactly on the borderline between satisfactory and unsatisfactory for language level A1, ..., C2?' In determining the performance standard for B1, the subject-area experts thus had to base themselves on a student who had mastered the *can-do statements* associated with the initial level of B1. For example, for listening comprehension, this would mean that the student 'can understand the main points of clear standard speech on familiar matters regularly encountered in work, school, leisure etc., including short narratives' (Council of Europe, 2011). In the first round, the borderline candidate's score for each cluster in the figure was marked using a black ballpoint pen. The results were discussed. To start the discussion, the results of the first assessment round were projected, and a few subject-area experts were asked to explain their assessments of one or two clusters. In general, the question was raised to two subject-area experts who found themselves at opposite extremes of the assessment spectrum, as well as to two experts located more in the middle. In the second assessment round, the subject-area experts were once again asked to mark the cut score for each cluster on the figure. In this round, the experts used a red ballpoint pen, thus clearly indicating whether they had adjusted their initial scores and, if so, where. To illustrate, a completed assessment form is presented in Figure 4.3.

Figure 4.3 Example of a completed assessment form



The 3DC standard setting procedure attempts to bring together the strengths of existing standard setting procedures. The 3DC method offers several advantages. First, the question posed to the subject-area experts is relatively easy (see also Angoff, 1988; Goodwin, 1999). In the *Angoff* method, subject-area experts are asked to indicate the probability of a correct answer for each test item. In the 3DC method, they need only determine the score that a borderline candidate would be expected to achieve on a larger subset of test items. Second, in practice, analyses of the correspondence between raters continue to be far too seldom (Berk, 1996). As compared to other methods, the 3DC method offers better opportunities for evaluating the correspondence within and between subject-area experts (see Subsection 4.4). Third, the 3DC method makes it possible to establish performance standards relatively efficiently and quickly. Standard setting procedures that call for evaluations of each test item are particularly time-consuming (Sireci, Hambleton & Pitoniak, 2004). Fourth, the 3DC method can be applied without problems related to polytomously scored test items. This would be more difficult, for instance, in the application of the *Bookmark* method, because the difficulty of a test item is related to the number of points, which is taken as the point of departure. It is confusing for subject-area experts when polytomously scored test items appear several times in rankings based on the level of difficulty. Finally, the 3DC method can be used in combination with various item response theory models. This is not usually the case with alternative methods. For example, it would not be a straightforward task to rank test items according to level of difficulty if the students' answers to the test items are analysed according to the two- or three-parameter item response theory model.

4.4 Statistical analyses

After applying the 3DC standard-setting procedure, for each subject-area expert, we had cut scores for each cluster, as well as a cut score for the full test. Figure 4.4 presents the exact appearance of the data matrix. On the left side of Figure 4.4 are the numbers (or names) of the

subject-area experts participating in the standard-setting procedure. These are directly followed by the cut scores for the various clusters that each of the subject-area experts had chosen. Finally, the cut score for the full test is presented. This score is equal to the sum of the cut scores for each cluster, $C_{\text{total}} = \sum_{k=1}^{K} C_k$, and it was not established directly by the subject-area experts. Several descriptive statistics are presented at the bottom of Figure 4.4. First, the mode is used to indicate the cut score that was selected most frequently. Amongst other patterns, we can see that the performance standard for the first cluster was most frequently located at Score 6, and the standard for the second cluster was most frequently located at Score 8. These scores are shaded in orange. The frequencies associated with the various modes were then reflected. In the first cluster, Score 6 was selected by 7 of the 20 subject-area experts. In the third cluster, a larger group of subject-area experts agree with each other: 13 of the 20 subject-area experts set the performance standard for this cluster at Score 6. This is reflected in the number of cells that are shaded in orange. The results for each cluster were followed by the performance standard for the full test, as derived from the 20 × 5 = 100 individual assessments. The example presented in Figure 4.4 reveals the performance standard for the full test at $\left[\bar{C}_{\text{total}}\right] = 36.$ After the first assessment round, the data matrix in Figure 4.4 could be presented to the subject-area experts and used as input for the discussion. The definitive data matrix was created after the second round of assessments. This data matrix provides the input for the analysis of correspondence between raters. An initial indication of the extent to which the subject-area experts agreed with each other is subsequently reflected in the number of cells shaded in orange. More orange indicates greater correspondence between the subject-area experts.

Figure 4.4 Data matrix after the application of the standard setting procedure

Expert	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Total
1	5	8	6	9	6	34
2	5	8	6	10	7	36
3	7	9	7	11	7	41
4	7	8	6	10	7	38
5	6	5	4	7	5	27
6	6	6	6	9	6	33
7	5	8	6	9	6	34
8	6	7	7	8	7	35
9	6	8	6	9	6	35
10	6	9	6	9	7	37
11	6	9	5	10	8	38
12	7	8	7	11	7	40
13	5	8	6	9	6	34
14	7	9	6	10	7	39
15	7	9	7	10	8	41
16	6	7	6	9	7	35
17	5	8	6	9	6	34
18	8	8	6	11	5	38
19	7	8	6	10	8	39
20	5	7	5	9	7	33
Mode	6	8	6	9	7	
Frequency	7	10	13	9	9	
Cut score						36
Maximum						49

Rater correspondence was analysed according to the 20 × 5 = 100 individual assessments, which are displayed in Figure 4.4 as an example. The selected cut scores for each cluster can be used in the analysis. One disadvantage to this approach, however, is that it ignores the fact that the scoring scales for the clusters are not directly comparable. Depending upon the level of difficulty of the test items in a cluster, one cluster might require a higher ability than another cluster in order to achieve Score 7. For this reason, it is better to convert the cut scores that the subject-area experts selected for each cluster (prior to the analysis) into the scoring scale for the full test. This conversion is simple if the test items are calibrated according to a model from item response theory (see Section 4.3). Various measures can be used to analyse inter-rater correspondence at multiple levels (see e.g. Gwet, 2010; Le Breton & Sentor 2008; Shrout & Fleiss, 1979; Sim & Wright, 2005; Tinsley & Weiss, 2000; Uebersax, 1992). To analyse the results of the conference at the level of the individual subject-area experts, we examined (a) the empirical plausibility of the assessment, (b) the impact of an assessment on $\left[\bar{C}_{\text{total}}\right]$ and (c) the consistency between assessments. The result was evaluated using *Gower's similarity coefficient* for absolute agreement and the *Finn coefficient* for relative agreement. At the same time, a χ^2-test was used to determine the extent to which the group of subject-area experts could be considered a realistic sample of an imaginary population of all possible subject-area experts.

In the evaluation of the empirical plausibility of an assessment, a comparison was made between the pattern of observed borderline scores and the expected pattern of borderline scores given C_{total}. The difference is expressed as a χ_1^2 distance. The conditional expectation for each cluster was determined according to the simulation technique described in Section 4.3. The χ_1^2 distance for a subject-area expert was calculated by using the standard χ^2 formula, with the statistical test performed in the manner described in Verhelst (2009). The outcome shows the likelihood that a given assessment would be observed in actual educational practice.
The consistency between the assessments of one subject-area expert with those of the other subject-area experts is determined according to the *ranking similarity index* (RSI). This index is calculated as the average correlation of a subject-area expert with the rest of the subject-area experts. The impact of an individual assessment of $\left[\bar{C}_{\text{total}}\right]$ is determined by disregarding the assessments of the subject-area experts concerned in the calculations. The result reveals the extent to which the assessment of the subject-area expert determines the level at which the performance standard is ultimately set. The three described measurements are calculated in order to detect extreme or aberrant subject-area experts. High χ_1^2 distance, large impact and/or low RSI may give cause to eliminate the assessments of a certain subject-area experts. Restraint is advised in the elimination of assessment, however, as 'aberrant' behaviour does not necessarily reflect unwillingness or incompetence. It might be legitimate in light of the rater's professional knowledge, function and background. Exploratory analyses revealed several examples of behaviour that could be considered aberrant. In the light of the manner in which the subject-area experts were selected (see Section 4.2) and the extensive instructions given during the conference, this behaviour was unlikely to be the result of unwillingness or incompetence. In order to avoid debate concerning whether to include specific assessments, it was decided to eliminate one randomly selected maximum assessment and one randomly selected minimum assessment from the analyses, even if it was not necessarily based on χ_1^2 distance, impact or RSI.

The final result was evaluated according to a relative and an absolute dimension for inter-rater agreement. High inter-rater agreement is important, as it determines the legitimacy of the outcome. The following standard values are proposed in the literature: (unsatisfactory) < .60, (satisfactory) .61 - .80, (good) > .80 (Landis & Koch, 1977; Evers, Lucassen, Meijer & Sijtsma, 2010). The *Finn coefficient* is used as a relative measure of inter-rater correspondence (Finn, 1970). This dimension is similar to the intra-class correlation coefficient, but its results are more realistic for data with a low measurement level. The measurement level during the conference

was relatively low, given the limited number of assessment categories. *Gower's similarity coefficient* was used as a measure of absolute inter-rater agreement (Gower, 1971). The main difference between the two measurements is that, if the assessments of all subject-area experts on all clusters are identical, the *Finn coefficient* cannot be determined, while *Gower's similarity coefficient* in that case would be equal to 1. One disadvantage of *Gower's similarity coefficient*, however, is that the outcome can provide a distorted picture if not all of the assessment categories are used. For this reason, we calculated both *Gower's similarity coefficient* and the *Finn coefficient*. Finally, we examined whether the distribution that we found in the assessments of the subject-area experts corresponded to the distribution in test results that could be expected if the same student with a level of competence equal to $\left[\bar{c}_{\text{total}}\right]$ would take the test an infinite number of times. In performing the χ_2^2 tests, the borderline scores of the individual subject-area experts and the performance standard were converted to the skill scale. A *p*-value greater than .05 would indicate that the level of agreement that we observed in the group of subject-area experts is realistic, given the reliability of the test on Point $\left[\bar{c}_{\text{total}}\right]$. After all, some variation in the assessments of subject-area experts is to be expected on the basis of a test's reliability. Moreover, in a theoretical test-retest situation, students would not always answer the same number of test items correctly on a given test. If the assessments of the subject-area experts differ substantially, however, and if the outcome of the χ_2^2 test is significant at the 5% level, this may indicate insufficient support for the performance standard.

5 Results

5 Results

During the conference, 24 tests were assessed by 56 subject-area experts. The data were analysed after the conclusion of the conference. First, the position of the performance standard was determined for each test. For the performance standard, we used the trimmed mean of the sum of the individual assessments for each cluster. The trimmed mean involves the calculation of the mean after discarding one randomly selected maximum assessment and one randomly selected minimum assessment. We then determined the extent of support for the performance standard. To do this, at the individual level, we examined (a) the empirical plausibility of the assessment, (b) consistency with other assessments and (c) the impact of the assessment on the performance standard. At the group level, we calculated the 70% confidence interval for the performance standard. We also calculated *Gower's similarity coefficient* for absolute agreement and *the Finn coefficient* for relative agreement. Finally, a χ_2^2 test was used to determine the extent to which the group of subject-area experts could be considered a realistic sample of an imaginary population of all possible subject-area experts. Chapter 4 provides a comprehensive description of the procedure that was followed in the analysis of the data collected during the conference. In this chapter, we present the results. In Sections 5.1, 5.2 and 5.3, we discuss the most important characteristics and results for German, English and French in each type of education. In each section, we also elaborate on several noteworthy results. These explanations differ in each section, due to the decision to report only the most interesting results. The full results are presented in Tables 5.1–5.24. In Section 5.4, we discuss the practical implications of the outcomes. The results are summarised and discussed according to the percentage of students in the Netherlands achieving a particular CEFR level.

5.1 German

For German, five central examinations and four listening comprehension tests were assessed by a panel of 18 subject-area experts. Four subject-area experts were not able to attend all of the standard-setting sessions. In practical terms, this meant that there was a fixed expert panel of 14 people, which was usually supplemented by two of the four other people. In the following sections, we present the results for the basic vocational track of pre-vocational secondary education (VMBO; 5.1.1); the middle-management vocational track of VMBO (5.1.2); the combined theoretical and vocational track of VMBO (5.1.3); senior general secondary education (HAVO; 5.1.4); and university preparatory education (VWO; 5.1.5). A summary of the results is presented in Table 5.25a.

5.1.1 Pre-vocational secondary education (VMBO) Basic vocational track

During the conference, the experts assessed two tests that are used in the basic vocational track of VMBO: the listening comprehension test from 2013 and one of the variants included in the electronic examination from 2012. According to the expert panel, the listening comprehension test is suitable for determining whether the listening skills of students are located at CEFR level B1. The test contained 36 dichotomously scored items. For the standard-setting procedure, the 36 test items were divided into four clusters with the following scoring scales: Cluster 1 (0-9), Cluster 2 (0-8), Cluster 3 (0-10) and Cluster 4 (0-9). Table 5.1 provides an analysis of the data collected in the second round of assessments. First, we present the cut scores that were proposed by the various subject-area experts. The assessments for each cluster follow in Columns C1–C6. Within each cluster, the assessments that correspond to the mode are indicated in bold. In the last four columns, the behaviour of each subject-area expert is examined in relation to an item-response theory model and other subject-area experts. We first

report the χ_1^2 distance for each subject-area expert, calculated according to the observed and expected pattern of borderline scores (Section 4), along with the corresponding critical value for the α =.10 significance. This is followed by the *Ranking Similarity Index* (RSI) and the impact. Finally, at the bottom of Table 5.1, we report the 70% confidence interval, *Gower's similarity coefficient,* the *Finn coefficient* and the outcome of the χ_2^2 test across all subject-area experts.

Table 5.1 Results of the standard setting procedure for the listening comprehension test for German VMBO [Bas.Voc.] (CEFR = B1, $\lceil \bar{C}_{\text{total}} \rceil$ = 21)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	22	**6**	**5**	6	**5**	---	---	.04	2.18	.31	0
2	21	**6**	**5**	**5**	**5**	---	---	.13	2.42	.50	0
3	23	**6**	**5**	6	6	---	---	.05	2.14	-.18	0
4	19	5	**5**	**5**	4	---	---	.14	2.78	.38	0
5	22	7	**5**	**5**	**5**	---	---	.42	2.18	.52	0
6	22	**6**	**5**	6	**5**	---	---	.04	2.18	.31	0
7	25	**6**	6	7	6	---	---	.08	1.80	-.49	0
8	22	**6**	**5**	6	**5**	---	---	.04	2.18	.31	0
9	22	**6**	**5**	6	**5**	---	---	.04	2.18	.31	0
10	20	**6**	**5**	4	**5**	---	---	.54	2.56	.43	0
11	21	**6**	**5**	**5**	**5**	---	---	.13	2.42	.50	0
12	22	**6**	6	**5**	**5**	---	---	.30	2.18	.44	0
13	20	**6**	**5**	**5**	4	---	---	.30	2.56	.56	0
14	23	7	6	**5**	**5**	---	---	.51	2.14	.54	0
15	15	3	**5**	4	3	---	---	.98	3.36	-.03	1
16	21	**6**	**5**	**5**	**5**	---	---	.13	2.42	.50	0

70%-BI = (20, 23) ; *Gower* = .95 ; *Finn* = .96 ; χ_2^2 (13, *N* = 14) = 2.05, *p* = 1.00

According to the subject-area experts, the performance standard for CEFR level B1 is somewhere between Score 15 and Score 25, as shown in Table 5.1. Whereas one subject-area expert argued that a student should answer 15 ÷ 36 = 42% of the test items correctly in order to demonstrate CEFR level B1, another subject-area expert proposed that CEFR level B1 could not be confirmed unless a student answers 25 ÷ 36 = 69% of the test items correctly. At first glance, this appears to be a relatively large difference. Detailed analysis shows that, compared to the other subject-area experts, Subject-area Expert 15 was remarkably mild in selecting the borderline scores for each cluster. The RSI for this subject-area expert (-.03) is relatively low, and the impact shows that the performance standard increases by one point if we exclude this subject-area expert from the analysis. Subject-area Expert 7 set the bar quite high, relative to the other subject-area experts. The small χ_1^2 distance indicates that the assessment of this subject-area expert is plausible from an empirical perspective. The decision to use a trimmed mean in the analysis automatically eliminates the assessments of the two subject-area experts mentioned above from the analysis. If repeated, therefore, the performance standard for CEFR level B1 would thus fall within the interval $\left(\lceil \bar{C}_{\text{total}} \rceil - 1.036 \times \sigma_{C_{\text{total}}} ; \lceil \bar{C}_{\text{total}} \rceil + 1.036 \times \sigma_{C_{\text{total}}}\right) = (20;23)$ in 70% of the cases. This interval can be considered small. *Gower's similarity coefficient* (.95) and the *Finn coefficient* (.96) are thus high. The distribution in the group of subject-area experts is small, given the reliability of the test at the location of the performance standard (or borderline score). This can be observed in the results of the χ_2^2 test: χ_2^2 (13, N = 14) = 2.05, p = 1.00. There was thus sufficient support among the subject-area experts to regard the listening comprehension of students in the basic vocational track of VMBO as B1 beginning with Score 21.

In addition to a listening comprehension test, the electronic final examination from 2012 was submitted for assessment during the conference. The subject-area experts classified the electronic final examination at CEFR level B1. In the electronic final examination, the CEFR level refers to both reading and listening (see Chapter 3). Nine variants of the examinations were available, as compiled from an item bank calibrated according to OPLM. Version 1a was used during the conference. This variant contained 31 items, including 17 listening items, 13 reading items and 1 writing assignment. The writing assignment was not included in the standard-setting procedure. Several items in the electronic examination were scored polytomously. If we disregard the writing assignment, the minimum score that could be achieved in the examination would be 0, with a maximum score of 32. The examination was divided into four clusters with the following scoring scales: Cluster 1 (0-8), Cluster 2 (0-8), Cluster 3 (0-7) and Cluster 4 (0-9). The results of the second round of assessments are displayed in Table 5.2. Examination of the assessments for each cluster (Columns C1–C6) reveals considerable agreement amongst the subject-area experts. The percentage of absolute agreement varies from 50% for Cluster 2 to 69% for Clusters 1, 3 and 4. As compared to Table 5.1, the χ_1^2 distances are relatively large. It might have been difficult for the expert panel to arrive at realistic assessments for the various question formats at the same time. The trimmed mean is 19, with a 70% confidence interval of (18; 20). This means that, on this test, students would have to answer 19 ÷ 32 = 59% of the test items correctly in order to achieve CEFR level B1. The subject-area experts provided sufficient support for this cut score. The *Finn coefficient* is .95 and *Gower's similarity coefficient* is .94. Given the reliability of the test, the distribution in the group of subject-area experts was small for the borderline score: χ_2^2 (13, N = 14) = 1.89, p = 1.00.

Table 5.2 Results of standard setting in the central final examination for German VMBO [Bas.Voc.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right]$ = 19)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	19	**6**	**5**	3	**5**	---	---	1.09	2.02	79	0
2	20	**6**	**5**	**4**	**5**	---	---	.72	2.01	.91	0
3	20	**6**	**5**	**4**	**5**	---	---	.72	2.01	.91	0
4	17	5	4	**4**	4	---	---	.61	2.44	.86	0
5	19	5	**5**	**4**	**5**	---	---	.19	2.02	.92	0
6	21	**6**	**5**	5	**5**	---	---	.72	1.93	.86	0
7	18	**6**	4	**4**	4	---	---	1.39	2.28	.90	0
8	20	**6**	4	**4**	6	---	---	1.09	2.01	.83	0
9	20	**6**	**5**	**4**	**5**	---	---	.72	2.01	.91	0
10	20	**6**	4	5	**5**	---	---	1.19	2.01	.82	0
11	19	5	4	**4**	6	---	---	.60	2.02	.67	0
12	19	**6**	4	**4**	**5**	---	---	1.12	2.02	.91	0
13	19	5	**5**	**4**	**5**	---	---	.19	2.02	.92	0
14	23	7	6	5	**5**	---	---	1.14	1.47	.81	0
15	19	**6**	**5**	**4**	4	---	---	1.08	2.02	.82	0
16	16	**6**	3	2	**5**	---	---	2.74	2.66	.81	1

70%-BI = (18, 20) ; *Gower* = .94 ; *Finn* = .95 ; χ_2^2 (13, N = 14) = 1.89, p = 1.00

5.1.2 Pre-vocational secondary education (VMBO) Middle-management vocational track

During the conference, one test from the middle-management vocational track of VMBO was submitted to the subject-area experts for assessment. This was the computer based examination from 2012. The subject-area experts found this examination suitable for determining whether the reading and listening comprehension of students meets the requirements for CEFR level B1. As with the examination used in the basic vocational track of VMBO, nine variants of the examinations were available, as compiled from an item bank calibrated according to OPLM. Version 1a was used during the conference. This variant contained 41 items, including 16 listening items, 24 reading items and 1 writing assignment. The writing assignment was not included in the standard setting procedure. For 39 of the 40 remaining items, the maximum score that students could achieve was 1. For one item, students could earn 2 points. The scoring scale of the examinations used during the conference thus ranged from 0 to 41. The examination is divided into five clusters with the following scoring scales: Cluster 1 (0-7), Cluster 2 (0-8), Cluster 3 (0-11), Cluster 4 (0-7) and Cluster 5 (0-8). Table 5.3 provides an analysis of the data collected in the second round of assessments. According to the subject-area experts, the performance standard for CEFR level B1 is somewhere between Score 17 and Score 29. Although this interval might appear to be quite large at first glance, the size of the interval was largely determined by one remarkably low cut score. The suggestion of Subject-area Expert 13 was five points lower than the mildest suggestion from the other 12 subject-area experts. The χ_1^2 distance reveals that the assessment of Subject-area Expert 13 is empirically unlikely. For example, if we translate the cut scores for each cluster to the scoring scale of the examination, this subject-area expert would argue that a student would have to answer 85% of the items correctly in order to demonstrate CEFR level B1 in one case, while 24% would suffice in another case. The suggestion of Subject-area Expert 13 does not affect the trimmed mean. The same applies to the suggestions of Subject-area Expert 3 and Subject-area Expert 11. Proceeding from the trimmed mean, the performance standard for CEFR level B1 on the electronic examination that was administered in the middle-management vocational track of VMBO in 2012 would correspond to Score 25, with a 70% confidence interval of (23; 27). The *Finn coefficient* (.92) and *Gower's similarity coefficient* (.92) indicate sufficient support in the expert panel for this cut score.

Table 5.3 Results of standard setting in the central final examination for German VMBO [Adv.Voc.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right] = 25$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	24	**5**	**5**	**5**	5	4	---	1.89	3.07	.90	0
2	22	4	**5**	**5**	4	4	---	.99	3.33	.88	0
3	29	**5**	7	6	**6**	**5**	---	1.15	2.15	.70	0
4	27	**5**	6	6	5	**5**	---	1.06	2.50	.81	0
5	27	6	**5**	**5**	**6**	**5**	---	2.91	2.50	.89	0
6	24	**5**	**5**	**5**	5	4	---	1.89	3.07	.90	0
7	24	**5**	**5**	4	5	**5**	---	2.91	3.07	.89	0
8	25	4	6	4	**6**	**5**	---	2.37	2.90	.74	0
9	26	**5**	**5**	**5**	**6**	**5**	---	2.02	2.72	.88	0
10	25	4	4	6	**6**	**5**	---	1.36	2.90	.61	0
11	29	6	6	6	**6**	**5**	---	1.78	2.15	.89	0
12	25	**5**	**5**	**5**	**6**	4	---	2.14	2.90	.88	0
13	17	**5**	3	3	4	2	---	6.73	4.22	.77	1

70%-BI = (23, 27) ; *Gower* = .92 ; *Finn* = .92 ; χ^2_2 (10, *N* = 11) = 4.02, *p* = .95

5.1.3 Pre-vocational secondary education (VMBO) Combined theoretical and vocational track

For the mixed and theoretical tracks of VMBO, the listening comprehension test from 2012 was selected. According to the expert panel, this test is suitable for determining whether the listening comprehension of students meets the requirements for CEFR level B2. The listening comprehension test from 2012 contained 37 dichotomously scored test items. Prior to the conference, the 37 test items were divided into five clusters of varying size. The first cluster contained five test items, the second cluster six, the third cluster seven, the fourth cluster ten and, in the fifth cluster, nine. The clusters were assessed by 16 subject-area experts. The most important results are presented in Table 5.4. Six of the 16 subject-area experts found that the performance standard for CEFR level B2 should be set at Score 26. The other subject-area experts set the border at scores ranging from 22 to 29. As compared to the other experts, Subject-area Experts 1 and 16 set the border very high and very low, respectively. The way in which Subject-area Expert 8 set the cut score is remarkable, considering the behaviour of students in Dutch education. In practice, very few students actually achieve the following scores of the various clusters: (Cluster 1) Score 2, (Cluster 2) Score 2, (Cluster 3) Score 6, (Cluster 4) Score 8, and (Cluster 5) Score 8. The impact scores indicate that the omission of individual subject-area experts from the analysis does not affect the final result. The trimmed mean is 26. The performance standard for CEFR level B2 is thus set at Score 26 (25; 27). There was sufficient agreement amongst the subject-area experts to position the performance standard here. The 70% confidence interval is relatively small, and the *Gower's similarity coefficient* (.92) and the *Finn coefficient* (.90) can be regarded as high. The group of subject-area experts can also be considered a realistic sample of an imaginary population of all possible subject-area experts: χ^2_2 (13, N = 14) = 1.93, p = 1.00.

Table 5.4 Results of standard setting in the listening comprehension test for German VMBO [Comb./Th.] (CEFR = B1, $\lceil \bar{C}_{\text{total}} \rceil = 26$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	29	4	**5**	**6**	**7**	**7**	---	.08	1.39	-.24	0
2	26	**3**	**5**	**6**	6	6	---	.48	1.99	.47	0
3	26	**3**	4	5	**7**	**7**	---	.16	1.99	.43	0
4	27	**3**	4	**6**	**7**	**7**	---	.21	1.83	.50	0
5	26	**3**	4	5	**7**	**7**	---	.16	1.99	.43	0
6	25	**3**	**5**	5	5	**7**	---	.53	2.16	.22	0
7	27	**3**	**5**	**6**	**7**	6	---	.46	1.83	.53	0
8	26	2	2	**6**	8	8	---	2.53	1.99	.27	0
9	25	**3**	4	5	**7**	6	---	.25	2.16	.48	0
10	28	**3**	**5**	**6**	**7**	**7**	---	.21	1.59	.55	0
11	26	4	4	5	6	**7**	---	.11	1.99	-.66	0
12	27	**3**	**5**	**6**	**7**	6	---	.46	1.83	.53	0
13	27	**3**	4	**6**	6	8	---	.34	1.83	.29	0
14	26	**3**	**5**	5	**7**	6	---	.43	1.99	.45	0
15	24	**3**	4	5	6	6	---	.08	2.36	.55	0
16	22	**3**	2	**6**	6	5	---	1.41	2.73	.12	0

70%-BI = (25, 27) ; *Gower* = .92 ; *Finn* = .90 ; χ_2^2 (13, N = 14) = 1.93, p = 1.00

In addition to a listening comprehension test, the final examination from 2012 was submitted for assessment during the conference. The subject-area experts classified the examination at CEFR level B1. In contrast to the electronic final examinations for the basic and middle-management vocational track of VMBO, the paper-based final examination for the combined theoretical and vocational track of VMBO contained only reading items. The test contained 43 items, most of which were scored dichotomously. For substantive and/or psychometric reasons, three items were eliminated from the examination before it was divided into clusters. Five clusters were created with the following scoring scales: Cluster 1 (0-8), Cluster 2 (0-9), Cluster 3 (0-10), Cluster 4 (0-8) and Cluster 5 (0-9). On the full test, students could earn a maximum of 44 points. Table 5.5 provides an analysis of the data collected during the second round of assessments. The suggestions of the subject-area experts varied considerably. A relatively large share of the subject-area experts positioned the performance standard at Scores 20 or 21, although other subject-area experts set the performance standard substantially higher (27) or lower (15). The percentage of absolute agreement fluctuates around .50. None of the scoring patterns demonstrated significant difference (at the 10% level) from what could be expected under the assumptions of the item response model. If we disregard one randomly selected maximum assessment and one randomly selected minimum assessment, the performance standard for CEFR level B1 would be located at Score 21, with a 70% confidence interval of (19; 24). *Gower's similarity coefficient* (.90) and the *Finn coefficient* (.89) were relatively high. The distribution in the group of subject-area experts is realistic, given the reliability of the test at the borderline score. χ_2^2 (13, N = 14) = 6.74, p = .92.

Table 5.5 Results of standard setting in the central final examination for German VMBO [Comb./Th.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right] = 21$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	19	**5**	4	**4**	3	**3**	---	1.39	4.03	.59	0
2	23	**5**	**5**	5	**4**	4	---	.90	3.34	.67	0
3	27	**5**	6	6	5	5	---	.62	2.60	.65	0
4	21	4	6	**4**	**4**	**3**	---	1.48	3.76	.42	0
5	19	4	6	5	3	1	---	2.15	4.03	-.08	0
6	23	4	**5**	**4**	**4**	6	---	2.16	3.34	.46	0
7	21	4	**5**	**4**	3	5	---	1.01	3.76	.49	0
8	20	4	**5**	**4**	**4**	**3**	---	1.71	3.89	.59	0
9	27	6	6	5	**4**	6	---	1.07	2.60	.55	0
10	20	**5**	**5**	**4**	3	**3**	---	.90	3.89	.55	0
11	20	4	**5**	**4**	3	4	---	.58	3.89	.68	0
12	20	4	4	3	**4**	5	---	3.28	3.89	.63	0
13	21	3	**5**	**4**	**4**	5	---	2.19	3.76	.33	0
14	24	**5**	6	5	**4**	4	---	.56	3.15	.64	0
15	15	**5**	3	3	1	**3**	---	2.09	4.78	.32	1
16	21	**5**	5	**4**	3	4	---	.80	3.76	.63	0

70%-BI = (19, 24) ; *Gower* = .90 ; *Finn* = .89 ; χ_2^2 (13, *N* = 14) = 6.74, *p* = .92

5.1.4 Senior general secondary education (HAVO)

A performance standard was established for students in senior general secondary education (HAVO), based on the listening comprehension test that was administered in schools in 2013. The substance subject-area experts found that this test was best suited to the *can-do statements* for CEFR level C1. The test contained 40 items, all of which were scored dichotomously. Prior to the conference, the 40 test items were divided into four clusters of varying size. Eight test items were placed in the first cluster, with 12 in the second cluster, 9 in the third cluster and 11 in the fourth cluster. The cut scores selected by the subject-area experts for the various clusters and for the full tests are displayed in Table 5.6. This table also provides an analysis of inter-rater agreement at the individual level. The percentage of absolute agreement in the various clusters is quite high. For each cluster, at least 60% of the subject-area experts selected the same cut score. None of the patterns of cut scores differ significantly from what might be expected based on the behaviour of students in Dutch education. Although Subject-area Experts 16 and 10 set the performance standard very low and very high, respectively, they had no significant effect on the final result. The impact scores are 0. If we were to determine the location of the performance standard for CEFR level C1 based on the suggestions of the subject-area experts, we would arrive at Score 29. The corresponding 70% confidence interval is small: $\left(\left[\bar{C}_{\text{total}}\right] - 1.036 \times \sigma_{C_{\text{total}}}; \left[\bar{C}_{\text{total}}\right] + 1.036 \times \sigma_{C_{\text{total}}}\right) = (28; 30)$. Thus, *Gower's similarity coefficient* (.95) and the *Finn coefficient* (.97) are high. The distribution in the group of subject-area experts is small, given the reliability of the test at the borderline score. χ_2^2 (13, *N* = 14) = 1.60, *p* = 1.00. In summary, this means that there is sufficient support for considering the listening comprehension of students in HAVO as consistent with the C1 level, beginning with Score 29. Students would have to answer 72.5% of the test items correctly in order to achieve this level.

Table 5.6 Results of the standard setting procedure for the listening comprehension test for German HAVO (CEFR = C1, $\lceil \bar{C}_{\text{total}} \rceil = 29$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	29	**6**	**9**	**6**	**8**	---	---	.09	1.57	.55	0
2	28	**6**	**9**	**6**	7	---	---	.01	1.68	.00	0
3	29	7	**9**	**6**	7	---	---	.22	1.57	.00	0
4	31	**6**	10	7	**8**	---	---	.05	1.30	-.19	0
5	29	**6**	**9**	**6**	**8**	---	---	.09	1.57	.55	0
6	29	**6**	8	**6**	9	---	---	.59	1.57	.57	0
7	28	**6**	8	**6**	**8**	---	---	.27	1.68	.55	0
8	29	**6**	**9**	**6**	**8**	---	---	.09	1.57	.55	0
9	29	**6**	**9**	**6**	**8**	---	---	.09	1.57	.55	0
10	32	7	10	7	**8**	---	---	.03	1.07	-.09	0
11	29	**6**	**9**	**6**	**8**	---	---	.09	1.57	.55	0
12	28	**6**	8	**6**	**8**	---	---	.27	1.68	.55	0
13	29	5	**9**	7	**8**	---	---	.31	1.57	.09	0
14	29	5	**9**	**6**	9	---	---	.57	1.57	.37	0
15	31	**6**	**9**	7	9	---	---	.27	1.30	.44	0
16	23	4	8	5	6	---	---	.12	2.33	.00	0

70%-BI = (28, 30) ; *Gower* = .95 ; *Finn* = .97 ; χ_2^2 (13, *N* = 14) = 1.60, *p* = .92

The performance standard for reading comprehension at the HAVO level was based on the first-term German examination from 2012. The subject-area experts rated this examination, which was administered on paper, at CEFR level B2. The test consisted of 42 items, most of which were scored dichotomously. For eight of the test items students could achieve more than one point. Thus, the maximum score that could be achieved on the final examination was 50. For the standard-setting procedure, the 42 test items were divided into five clusters with the following scoring scales: Cluster 1 (0-10), Cluster 2 (0-9), Cluster 3 (0-10) and Cluster 5 (0-12). During the conference, the five clusters were submitted to 15 different subject-area experts for assessment. Table 5.7 provides an analysis of the data collected during the second round of assessments. As shown in these results, one suggestion was extremely high, and another was extremely low. Disregarding these two suggestions, as in the calculation of the trimmed mean, the expert panel set the performance standard for CEFR level B2 for this examination between Score 26 and Score 31. The lower and upper limits are quite close to each other. One subject-area expert proposed that a student should answer 26 ÷ 50 = 52% of the test items correctly in order to demonstrate CEFR level B2, another subject-area expert demanded slightly more: 31 ÷ 50 = 62%. The trimmed mean is 29 (27:31). There was sufficient support amongst the subject-area experts to locate the performance standard at this point. Both the absolute and the relative agreement between subject-area experts are high: *Gower's similarity coefficient* is .95 and the *Finn coefficient* is .97.

Table 5.7 Results of standard setting in the central final examination for German HAVO (CEFR = B2, $\left[\bar{C}_{\text{total}}\right] = 29$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6		α=.10	RSI	Impact
1	35	8	6	**6**	6	9	---	.50	2.12	.66	-1
2	28	**7**	**5**	5	**5**	6	---	.34	3.13	.61	0
3	31	**7**	**5**	**6**	**5**	8	---	.32	2.68	.49	0
4	29	**7**	**5**	5	**5**	**7**	---	.42	2.98	.70	0
5	28	**7**	**5**	**6**	**5**	5	---	.31	3.13	.06	0
6	28	6	**5**	5	**5**	**7**	---	.49	3.13	.57	0
7	30	**7**	**5**	**6**	**5**	**7**	---	.10	2.86	.62	0
8	26	6	4	5	**5**	6	---	.17	3.41	.46	0
9	31	**7**	**5**	**6**	6	**7**	---	.19	2.68	.36	0
10	31	**7**	**5**	**6**	**5**	8	---	.32	2.68	.49	0
11	28	**7**	**5**	5	**5**	6	---	.34	3.13	.61	0
12	26	**7**	4	4	**5**	6	---	.70	3.41	.56	0
13	28	6	**5**	5	**5**	**7**	---	.49	3.13	.57	0
14	30	**7**	**5**	**6**	**5**	**7**	---	.10	2.86	.62	0
15	24	8	3	**6**	3	4	---	1.68	3.69	-.22	0

70%-BI = (27, 31) ; *Gower* = .95 ; *Finn* = .97 ; χ^2_2 (12, *N* = 13) = 3.20, *p* = .99

5.1.5 University preparatory education (VWO)

During the conference, the experts assessed two tests that are used in university preparatory education (VWO): the listening comprehension test from 2012 and the central examination from 2010. There was sufficient support for rating the listening comprehension test for VWO at CEFR level C2. The test that was administered to VWO students in 2012 consisted of 38 multiple-choice items, which were scored dichotomously. All test items were included in the process of establishing the performance standard. Five clusters were composed with the following scoring scales: Cluster 1 (0-5), Cluster 2 (0-5), Cluster 3 (0-9), Cluster 4 (0-9) and Cluster 5 (0-10). The five clusters together yielded a minimum score of 0 on the full test and a maximum score of 38. The most important results from the conference are presented in Table 5.8. There was considerable consensus amongst the 17 subject-area experts in the assessment. Only in the fifth cluster did their opinions vary. Some subject-area experts proposed that a student should answer 5 of the 10 test items correctly (expected overall score = 19), other experts set the bar at 8 of the 10 test items (expected overall score = 30). The pattern of cut scores suggested by Subject-area Expert 16 differs (at the 10% significance level) from what could be expected according to the item response model. No individual assessment differed so much from the others that the performance standard would change when disregarding the assessment in question; all impact scores were 0. Disregarding one randomly selected maximum assessment and one randomly selected minimum assessment, the performance standard for CEFR level C2 would be located at Score 28, with a 70% confidence interval of (26; 29). The level of consensus within the expert panel was sufficient to set the performance standard at this scoring point. The upper and lower limits of the 70% confidence interval are relatively close to each other, with relatively high values for *Gower's similarity coefficient* (.92) and the *Finn coefficient* (.89). The distribution within the expert panel is realistic, given the reliability of the test at the cut score: χ^2_2 (14, N = 15) = 3.50, p = .99. In practice, this means that VWO students can achieve the performance standard for CEFR level C2 by answering 28 ÷ 38 = 74% of the test items correctly.

Table 5.8 Results of standard setting in the listening comprehension test for German VWO (CEFR = C2, $\left[\bar{C}_{\text{total}}\right] = 28$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	30	**5**	**4**	7	7	**7**	---	.25	1.46	.70	0
2	28	**5**	**4**	**6**	**6**	**7**	---	.28	1.81	.76	0
3	30	**5**	5	**6**	7	**7**	---	.61	1.46	.69	0
4	27	4	**4**	**6**	7	6	---	.34	2.03	.01	0
5	29	**5**	**4**	**6**	**6**	8	---	.25	1.62	.71	0
6	28	**5**	**4**	5	**6**	8	---	.42	1.81	.66	0
7	29	**5**	**4**	**6**	**6**	8	---	.25	1.62	.71	0
8	24	4	**4**	5	5	6	---	.46	2.58	.46	0
9	27	**5**	3	**6**	**6**	**7**	---	.28	2.03	.53	0
10	27	**5**	**4**	**6**	5	**7**	---	.65	2.03	.71	0
11	28	**5**	3	**6**	**6**	8	---	.31	1.81	.44	0
12	29	**5**	**4**	7	**6**	**7**	---	.42	1.62	.73	0
13	27	**5**	**4**	**6**	**6**	6	---	.50	2.03	.76	0
14	27	**5**	**4**	5	**6**	**7**	---	.37	2.03	.73	0
15	30	**5**	**4**	**6**	7	8	---	.09	1.46	.70	0
16	27	**5**	5	4	8	5	---	2.29	2.03	.48	0
17	25	4	**4**	**6**	**6**	5	---	.75	2.40	.32	0

70%-BI = (26, 29) ; *Gower* = .92 ; *Finn* = .89 ; χ_2^2 (14, *N* = 15) = 3.50, *p* = 1.00

The performance standard for reading comprehension in VWO was determined based on the central examination from 2010. The expert panel considered this exam suitable for determining whether the reading comprehension of students corresponds to the descriptors formulated for CEFR level C1. The examination from 2010 consisted of 46 items, all but four of which were scored dichotomously. Students could achieve a maximum of 51 points on the final examination. Prior to the conference, the items were divided into six clusters with the following scoring scales: Cluster 1 (0-8), Cluster 2 (0-9), Cluster 3 (0-11), Cluster 4 (0-8) and Cluster 6 (0-7). The clusters were assessed by 16 subject-area experts. Table 5.9 provides an analysis of the data collected during the second round of assessments. Based on the trimmed mean and the corresponding 70% confidence interval, the subject-area experts apparently agreed with each other to a large extent: $\left[\bar{C}_{\text{total}}\right] = 29$ (26; 32). The lower limit of the 70% confidence interval corresponds to 26 ÷ 51 = 51% correct answers, and the upper limit corresponds to 32 ÷ 51 = 63% correct answers. *Gower's similarity coefficient* for absolute inter-rater agreement is .91.
The *Finn coefficient* for relative inter-rater agreement is .91. With the exclusion of Subject-area Expert 16 and one *randomly* selected maximum assessment, the expert panel can be regarded as a realistic sample of the imaginary population of all possible subject-area experts:
χ_2^2 (13, N = 14) = 8.45, p = .81.

Table 5.9 Results of standard setting in the central final examination for German VWO (CEFR = C1, $\lceil \bar{C}_{\text{total}} \rceil = 29$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	33	5	6	7	**5**	5	**5**	.57	2.99	.57	0
2	27	5	4	**6**	4	**4**	4	.74	4.01	.49	0
3	32	5	**5**	8	**5**	**4**	**5**	1.03	3.20	.48	0
4	27	**4**	**5**	**6**	4	5	3	.28	4.01	.22	0
5	28	**4**	4	7	**5**	**4**	4	1.00	3.84	.47	0
6	30	6	**5**	**6**	4	**4**	**5**	1.60	3.51	.49	0
7	24	**4**	4	4	**5**	**4**	3	2.03	4.52	.47	1
8	28	**4**	**5**	**6**	**5**	**4**	4	1.04	3.84	.58	0
9	29	**4**	**5**	**6**	**5**	**4**	**5**	1.43	3.67	.61	0
10	32	5	**5**	7	**5**	6	4	.09	3.20	.20	0
11	26	**4**	4	4	4	5	**5**	1.93	4.22	.44	0
12	28	**4**	**5**	7	4	**4**	4	.60	3.84	.47	0
13	30	**4**	**5**	**6**	**5**	5	**5**	.85	3.51	.56	0
14	33	5	**5**	7	**5**	6	**5**	.18	2.99	.47	0
15	33	**4**	**5**	8	**5**	6	**5**	.33	2.99	.11	0
16	13	**4**	2	1	0	**4**	2	5.90	6.49	-.22	1

70%-BI = (26, 32) ; *Gower* = .92 ; *Finn* = .91 ; χ_2^2 (13, *N* = 14) = 8.45, *p* = .81

5.2 English

For English, five central examinations and four listening comprehension tests were assessed by a fixed panel of 19 subject-area experts. In the following subsections, we present the results for the basic vocational track of pre-vocational secondary education (VMBO; 5.2.1); the middle-management vocational track of VMBO (5.2.2); the combined theoretical and vocational track of VMBO (5.2.3); senior general secondary education (HAVO; 5.2.4); and university preparatory education (VWO; 5.2.5). A summary of the results is presented in Table 5.25b.

5.2.1 Pre-vocational secondary education (VMBO) Basic vocational track

During the conference, the experts assessed two tests that are used in the basic vocational track of VMBO: the listening comprehension test from 2012 and the central examination from 2012. According to the expert panel, the listening comprehension test is suitable for determining whether the listening skills of students are located at CEFR level B1. The test contained 34 dichotomously scored items. For the standard setting procedure, the 34 test items were divided into four clusters with the following scoring scales: Cluster 1 (0-9), Cluster 2 (0-9), Cluster 3 (0-11) and Cluster 4 (0-5). The subject-area experts set the performance standard for CEFR level B1 on this test at Score 22. This means that students would have to earn at least 22 of the 34 points in order to demonstrate CEFR level B1. This number was derived by summing up and averaging the assessments of the subject-area experts for each cluster and averaging these numbers. The highest and lowest overall scores were removed. Table 5.10 provides a detailed analysis of the data collected during the second round of assessments. The distribution in the assessments of the subject-area experts was used to calculate the precision of the performance standard. If we express the level of precision as a 70% confidence interval, we arrive at the following formula: $\left(\lceil \bar{C}_{\text{total}} \rceil - 1.036 \times \sigma_{C_{\text{total}}} ; \lceil \bar{C}_{\text{total}} \rceil + 1.036 \times \sigma_{C_{\text{total}}}\right) = (20; 23)$. When examining the

impact of the individual assessments on the performance standard, it is interesting to note that the exclusion of Subject-area Experts 1, 6, 15, 16 or 17 would raise the performance standard by one point. As compared to the other experts, these subject-area experts considered the listening comprehension test relatively difficult, and they were therefore of the opinion that students should not be expected to score this high in order to meet the requirements of CEFR level B1. The expert panel can be regarded as a realistic sample of an imaginary population of all possible subject-area experts: χ^2_2 (16, N = 17) = 4.770, p = 1.00. *Gower's similarity coefficient* for absolute inter-rater agreement is .92. The *Finn coefficient* for relative inter-rater agreement is .92.

Table 5.10 Results of standard setting in the listening comprehension test for English VMBO [Bas. Voc.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right]$ = 22)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	17	**6**	5	5	1	---	---	.40	3.04	-.54	1
2	22	**6**	**6**	**7**	**3**	---	---	.44	2.37	.48	0
3	23	7	7	6	**3**	---	---	.29	2.02	.51	0
4	22	**6**	**6**	**7**	**3**	---	---	.44	2.37	.48	0
5	22	**6**	**6**	**7**	**3**	---	---	.44	2.37	.48	0
6	19	**6**	5	5	**3**	---	---	.77	2.84	.44	1
7	23	7	7	6	**3**	---	---	.29	2.02	.51	0
8	22	7	**6**	6	**3**	---	---	.31	2.37	.45	0
9	23	**6**	7	**7**	**3**	---	---	.42	2.02	.52	0
10	22	**6**	7	6	**3**	---	---	.55	2.37	.56	0
11	23	7	8	5	**3**	---	---	.96	2.02	.33	0
12	22	**6**	7	6	**3**	---	---	.55	2.37	.56	0
13	22	7	**6**	**7**	2	---	---	.07	2.37	-.40	0
14	22	**6**	**6**	**7**	**3**	---	---	.44	2.37	.48	0
15	19	5	**6**	5	**3**	---	---	1.10	2.84	.54	1
16	20	**6**	7	5	2	---	---	.63	2.66	.24	1
17	18	5	5	6	2	---	---	.18	3.01	.38	1
18	24	7	7	**7**	**3**	---	---	.12	1.85	.56	0
19	23	8	**6**	**7**	2	---	---	.19	2.02	-.65	0

70%-BI = (20, 23) ; *Gower* = .92 ; *Finn* = .92 ; χ^2_2 (16, N = 17) = 4.77, p = 1.00

In addition to a listening comprehension test, the electronic examination from 2012 was submitted to the expert panel for assessment during the conference. The subject-area experts classified the electronic examination at CEFR level B1. In the electronic examination, the CEFR level refers to both reading and listening (see Section 3). Nine variants of the examinations were available, as compiled from an item bank calibrated according to OPLM. Version 1a was used during the conference. This variant contained 29 items, including 13 listening items, 15 reading items and 1 writing assignment. The writing assignment was not included in the standard setting procedure. Several items in the electronic examination were scored polytomously. If we disregard the writing assignment, the minimum score that could be achieved in the examination would be 0, with a maximum score of 32. Prior to the conference, the examination was divided into four clusters with the following scoring scales: Cluster 1 (0-7), Cluster 2 (0-8), Cluster 3 (0-8) and Cluster 4 (0-9). During the conference, the four clusters were submitted to 19 different subject-area experts for assessment. Table 5.11 provides an analysis of the data collected during the second round of assessments. The subject-area experts located the performance standard for CEFR level B1 at Score 24. As shown in Table 5.11, the subject-area experts selected this

performance standard with considerable consensus. The 70% confidence interval is only one point above and below the selected performance standard. The different measures of inter-rater agreement confirm this image. *Gower's similarity coefficient* for absolute inter-rater agreement is .95, and the *Finn coefficient* for relative inter-rater agreement is .96. The distribution in the group of subject-area experts is small, given the reliability of the test at the borderline score. χ^2_2 (16, *N* = 17) = 2.67, *p* = 1.00. The impact of the individual assessments on the performance standard is shown in the last column of Table 5.11. The impact for all subject-area experts is 0. This confirms that there was sufficient support within the expert panel to set the performance standard for CEFR level B1 at Score 24.

Table 5.11 Results of standard setting in the central final examination for English VMBO [Bas.Voc.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right] = 24$

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	23	4	**6**	**6**	**7**	---	---	.05	1.83	-.06	0
2	23	4	**6**	**6**	**7**	---	---	.05	1.83	-.06	0
3	25	**5**	**6**	**6**	8	---	---	.06	1.36	.12	0
4	25	**5**	**6**	**6**	8	---	---	.06	1.36	.12	0
5	22	4	5	**6**	**7**	---	---	.11	1.96	.23	0
6	22	4	5	**6**	**7**	---	---	.11	1.96	.23	0
7	24	4	**6**	7	**7**	---	---	.24	1.55	.04	0
8	25	**5**	7	**6**	**7**	---	---	.15	1.36	-.35	0
9	24	**5**	**6**	**6**	**7**	---	---	.03	1.55	-.06	0
10	24	**5**	**6**	**6**	**7**	---	---	.03	1.55	-.06	0
11	24	**5**	5	7	**7**	---	---	.33	1.55	.18	0
12	23	4	**6**	**6**	**7**	---	---	.05	1.83	-.06	0
13	25	**5**	**6**	**6**	8	---	---	.06	1.36	.12	0
14	24	**5**	**6**	**6**	**7**	---	---	.03	1.55	-.06	0
15	21	4	5	**6**	6	---	---	.13	2.14	.13	0
16	27	6	7	**6**	8	---	---	.15	.93	-.21	0
17	24	**5**	**6**	**6**	**7**	---	---	.03	1.55	-.06	0
18	22	4	5	**6**	**7**	---	---	.11	1.96	.23	0
19	25	**5**	**6**	7	**7**	---	---	.12	1.36	.06	0

70%-BI = (23, 25) ; *Gower* = .95 ; *Finn* = .96 ; χ^2_2 (16, *N* = 17) = 2,67, *p* = 1.00

5.2.2 Pre-vocational secondary education (VMBO) Middle-management vocational track

During the conference, one of the tests from the middle-management vocational track of VMBO was submitted to the subject-area experts for assessment. This was the electronic examination from 2012. The subject-area experts found this examination suitable for determining whether the reading and listening comprehension of students meets the requirements for CEFR level B1. As with the examination used in the basic vocational track of VMBO, nine variants of the examinations were available, as compiled from an item bank calibrated according to OPLM. Version 1a was used during the conference. This variant contained 33 items, including 14 listening items, 18 reading items and 1 writing assignment. The writing assignment was not included in the standard setting procedure. Most of the items were scored dichotomously. Disregarding the writing assignment, students could earn a total of 39 points on the examination. These 39 points were divided into five clusters with the following scoring scales: Cluster 1 (0-10), Cluster 2 (0-7), Cluster 3 (0-8), Cluster 4 (0-7) and Cluster 5 (0-7).

The clusters were assessed by 19 subject-area experts. The most important results are presented in Table 5.12. The performance standard is located at $\lceil \bar{c}_{\text{total}} \rceil = 24$, with the corresponding 70% confidence interval between Score 22 and Score 26. These results also indicate that Subject-area Expert 6 estimated the performance required to demonstrate CEFR level B1 substantially lower than did the other subject-area experts. The impact value of the assessment of Subject-area Expert 6 on the performance standard is 1. Nevertheless, there is sufficient support amongst the subject-area experts to set the performance standard for CEFR level B1 at Score 24. *Gower's similarity coefficient* for absolute inter-rater agreement is .91, and the *Finn coefficient* for relative inter-rater agreement is .90. The distribution in the group of subject-area experts is realistic, given the reliability of the test at the cut score. χ^2_2 (16, *N* = 17) = 7.75, *p* = .96.

Table 5.12 Results of standard setting in the central final examination for English VMBO [Adv.Voc.] (CEFR = B1, $\lceil \bar{c}_{\text{total}} \rceil = 24$)

Expert	c_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	23	**5**	**5**	5	3	**5**	---	.90	2.71	.82	0
2	23	**5**	**5**	**6**	3	4	---	.44	2.71	.79	0
3	22	**5**	**5**	**6**	2	4	---	.63	2.87	.22	0
4	25	**5**	**5**	5	5	**5**	---	2.18	2.36	.78	0
5	26	6	6	**6**	3	**5**	---	.22	2.19	.67	0
6	19	**5**	4	4	3	3	---	.40	3.38	.60	1
7	27	6	4	7	**4**	6	---	1.80	2.08	.79	0
8	28	6	6	7	**4**	**5**	---	.33	1.92	.79	0
9	23	**5**	4	**6**	**4**	4	---	1.43	2.71	.78	0
10	22	**5**	**5**	5	3	4	---	.34	2.87	.86	0
11	22	2	**5**	**6**	**4**	**5**	---	4.32	2.87	.72	0
12	24	**5**	**5**	5	**4**	**5**	---	1.27	2.53	.82	0
13	27	6	6	**6**	**4**	**5**	---	.35	2.08	.83	0
14	27	6	6	**6**	**4**	**5**	---	.35	2.08	.83	0
15	23	**5**	4	5	**4**	**5**	---	1.90	2.71	.84	0
16	26	**5**	**5**	7	**4**	**5**	---	1.18	2.19	.81	0
17	23	**5**	**5**	5	3	**5**	---	.90	2.71	.82	0
18	25	**5**	4	**6**	**4**	6	---	2.38	2.36	.86	0
19	26	**5**	**5**	**6**	**4**	6	---	1.64	2.19	.85	0

70%-BI = (22, 26) ; *Gower* = .91 ; *Finn* = .90 ; χ^2_2 (16, *N* = 17) = 7.75, *p* = .96

5.2.3 Pre-vocational secondary education (VMBO) Combined theoretical and vocational track

For the mixed and theoretical tracks of VMBO, the listening comprehension test from 2012 was selected. According to the expert panel, this test is suitable for determining whether the listening comprehension of students is located at CEFR level B2. The listening comprehension test from 2012 consisted of 36 dichotomously scored items, which were divided into six relatively small clusters before the conference. The first cluster contained 7 test items, with 6 in the second cluster, 6 in the third cluster, 4 in the fourth cluster, 9 in the fifth cluster and 4 in the sixth cluster. The clusters were assessed by 19 subject-area experts. The most important results are presented in Table 5.13. According to the subject-area experts, the performance standard for CEFR level B2 is somewhere between Score 21 and Score 26. Whereas one subject-area expert argued that a student should answer 21 ÷ 36 = 58% of the test items correctly in order to demonstrate CEFR level B2, another subject-area expert proposed that CEFR level B2 could not

be confirmed unless a student answers 26 ÷ 36 = 72% of the test items correctly. Disregarding one *randomly* selected maximum assessment and one *randomly* selected minimum assessment, the performance standard for CEFR level B2 on this test would be located at $\lceil \bar{C}_{\text{total}} \rceil$ = 24 (23; 26). The performance standard would be in exactly the same place if all of the assessments were included. This underlines the unanimity amongst the subject-area experts in the assessment of this test. This consensus is also reflected in *Gower's similarity coefficient* (.94) and the *Finn coefficient (*.93).

Table 5.13 Results of standard setting in the listening comprehension test for English VMBO [Comb./Th.] (CEFR = B2, $\lceil \bar{C}_{\text{total}} \rceil$ = 24)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	24	**5**	**4**	**4**	**3**	5	**3**	.15	2.60	.43	0
2	25	**5**	**4**	**4**	**3**	**6**	**3**	.10	2.38	.64	0
3	25	**5**	**4**	**4**	**3**	**6**	**3**	.10	2.38	.64	0
4	24	**5**	3	**4**	**3**	**6**	**3**	.21	2.60	.58	0
5	23	**5**	3	**4**	**3**	5	**3**	.14	2.85	.44	0
6	23	4	**4**	**4**	**3**	**6**	2	.81	2.85	.45	0
7	24	4	**4**	**4**	**3**	**6**	**3**	.27	2.60	.31	0
8	25	6	**4**	**4**	**3**	5	**3**	.41	2.38	.33	0
9	21	4	3	3	2	**6**	**3**	.73	3.31	.34	0
10	26	**5**	**4**	5	**3**	**6**	**3**	.28	2.14	.59	0
11	22	**5**	3	3	**3**	5	**3**	.13	3.08	.05	0
12	25	**5**	3	**4**	**3**	7	**3**	.53	2.38	.54	0
13	26	7	3	**4**	**3**	**6**	**3**	1.13	2.14	.37	0
14	24	**5**	3	**4**	**3**	**6**	**3**	.21	2.60	.58	0
15	26	**5**	**4**	5	**3**	**6**	**3**	.28	2.14	.59	0
16	23	4	3	**4**	**3**	**6**	**3**	.30	2.85	.33	0
17	25	**5**	**4**	**4**	**3**	**6**	**3**	.10	2.38	.64	0
18	25	**5**	**4**	**4**	**3**	**6**	**3**	.10	2.38	.64	0
19	26	4	**4**	5	4	**6**	**3**	.57	2.14	-.16	0

70%-BI = (23, 26) ; *Gower* = .94 ; *Finn* = .93 ; χ_2^2 (16, *N* = 17) = 3.02, *p* = 1.00

The performance standard for reading comprehension in the combined theoretical and vocational track of pre-vocational secondary education (VMBO) were determined based on the English examination from 2012. The subject-area experts classified the examination at CEFR level B2. The examination contains a total of 30 questions. In addition to a writing assignment, which was not included in the standard setting procedure, the examination consisted entirely of reading items. For substantive and/or psychometric reasons, one of these reading items was not included in the standard-setting procedure. In all, students could achieve 33 points on the remaining 28 items. Prior to the conference, the items were divided into five clusters with the following scoring scales: Cluster 1 (0-8), Cluster 2 (0-7), Cluster 3 (0-6), Cluster 4 (0-7) and Cluster 5 (0-5). Table 5.14 provides an analysis of the data collected during the second round of assessments. According to the expert panel, students would have to earn at least 25 of the 33 points in order to demonstrate CEFR level B2. In practice, this means that students in the combined theoretical and vocational track of VMBO meet the performance standard for CEFR level B2 if they earn 25 ÷ 33 = 76% of the points. If we use a 70% confidence interval to consider the uncertainty surrounding this performance standard, we see that students must earn between 24 ÷ 33 = 73% and 26 ÷ 33 = 79% of the points in order to demonstrate CEFR level B2.

The confidence interval can be regarded as very small. This can be observed in the measures of inter-rater agreement: *Gower's similarity coefficient* is .96 and the *Finn coefficient* is .97. The distribution in the group of subject-area experts is small, given the reliability of the test at the cut score. χ^2_2 (16, *N* = 17) = 1.47, *p* = 1.00.

Table 5.14 Results of standard setting in the central final examination for English VMBO [Comb./ Th.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right]$ = 25)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
2	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
3	25	**7**	**5**	5	**5**	**3**	---	.19	1.70	.41	0
4	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
5	26	**7**	**5**	5	6	**3**	---	.13	1.43	.25	0
6	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
7	25	**7**	**5**	**4**	6	**3**	---	.12	1.70	.41	0
8	25	**7**	**5**	**4**	6	**3**	---	.12	1.70	.41	0
9	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
10	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
11	26	**7**	6	**4**	6	**3**	---	.22	1.43	.13	0
12	25	**7**	**5**	5	**5**	**3**	---	.19	1.70	.41	0
13	25	**7**	6	**4**	**5**	**3**	---	.25	1.70	.27	0
14	26	**7**	**5**	5	6	**3**	---	.13	1.43	.25	0
15	25	**7**	**5**	5	**5**	**3**	---	.19	1.70	.41	0
16	25	**7**	6	**4**	**5**	**3**	---	.25	1.70	.27	0
17	24	**7**	**5**	**4**	**5**	**3**	---	.08	1.91	.65	0
18	25	**7**	**5**	5	**5**	**3**	---	.19	1.70	.41	0
19	26	8	**5**	**4**	6	**3**	---	.38	1.43	.60	0

70%-BI = (24, 26) ; *Gower* = .96 ; *Finn* = .97 ; χ^2_2 (16, *N* = 17) = 1.47, *p* = 1.00

5.2.4 Senior general secondary education (HAVO)

A performance standard was established for students in senior general secondary education, based on the listening comprehension test that was administered in schools in 2012. The listening comprehension test from 2012 contained 39 dichotomously scored test items. For the standard-setting procedure, the test was divided into four clusters, consisting of 11, 8, 12 and 8 items, respectively. According to the expert panel, this test is suitable for determining whether the listening comprehension of students is located at CEFR level C1. In order to demonstrate CEFR level C1, the expert panel deemed that students should be able to answer at least 32 of the 39 test items correctly. This corresponds to 32 ÷ 39 = 82% of the test items. Table 5.15 provides a detailed analysis of the data collected during the conference. According to this analysis, the expert panel estimated the CEFR level unanimously. The χ^2_1 distances are small, the impact of individual expert assessments on the final result is 0 in all cases, and the variation of the assessments is small. The lower limit of the 70% confidence interval is located at Score 31, with the upper limit at Score 33. With values of .93 (*Gower's similarity coefficient*) and .94 (*Finn coefficient*), the two measures for inter-rater agreement can be regarded as high.

Table 5.15 Results of standard setting in the listening comprehension test for English HAVO (CEFR = B1, $\left[\bar{C}_{\text{total}}\right] = 32$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	30	**9**	6	9	6	---	---	.04	1.34	.17	0
2	34	10	**7**	**10**	**7**	---	---	.09	.79	.35	0
3	35	10	**7**	11	**7**	---	---	.00	.68	.30	0
4	31	**9**	6	9	**7**	---	---	.15	1.21	.16	0
5	31	**9**	6	9	**7**	---	---	.15	1.21	.16	0
6	31	10	6	9	6	---	---	.22	1.21	.04	0
7	31	**9**	6	**10**	6	---	---	.02	1.21	-.33	0
8	29	**9**	6	9	5	---	---	.24	1.57	-.09	0
9	33	10	6	**10**	**7**	---	---	.12	.91	-.05	0
10	33	**9**	**7**	**10**	**7**	---	---	.12	.91	.30	0
11	31	**9**	**7**	**10**	5	---	---	.41	1.21	-.05	0
12	33	**9**	**7**	11	6	---	---	.13	.91	-.08	0
13	33	**9**	**7**	**10**	**7**	---	---	.12	.91	.30	0
14	32	**9**	**7**	**10**	6	---	---	.12	1.02	.14	0
15	33	**9**	**7**	**10**	**7**	---	---	.12	.91	.30	0
16	30	8	**7**	9	6	---	---	.29	1.34	.26	0
17	33	**9**	**7**	**10**	**7**	---	---	.12	.91	.30	0
18	32	**9**	5	11	**7**	---	---	.39	1.02	-.33	0
19	32	**9**	**7**	9	**7**	---	---	.27	1.02	.35	0

70%-BI = (31, 33) ; *Gower* = .93 ; *Finn* = .94 ; χ_2^2 (16, *N* = 17) = 3.80, *p* = 1.00

The performance standard for reading comprehension for HAVO students was based on the English examination from 2012. The expert panel considered this exam suitable for determining whether the reading comprehension of students corresponds to the descriptors formulated for CEFR level C1. In all, the examination consisted of 45 items, all of which related to reading comprehension. For substantive and/or psychometric reasons, three test items were not included in the standard setting procedure. In all, students could achieve 49 points on the remaining 42 items. The items were divided into five clusters with the following scoring scales: Cluster 1 (0-8), Cluster 2 (0-11), Cluster 3 (0-8), Cluster 4 (0-12) and Cluster 5 (0-10). During the conference, the five clusters were submitted to 19 different subject-area experts for assessment. The most important results from the conference are presented in Table 5.16. As shown in this table, Subject-area Expert 19 set the bar relatively high in comparison to the other subject-area experts. Whereas most of the subject-area experts proposed that a student should be able to answer approximately 35 ÷ 49 = 71% of the test items correctly in order to meet the requirements of CEFR level C1, Subject-area Expert 19 would require students to answer at least 41 ÷ 49 = 84% of the items correctly to demonstrate this level. The impact value of the assessment of Subject-area Expert 19 on the final result was -1. This means that the position of the performance standard would be 1 point lower if we eliminated Subject-area Expert 19 from the analysis. If we disregard one *randomly* selected maximum assessment and one randomly selected minimum assessment, the performance standard for CEFR level C1 would be located at Score 36, with a 70% confidence interval of (34; 37). There was sufficient consensus within the expert panel to set the performance standard at this point. The distribution within the expert panel is realistic, given the reliability of the test at the borderline score. χ_2^2 (16, *N* = 17) = 6.58, *p* = .98, and *Gower's similarity coefficient* and the *Finn coefficient* both exceed .90.

Table 5.16 Results of standard setting in the central final examination for English HAVO (CEFR = C1, $\lceil \bar{C}_{\text{total}} \rceil = 36$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	35	5	**8**	**6**	**9**	**7**	---	.07	1.99	.27	0
2	34	5	6	**6**	10	**7**	---	.63	2.16	.17	0
3	39	7	9	**6**	10	**7**	---	.53	1.47	.58	0
4	36	**6**	**8**	**6**	**9**	**7**	---	.23	1.83	.72	0
5	32	**6**	7	5	8	6	---	.63	2.43	.73	0
6	33	**6**	6	**6**	**9**	6	---	.74	2.30	.55	0
7	34	5	**8**	**6**	**9**	6	---	.02	2.16	.51	0
8	34	**6**	7	**6**	8	**7**	---	.64	2.16	.66	0
9	35	**6**	**8**	**6**	**9**	6	---	.26	1.99	.69	0
10	36	**6**	**8**	**6**	**9**	**7**	---	.23	1.83	.72	0
11	35	**6**	**8**	5	**9**	**7**	---	.50	1.99	.58	0
12	37	**6**	**8**	7	**9**	**7**	---	.23	1.73	.52	0
13	34	5	**8**	**6**	**9**	6	---	.02	2.16	.51	0
14	39	**6**	9	**6**	10	8	---	.16	1.47	.39	0
15	38	7	**8**	**6**	10	**7**	---	.58	1.56	.69	0
16	35	**6**	7	**6**	**9**	**7**	---	.42	1.99	.65	0
17	34	5	**8**	**6**	**9**	6	---	.02	2.16	.51	0
18	37	**6**	**8**	**6**	11	6	---	.44	1.73	.24	0
19	41	7	9	7	10	8	---	.27	1.14	.68	-1

70%-BI = (34, 37) ; *Gower* = .94 ; *Finn* = .95 ; χ^2_2 (16, *N* = 17) = 6.58, *p* = .98

5.2.5 University preparatory education (VWO)

During the conference, the experts assessed two tests that are used in university preparatory education (VWO): the listening comprehension test from 2012 and the first-term examination from 2012. There was sufficient support for rating the listening comprehension test for VWO at CEFR level C1. The test that was administered to VWO students in 2012 consisted of 36 multiple-choice items, which were scored dichotomously. All test items were included in the process of establishing the performance standard. The test was divided into five clusters with the following scoring scales: Cluster 1 (0-6), Cluster 2 (0-6), Cluster 3 (0-8), Cluster 4 (0-9) and Cluster 5 (0-7). Table 5.17 provides a detailed analysis of the data collected during the conference. The suggestions by the subject-area experts are quite close to each other.
The percentage of absolute agreement ranges from .68 (Cluster 1) to .90 (Cluster 3). According to the expert panel, the performance standard for CEFR level C1 should be positioned at Score 23 (23; 24). Students thus answered 23 ÷ 36 = 64% of the test items correctly. As might be expected based on the percentage of absolute agreement, the selection for this performance standard received broad support within the expert panel. Both the absolute and the relative agreement between subject-area experts were high: *Gower's similarity coefficient* is .96 and the *Finn coefficient* is .97. The group of subject-area experts can also be considered a realistic sample of an imaginary population of all possible subject-area experts: χ^2_2 (16, *N* = 17) = 1.52, *p* = 1.00.

Table 5.17 Results of standard setting in the listening comprehension test for English VWO (CEFR = C1, $\lceil \bar{C}_{\text{total}} \rceil = 23$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	22	**4**	**4**	**5**	5	4	---	.19	2.84	-.02	1
2	25	**4**	**4**	6	**6**	**5**	---	.10	2.14	.53	0
3	23	**4**	**4**	**5**	5	**5**	---	.10	2.59	.26	0
4	22	3	3	**5**	**6**	**5**	---	.54	2.84	.61	1
5	21	3	**4**	**5**	5	4	---	.07	3.01	-.06	1
6	24	**4**	**4**	**5**	**6**	**5**	---	.13	2.32	.61	0
7	23	3	**4**	**5**	**6**	**5**	---	.23	2.59	.45	0
8	24	3	**4**	6	**6**	**5**	---	.31	2.32	.34	0
9	22	3	3	**5**	**6**	**5**	---	.54	2.84	.61	1
10	24	**4**	**4**	**5**	**6**	**5**	---	.13	2.32	.61	0
11	26	**4**	**4**	6	**6**	6	---	.24	1.92	.38	0
12	24	**4**	**4**	**5**	**6**	**5**	---	.13	2.32	.61	0
13	23	**4**	**4**	**5**	5	**5**	---	.10	2.59	.26	0
14	24	**4**	**4**	**5**	**6**	**5**	---	.13	2.32	.61	0
15	24	3	**4**	6	**6**	**5**	---	.31	2.32	.34	0
16	23	**4**	**4**	**5**	**6**	4	---	.29	2.59	.26	0
17	24	**4**	**4**	**5**	**6**	**5**	---	.13	2.32	.61	0
18	24	**4**	**4**	**5**	**6**	**5**	---	.13	2.32	.61	0
19	23	**4**	**4**	**5**	5	**5**	---	.10	2.59	.26	0

70%-BI = (23, 24) ; *Gower* = .96 ; *Finn* = .97 ; χ^2_2 (16, *N* = 17) = 1.52, *p* = 1.00

The performance standard for reading comprehension in VWO was determined based on the central examination for English from 2011. The examination contained a total of 43 reading items. The expert panel considered this exam suitable for determining whether the reading comprehension of students corresponds to the descriptors formulated for CEFR level C1.
The examination from 2012 consisted of 43 items, all but seven of which were scored dichotomously. Students could achieve a maximum of 53 points on the examination. All of the items in this examination were included in the standard setting procedure. Prior to the conference, the items were distributed across five clusters with the following scoring scales: Cluster 1 (0-13), Cluster 2 (0-10), Cluster 3 (0-9), Cluster 4 (0-10) and Cluster 5 (0-11). During the conference, the clusters were assessed by 19 subject-area experts. The most important results are presented in Table 5.18. According to these results, the expert panel deemed that students should earn 34 of the 53 points on the VWO examination in order to achieve CEFR level C1. If we take into account the 70% confidence interval around the performance standard, this would correspond to 31 ÷ 53 = 58% and 36 ÷ 53 = 68% of the points. At first glance, the performance standard for CEFR level C1 in VWO appears quite low in comparison to HAVO. On the VWO examination, VWO students have to answer fewer items correctly in order to meet the requirements for CEFR level C1 than do HAVO students on the HAVO examination. This is because the VWO examination is more difficult than the HAVO examination. There was sufficient support amongst the subject-area experts for setting the performance standard at Score 34. *Gower's similarity coefficient* for absolute inter-rater agreement is .91, and the *Finn coefficient* for relative inter-rater agreement has the same value.

Table 5.18 Results of standard setting in the central final examination for English VWO (CEFR = C1, $[\bar{C}_{\text{total}}] = 34$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	32	9	5	**6**	**6**	6	---	.63	2.73	.35	0
2	35	**10**	**6**	7	7	**5**	---	.10	2.35	-.40	0
3	36	**10**	**6**	7	**6**	7	---	.55	2.16	.30	0
4	34	9	7	**6**	**6**	6	---	.53	2.48	.31	0
5	31	9	5	**6**	**6**	**5**	---	.19	2.87	.25	0
6	30	**10**	**6**	5	5	4	---	.49	3.04	.14	0
7	33	**10**	4	**6**	5	8	---	3.08	2.56	.36	0
8	36	11	8	**6**	**6**	**5**	---	.53	2.16	.03	0
9	31	8	5	5	8	**5**	---	1.23	2.87	-.11	0
10	32	**10**	5	**6**	**6**	**5**	---	.10	2.73	.23	0
11	38	12	5	7	7	7	---	.93	2.01	-.07	-1
12	36	11	**6**	**6**	**6**	7	---	.62	2.16	.31	0
13	33	**10**	**6**	**6**	**6**	**5**	---	.04	2.56	.29	0
14	38	**10**	7	7	7	7	---	.18	2.01	.17	-1
15	34	9	**6**	**6**	7	6	---	.41	2.48	.08	0
16	31	9	**6**	5	**6**	**5**	---	.44	2.87	.27	0
17	33	**10**	**6**	**6**	**6**	**5**	---	.04	2.56	.29	0
18	31	**10**	**6**	**6**	4	**5**	---	.94	2.87	.23	0
19	36	**10**	7	5	**6**	8	---	1.64	2.16	.40	0

70%-BI = (31, 36) ; *Gower* = .91 ; *Finn* = .91 ; χ_2^2 (16, *N* = 17) = 7.64, *p* = 1.00

5.3 French

For French, three central examinations and three listening comprehension tests were assessed by a fixed panel of 20 subject-area experts. In the following subsections, we present the results for the combined theoretical and vocational track of pre-vocational secondary education (VMBO; 5.3.1); senior general secondary education (HAVO; 5.3.2); and university preparatory education (VWO; 5.3.3). A summary of the results is presented in Table 5.25c.

5.3.1 Pre-vocational secondary education (VMBO) Combined theoretical and vocational track

During the conference, the experts assessed two tests that are used in the combined theoretical and vocational track of pre-vocational secondary education (VMBO): the listening comprehension test from 2012 and the central examination from 2011. The listening comprehension test consisted of 35 test items, all of which were scored dichotomously. According to the expert panel, this test is suitable for determining whether the listening comprehension of students is located at CEFR level B2. The test was divided into five clusters. These clusters consisted of 7, 6, 5, 12 and 5 test items, respectively. Table 5.19 provides an analysis of the data collected in the second round of assessments. According to these figures, the expert panel determined that a student should earn at least 17 of the 35 points in order to demonstrate CEFR level A2. This performance standard is based on the trimmed mean of the assessments of the subject-area experts. This means that the minimum assessment (possibly selected at random) and the maximum assessment (possibly selected at random) were eliminated. The precision of the performance can be evaluated according to the variation, $\sigma_{C_{\text{total}}}$, in the assessments of

individual subject-area experts. The 70% confidence interval is calculated as follows, and it is equal to $\left(\left\lceil \bar{C}_{\text{total}} \right\rceil - 1.036 \times \sigma_{C_{\text{total}}}; \left\lceil \bar{C}_{\text{total}} \right\rceil + 1.036 \times \sigma_{C_{\text{total}}}\right) = (14; 21)$. This wide confidence interval indicates that the subject-area experts were relatively diverse in their estimates of the number of points a student should be expected to earn in order to demonstrate CEFR level A2. Expressed as the minimum percentage of correct answers that a student would be expected to have on the listening comprehension test from 2012, the assessments varied from 14 ÷ 35 = 40% to 21 ÷ 35 = 60%. The lower level of agreement in establishing the performance standard for this test is confirmed by the lower level of inter-rater agreement in comparison to the other tests. *Gower's similarity coefficient* for absolute inter-rater agreement is .87, and the *Finn coefficient* for relative inter-rater agreement is .82. Even though the values are somewhat lower, the level of inter-rater agreement is still more than adequate. The distribution within the expert panel can be regarded as realistic, given the reliability of the test at the cut score: χ^2_2 (17, *N* = 18) = 2.82, *p* = .23.

Table 5.19 Results of standard setting in the listening comprehension test for French VMBO [Comb./Th.] (CEFR = A2, $\left\lceil \bar{C}_{\text{total}} \right\rceil$ = 17)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	22	5	4	**3**	7	**3**	---	.12	2.74	.02	0
2	18	**3**	**3**	**3**	6	**3**	---	.38	3.57	.38	0
3	22	6	4	2	7	**3**	---	1.00	2.74	-.24	0
4	19	4	4	**3**	**5**	**3**	---	.69	3.36	.36	0
5	20	5	**3**	4	6	2	---	1.26	3.12	.25	0
6	17	**3**	**3**	**3**	**5**	**3**	---	.69	3.78	.45	0
7	15	2	**3**	**3**	4	**3**	---	1.95	4.15	.39	1
8	16	**3**	2	**3**	6	2	---	.76	3.97	.24	0
9	24	5	4	4	8	**3**	---	.14	2.28	.19	0
10	18	**3**	**3**	**3**	6	**3**	---	.38	3.57	.38	0
11	17	**3**	**3**	**3**	**5**	**3**	---	.69	3.78	.45	0
12	13	**3**	2	2	**5**	1	---	.68	4.49	.00	1
13	11	2	2	2	3	2	---	1.12	4.83	.49	1
14	12	**3**	2	2	3	2	---	1.21	4.74	.46	1
15	17	**3**	**3**	**3**	**5**	**3**	---	.69	3.78	.45	0
16	15	**3**	**3**	2	**5**	2	---	.16	4.15	.20	1
17	20	4	**3**	**3**	6	4	---	.74	3.12	.31	0
18	19	**3**	**3**	2	7	4	---	.99	3.36	-.07	0
19	21	5	**3**	**3**	8	2	---	.66	2.87	-.21	0
20	12	2	2	2	4	2	---	.51	4.74	.45	1

70%-BI = (14, 21) ; *Gower* = .87 ; *Finn* = .82 ; χ^2_2 (17, *N* = 18) = 2.82, *p* = .23

The performance standard for reading comprehension was established by presenting the examination for the combined theoretical and vocational track of VMBO from 2011 to the expert panel. This examination consisted of 41 test items, all of which related to reading comprehension. Most of the test items were valued with 1 point for a correct answer. For six test items, students could earn 2 points. The examination thus contained a maximum of 47 points that could be earned. Prior to the conference, the examination was divided into five clusters with the following scoring scales: Cluster 1 (0-10), Cluster 2 (0-11), Cluster 3 (0-8), Cluster 4 (0-9) and Cluster 5 (0-9). The results of the second round of assessments are displayed in Table 5.20. The expert panel considered the examination suitable for determining whether

the reading comprehension of a student corresponds to the descriptors formulated for CEFR level B1. While the expert panel classified the listening comprehension test as CEFR level A2, the experts chose to classify this examination one CEFR level higher. The trimmed mean is located at Score 37, with a 70% confidence interval of (35; 40). This corresponds to 37 ÷ 47 = 79% of the maximum number of points that can be earned. The level of inter-rater agreement can be regarded as sufficient. In Table 5.20, it is interesting to note that, as shown in the last column, none of the subject-area experts estimated the performance that should be delivered in order to demonstrate CEFR level B1 in such a way that the elimination of that assessment would shift the performance standard to another position. Furthermore, *Gower's similarity coefficient* and the *Finn coefficient* were both high, with values of .92.

Table 5.20 Results of standard setting in the central final examination for French VMBO [Comb./ Th.] (CEFR = B1, $\left[\bar{C}_{\text{total}}\right] = 37$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	37	9	**9**	**6**	**7**	6	---	.50	1.50	.29	0
2	35	**8**	7	5	8	**7**	---	.33	1.74	.19	0
3	36	**8**	**9**	**6**	8	5	---	.59	1.60	-.09	0
4	38	9	8	**6**	8	**7**	---	.31	1.31	.36	0
5	41	9	**9**	7	9	**7**	---	.18	.74	-.12	0
6	34	7	8	5	**7**	**7**	---	.09	1.95	.14	0
7	42	10	10	**6**	8	8	---	.33	.70	.33	0
8	39	**8**	**9**	**6**	8	8	---	.12	1.15	.10	0
9	39	10	**9**	**6**	**7**	**7**	---	.67	1.15	.42	0
10	34	**8**	7	5	**7**	**7**	---	.35	1.95	.40	0
11	36	9	7	**6**	**7**	**7**	---	.80	1.60	.42	0
12	39	**8**	**9**	**6**	8	8	---	.12	1.15	.10	0
13	33	**8**	8	5	6	6	---	.44	2.08	.37	0
14	36	**8**	8	5	8	**7**	---	.12	1.60	.18	0
15	34	**8**	8	5	**7**	6	---	.18	1.95	.35	0
16	38	**8**	**9**	**6**	8	**7**	---	.01	1.31	.32	0
17	36	7	**9**	**6**	**7**	**7**	---	.23	1.60	-.02	0
18	43	9	10	**6**	9	9	---	.37	.57	-.20	0
19	38	**8**	10	**6**	**7**	**7**	---	.25	1.31	.10	0
20	35	**8**	**9**	5	**7**	6	---	.23	1.74	.19	0

70%-BI = (35, 40) ; *Gower* = .92 ; *Finn* = .92 ; χ_2^2 (17, *N* = 18) = 16.20, *p* = .51

5.3.2 Senior general secondary education (HAVO)

For senior general secondary education (HAVO), the listening comprehension test from 2012 was selected. The test consisted of 35 dichotomously scored test items. According to the expert panel, this test is suitable for determining whether the listening comprehension of students is located at CEFR level B1. For the standard setting procedure, the test was divided into five clusters with the following scoring scales: Cluster 1 (0-8), Cluster 2 (04), Cluster 3 (08), Cluster 4 (011) and Cluster 5 (0-4). During the conference, the clusters were assessed by 20 subject-area experts. The most important results are presented in Table 5.21. Seven of the 20 subject-area experts found that the performance standard for CEFR level B1 should be set at Score 22. The other subject-area experts set the border at scores ranging from 17 to 25. As compared to the other experts, Subject-area Experts 7 and 13 placed the border relatively high and relatively low,

respectively. None of the patterns of cut scores differ significantly from what might be expected based on the behaviour of students in Dutch education. The impact scores indicate that the omission of individual subject-area experts from the analysis does not affect the final result. The trimmed mean is 21. The performance standard for CEFR level B1 is thus located at Score 21, with a 70% confidence interval of (19; 23). There was sufficient agreement amongst the subject-area experts to position the performance standard at this point. In addition to being a realistic sample of the population of all possible subject-area experts χ^2_2 (17, *N* = 18) = 6.16, *p* = .99, the expert panel had a considerably high level of agreement, as evidenced by *Gower's similarity coefficient* (.92) and the *Finn coefficient* (.89).

Table 5.21 Results of standard setting in the listening comprehension test for French HAVO (CEFR = B1, $\lceil \bar{C}_{\text{total}} \rceil$ = 21)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	23	**5**	**3**	6	**6**	**3**	---	.41	2.36	.52	0
2	22	**5**	**3**	**5**	**6**	**3**	---	.42	2.57	.65	0
3	19	4	**3**	4	5	**3**	---	1.06	3.28	.66	0
4	22	**5**	**3**	**5**	7	2	---	.44	2.57	.03	0
5	22	**5**	**3**	**5**	7	2	---	.44	2.57	.03	0
6	20	**5**	**3**	4	5	**3**	---	.85	3.01	.63	0
7	25	**5**	4	**5**	7	4	---	1.71	1.93	.66	0
8	20	**5**	**3**	4	5	**3**	---	.85	3.01	.63	0
9	24	6	**3**	**5**	7	**3**	---	.25	2.10	.60	0
10	22	**5**	**3**	**5**	**6**	**3**	---	.42	2.57	.65	0
11	22	**5**	2	**5**	7	**3**	---	.44	2.57	.20	0
12	20	**5**	**3**	4	5	**3**	---	.85	3.01	.63	0
13	17	**5**	**3**	3	4	2	---	1.44	3.65	.37	0
14	21	**5**	**3**	**5**	**6**	2	---	.33	2.83	.09	0
15	19	4	2	**5**	**6**	2	---	.28	3.28	-.30	0
16	19	**5**	2	4	5	**3**	---	.44	3.28	.40	0
17	22	**5**	**3**	**5**	**6**	**3**	---	.42	2.57	.65	0
18	23	6	**3**	6	5	**3**	---	.56	2.36	.38	0
19	18	4	**3**	3	5	**3**	---	1.61	3.43	.64	0
20	22	**5**	**3**	**5**	**6**	**3**	---	.42	2.57	.65	0

70%-BI = (19, 23) ; *Gower* = .92 ; *Finn* = .89 ; χ^2_2 (17, *N* = 18) = 6.16, *p* = .99

In addition to a listening comprehension test, the final examination from 2011 was submitted for assessment during the conference. The expert panel deemed the examination suitable for determining whether students are at the CEFR level B1 with respect to reading comprehension. This examination contained 40 reading items. On eight of the items, students could earn two or more points. The other items were scored dichotomously. The scoring scale of the examinations used during the conference ranged from 0 to 49. The test was divided into five clusters with the following scoring scales: Cluster 1 (0-14), Cluster 2 (0-9), Cluster 3 (0-10), Cluster 4 (0-7) and Cluster 5 (0-9). Table 5.22 provides an analysis of the data collected during the second assessment round. The performance standard for CEFR level B1 is set at Score 29. The measures of agreement between subject-area experts differed according to cluster. The assessments for each cluster are presented in Columns C1–C6. As shown in Column C4, the majority of the subject-area experts (i.e. 16 of 20) deemed that a borderline candidate should earn at least 4 of the 7 points on this cluster in order to demonstrate CEFR level B1. The variation in assessments

was greater for the other clusters. For example, in the second cluster (Column C2), the mode is equal to 5 points. This score was selected by only 8 of the 20 subject-area experts. Nevertheless, the 70% confidence interval around the performance standard was quite small, with a lower limit of 26 and an upper limit of 32. The two measurements of inter-rater agreement also indicated that there was sufficient support for positioning the performance standard at Score 29: *Gower's similarity coefficient* for absolute inter-rater agreement is .92, and the *Finn coefficient* for relative inter-rater agreement is .91. The distribution across the group of subject-area experts is realistic, given the reliability of the test at the cut score. χ^2_2 (17, N = 18) = 14.84, p = .61.

Table 5.22 Results of standard setting in the central final examination for French HAVO (CEFR = B1, $\left[\bar{C}_{\text{total}}\right]$ = 29)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ^2_1	α=.10	RSI	Impact
1	30	10	**5**	**6**	**4**	**5**	---	.75	2.91	.53	0
2	29	9	**5**	**6**	**4**	**5**	---	1.06	3.09	.55	0
3	30	10	6	**6**	**4**	4	---	.56	2.91	.13	0
4	28	10	4	**6**	**4**	4	---	.30	3.21	.50	0
5	35	**11**	6	8	5	**5**	---	.09	2.12	.09	-1
6	30	9	4	7	5	**5**	---	.98	2.91	.41	0
7	31	**11**	**5**	**6**	**4**	**5**	---	.61	2.77	.40	0
8	25	9	3	5	**4**	4	---	1.20	3.63	.47	0
9	32	**11**	**5**	7	**4**	**5**	---	.33	2.66	.35	-1
10	27	8	4	**6**	**4**	**5**	---	1.72	3.37	.59	0
11	30	**11**	4	7	**4**	4	---	.22	2.91	.04	0
12	24	9	2	5	**4**	4	---	2.07	3.81	.34	0
13	21	5	4	5	**4**	3	---	2.01	4.24	.37	0
14	32	**11**	**5**	7	5	4	---	.05	2.66	-.05	-1
15	22	7	3	5	**4**	3	---	.86	4.14	.43	0
16	25	8	4	5	**4**	4	---	1.08	3.63	.58	0
17	32	**11**	**5**	7	**4**	**5**	---	.33	2.66	.35	-1
18	32	10	**5**	7	5	**5**	---	.36	2.66	.51	-1
19	31	10	**5**	7	**4**	**5**	---	.47	2.77	.48	0
20	27	8	4	**6**	**4**	**5**	---	1.72	3.37	.59	0

70%-BI = (26, 32) ; *Gower* = .92 ; *Finn* = .91 ; χ^2_2 (17, N = 18) = 14.84, p = .61

5.3.3 University preparatory education (VWO)

During the conference, the experts assessed two tests that are used in university preparatory education (VWO): the listening comprehension test from 2012 and the central final examination from 2009. The listening comprehension test consisted of 36 test items. The maximum number of points that could be earned on the test was 36. The expert panel established a performance standard for CEFR level B2. The test was divided into the following clusters with the following scoring scales: Cluster 1 (0-4), Cluster 2 (0-9), Cluster 3 (0-6), Cluster 4 (0-8) and Cluster 5 (0-9). According to the expert panel, a student should earn at least 18 of the 36 points in order to demonstrate CEFR level B2. As shown in Table 5.23, the distribution around this point estimate is relatively large. Whereas Subject-area Expert 15 would expect a minimal bordeline candidate to earn 12 points, Subject-area Experts 1 and 20 proposed that CEFR level B2 could be demonstrated only with scores starting at 23. Because we work with a trimmed mean in the

analysis, Subject-area Expert 15 (the lowest assessment) and one random selection between Subject-area Experts 1 and 20 (the highest assessments) were eliminated. After they were eliminated, the 70% confidence interval around the performance standard was (16;21).
The measurements for inter-rater agreement could be regarded as quite high: .90 for *Gower's similarity coefficient* and .89 for the *Finn coefficient*. This means that there was sufficient support amongst the subject-area experts to set the performance standard for CEFR level B2 at Score 18 on the listening comprehension test from 2012. This is further confirmed by the χ_2^2 test: χ_2^2 (17, N = 18) = 1.20, p = .90.

Table 5.23 Results of standard setting in the listening comprehension test for French VWO (CEFR = B2, $\lceil \bar{C}_{\text{total}} \rceil$ = 18)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	23	1	6	4	5	7	---	1.53	2.55	.54	0
2	17	**2**	4	**3**	**4**	4	---	.30	3.78	.50	0
3	20	**2**	**5**	4	**4**	**5**	---	.09	3.21	.31	0
4	19	**2**	**5**	**3**	**4**	**5**	---	.29	3.37	.63	0
5	21	**2**	**5**	**3**	5	6	---	.79	3.02	.68	0
6	15	**2**	4	**3**	3	3	---	.34	4.18	.02	0
7	19	**2**	**5**	**3**	**4**	**5**	---	.29	3.37	.63	0
8	20	**2**	**5**	**3**	5	**5**	---	.69	3.21	.65	0
9	17	1	4	**3**	3	6	---	1.31	3.78	.48	0
10	18	**2**	4	**3**	**4**	**5**	---	.31	3.60	.61	0
11	17	1	**5**	**3**	3	**5**	---	.99	3.78	.50	0
12	17	1	4	**3**	**4**	**5**	---	.93	3.78	.66	0
13	13	1	3	2	3	4	---	.81	4.55	.69	0
14	20	**2**	**5**	4	**4**	**5**	---	.09	3.21	.31	0
15	12	1	3	2	3	3	---	.66	4.74	.68	0
16	19	**2**	**5**	**3**	**4**	**5**	---	.29	3.37	.63	0
17	18	**2**	4	**3**	**4**	**5**	---	.31	3.60	.61	0
18	17	1	4	4	3	**5**	---	.72	3.78	.15	0
19	18	**2**	4	**3**	3	6	---	.65	3.60	.30	0
20	23	3	**5**	4	5	6	---	.30	2.55	.25	0

70%-BI = (16, 21) ; *Gower* = .90 ; *Finn* = .89 ; χ_2^2 (17, N = 18) = 1.20, p = .90

The performance standard for reading comprehension in VWO was determined based on the central final examination from 2009. The test contained 45 items, most of which were scored dichotomously. The maximum number of points that could be achieved in the examination was 49. The expert panel deemed the tests suitable for determining CEFR level B2. The classification that was used during the conference for this examination was as follows: Cluster 1 (0-8) Cluster 2 (0-9), Cluster 3 (0-8), Cluster 4 (0-7), Cluster 5 (0-7) and Cluster 6 (0-10). The clusters were assessed by 20 subject-area experts. Table 5.24 provides an analysis of the data collected during the second round of assessments. The expert panel positioned the performance standard at Score 31 (28; 33). The 70% confidence interval corresponds to 28 ÷ 49 = 57% (lower limit) and 33 ÷ 49 = 67% (upper limit) of the total number of points that can be earned in the examination. *Gower's similarity coefficient* for absolute inter-rater agreement is .92, and the *Finn coefficient* for relative inter-rater agreement is .93. These values can be regarded as high. The subject-area experts were also relatively unanimous in their assessment of each cluster, as shown in Columns C1–C6 in Table 5.24. This confirms that there was sufficient support within the expert

panel to set the performance standard for CEFR level B1 at Score 31 for the final examination in French from 2009.

Table 5.24 Results of standard setting in the central final examination for French VWO (CEFR = B2, $\lceil \bar{C}_{total} \rceil = 31$)

Expert	C_{total}	C1	C2	C3	C4	C5	C6	χ_1^2	α=.10	RSI	Impact
1	36	**6**	6	6	6	**5**	7	.21	2.03	.25	-1
2	29	**6**	**5**	4	4	4	**6**	.87	3.29	.62	0
3	31	**6**	6	**5**	4	4	**6**	.68	2.95	.46	0
4	29	5	**5**	4	**5**	**5**	5	.26	3.29	.18	0
5	35	7	7	**5**	**5**	**5**	**6**	.67	2.23	.44	-1
6	30	5	**5**	**5**	4	**5**	**6**	.34	3.15	.43	0
7	37	7	7	6	**5**	**5**	7	.50	1.88	.50	-1
8	32	**6**	**5**	**5**	**5**	**5**	**6**	.33	2.74	.63	0
9	26	5	4	4	4	4	5	.39	3.87	.65	0
10	30	**6**	**5**	**5**	4	4	**6**	.71	3.15	.59	0
11	30	**6**	**5**	**5**	4	4	**6**	.71	3.15	.59	0
12	31	**6**	**5**	4	**5**	**5**	**6**	.62	2.95	.55	0
13	28	**6**	**5**	3	4	4	**6**	1.40	3.52	.55	0
14	30	5	**5**	**5**	**5**	**5**	5	.12	3.15	.10	0
15	32	**6**	**5**	**5**	**5**	**5**	**6**	.33	2.74	.63	0
16	29	5	4	**5**	**5**	**5**	5	.42	3.29	.20	0
17	29	5	**5**	**5**	4	**5**	5	.30	3.29	.34	0
18	32	5	6	**5**	**5**	**5**	**6**	.06	2.74	.09	0
19	28	**6**	3	**5**	**5**	4	5	1.54	3.52	.40	0
20	29	**6**	**5**	**5**	4	4	5	.69	3.29	.52	0

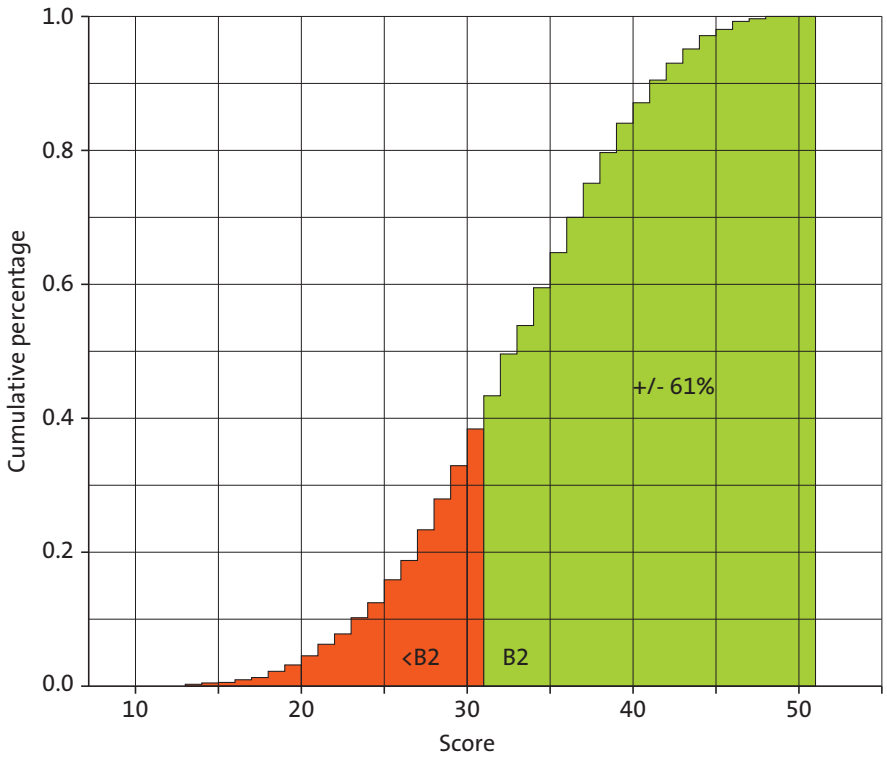
70%-BI = (28, 33) ; *Gower* = .92 ; *Finn* = .93 ; χ_2^2 (17, *N* = 18) = 8.69, *p* = .95

5.4 Practical implications

Based on the results presented in sections 5.1, 5.2 and 5.3, it is possible to identify at which CEFR levels we can expect to find the reading and listening comprehension of students. Test data were available for nearly all of the tests. As shown in Figure 5.1, based on the final examination in French, it is possible to determine how many students in university preparatory education (VWO) achieved a specific CEFR level. On the *x*-axis, we see that the scoring scale runs from 0 to 51. The *y*-axis shows the observed cumulative percentage of students with a particular score. We can infer that the cumulative percentage at Score 35 is equal to .60. This means that 60% of the students scored under 35 and that 40% scored equal to or above 35. Similarly, we can determine the percentage of students scoring below or above the performance standard. As reported in Section 5.3.3, the expert panel set the performance standard for CEFR level B2 at Score 31 on the VWO final examination in French. This score is highlighted in Figure 5.1. Approximately 61% of the students fall into the area shaded in green, thus performing at the B2 level or higher. The other students fall into the area shaded in red. The reading comprehension of these students in French does not correspond sufficiently to the descriptors for CEFR level B2. For all other final examinations and listening comprehension tests, the percentage of students with a specific CEFR level can be derived in exactly the same way from the empirical cumulative

distribution function, as has just been illustrated for the VWO final examinations in French from 2009.

Figure 5.1 Empirical cumulative distribution function for the VWO central final examination in French



Tables 5.25a, 5.25b and 5.25c provide a summary of the results of the conference, and they contain an estimate of the percentage of students in each type of education achieving a specific CEFR level. The percentage is expressed as a 95% confidence interval. The first figure indicates the lower limit of the interval, and the second figure shows the upper limit. We deliberately chose not to include any absolute values. The size of the databases varies quite strongly. For example, we have access to results from 19 875 students for the final examination in German for the combined theoretical and vocational track of secondary vocational preparatory education (VMBO). For the final examination in German for the basic vocational track of VMBO, results are available for only 151 students. The number of students affects the precision with which we can estimate the percentage of students who have achieved a particular CEFR level. The higher the number of available student results, the higher the precision. The margin of error was determined in a *bootstrap* procedure with 1000 replications. The nested structure of the data was ignored in the calculations. This means that the margin of error is probably underestimated and the confidence intervals are actually somewhat larger (see Cochran, 1977). As shown in Table 5.25, 39%–55% of the students taking German in the basic vocational track of VMBO could be expected to read at CEFR level B1. They should be able to provide correct answers to 19 items (59%) on the final examination from 2012.

Table 5.25a Summary results for German (underlined = predicted)

Type	CEFR	Skill	Year	*N*	Max.	$[\bar{C}_{total}]$	*Gower*	*Finn*	Achieved	Perc. correct
VMBO	B1	Reading	2012	151	32	19	.94	.95	(39; 55)	(56; 63)
[Bas.Voc.]		Listening	2012	0	36	21	.95	.96	NA*	(56; 64)
VMBO	B1	Reading	2012	472	41	25	.92	.92	(67; 76)	(56; 66)
[Adv.Voc.]		Listening	2011	62	36	18	----	----	(42; 68)	
VMBO	B1	Reading	2012	19875	44	21	.90	.89	(78; 80)	(43; 55)
[Comb./Th.]		Listening	2012	1774	37	20	----	----	(83; 88)	
	B2	Reading	----	----	----	----	----	----	----	
		Listening	2012	1774	37	26	.92	.90	(35; 40)	(68; 73)
Senior general	B2	Reading	2012	14948	50	29	.95	.97	(70; 72)	(54; 62)
secondary		Listening	2013	2885	40	23	----	----	(84; 88)	
education	C1	Reading	----	----	----	----	----	----	----	
(HAVO)		Listening	2013	2885	40	29	.95	.97	(38; 42)	(70; 75)
Pre-university	C1	Reading	2010	14411	51	29	.91	.91	(51; 54)	(51; 63)
education		Listening	2012	3370	38	21	----	----	(80; 84)	
(VWO)	C2	Reading	----	----	----	----	----	----	----	
		Listening	2012	3370	38	28	.92	.89	(31;35)	(68; 76)

NA = No test data are available for this type of education.

Table 5.25b Summary results for English (underlined = predicted)

Type	CEFR	Skill	Year	*N*	Max.	$[\bar{C}_{total}]$	*Gower*	*Finn*	Achieved	Perc. correct
VMBO	B1	Reading	2012	3912	32	24	.95	.96	(54; 58)	(72; 78)
[Bas.Voc.]		Listening	2012	641	34	22	.92	.92	(50; 59)	(59; 68)
VMBO	B1	Reading	2012	3777	39	24	.91	.90	(60; 64)	(56; 67)
[Adv.Voc.]		Listening	2012	984	35	19	----	----	(74; 80)	
VMBO	B2	Reading	2012	32000	33	25	.96	.97	(32; 34)	(73; 79)
[Comb./Th.]		Listening	2012	3872	36	24	.94	.93	(62; 66)	(64; 72)
Senior general secondary	C1	Reading	2012	32000	49	36	.94	.95	(33; 35)	(69; 76)
education (HAVO)		Listening	2012	8875	39	32	.93	.94	(30; 32)	(79; 85)
Pre-university education	C1	Reading	2011	29141	53	34	.91	.91	(52; 54)	(58; 68)
(VWO)		Listening	2012	6907	36	23	.96	.97	(73; 76)	(64; 67)

Table 5.25c Summary results for French (underlined = predicted)

Type	CEFR	Skill	Year	*N*	Max.	$[\bar{C}_{\text{total}}]$	*Gower*	*Finn*	Achieved	Perc. correct
VMBO	A2	Reading	----	----	----	----	----	----	----	
[Comb./Th.]		Listening	2012	636	35	17	.87	.82	(91; 96)	(40; 60)
	B1	Reading	2011	5518	47	37	.92	.92	(8; 11)	(74; 85)
		Listening	2012	636	35	28	----	----	(17; 25)	
Senior general secondary education (HAVO)	B1	Reading	2011	9342	49	29	.92	.91	(56; 59)	(53; 65)
		Listening	2012	2012	35	21	.92	.89	(69; 74)	(54; 66)
Pre-university education (VWO)	B2	Reading	2009	2257	49	31	.92	.93	(58; 63)	(57; 67)
		Listening	2012	3135	36	18	.90	.89	(89; 92)	(44; 58)

The percentage of students with a specific CEFR level can also be displayed visually. This has been done in Figures 5.2a, 5.2b and 5.2c for German, English and French, respectively. As with the previous tabular presentations of the results, comparisons have been made between reading comprehension and listening comprehension for each type of education. For English, this comparison is easy to make, because the listening comprehension test in each type of education was classified at the same CEFR level as the central examination. For German and French, the comparison was less easily made in some cases. For example, in the combined theoretical and vocational track of VMBO, the expert panel deemed the listening comprehension test in German suitable for establishing CEFR level B2, while they considered the final examination more consistent with the descriptors for CEFR level B1. In such cases, the percentage of students with a specific CEFR level was predicted according to the underlying item response theory model, whenever possible. For German in the combined theoretical and vocational track in VMBO, this means that we have attempted to estimate the percentage of students achieving CEFR levels B1 and B2 for both reading comprehension and listening comprehension. The performance standards for reading comprehension B1 and listening comprehension B2 were established during the conference and the percentage of students with these CEFR levels was derived from the empirical cumulative distribution functions. The item response theory model was used to convert the performance standard for listening comprehension B1 from the test that was administered in the basic vocational track of VMBO to the test that was administered in the combined theoretical and vocational track of VMBO. Students in the combined theoretical and vocational track of VMBO could be expected to satisfy the requirements for CEFR level B1 if they are able to answer 20 of the 37 items on the listening comprehension test from 2012 correctly. According to the empirical cumulative distribution function, between 83% and 88% of the students taking this test achieved or exceeded this score. The performance standard for reading level B2 could not be determined for the combined theoretical and vocational track of VMBO, as the central final examinations are not based on a single underlying measurement scale. The same process was followed for the other types of education. The predictions in Tables 5.25a, 5.25b and 5.25c are underlined, and they are highlighted in Figures 5.2a, 5.2b, and 5.2c. In general, we see that a slightly higher number of students achieved the prevailing CEFR level in listening comprehension than was the case for reading comprehension.

Figure 5.2a Percentage of students with a specific CEFR level in German, by type of education

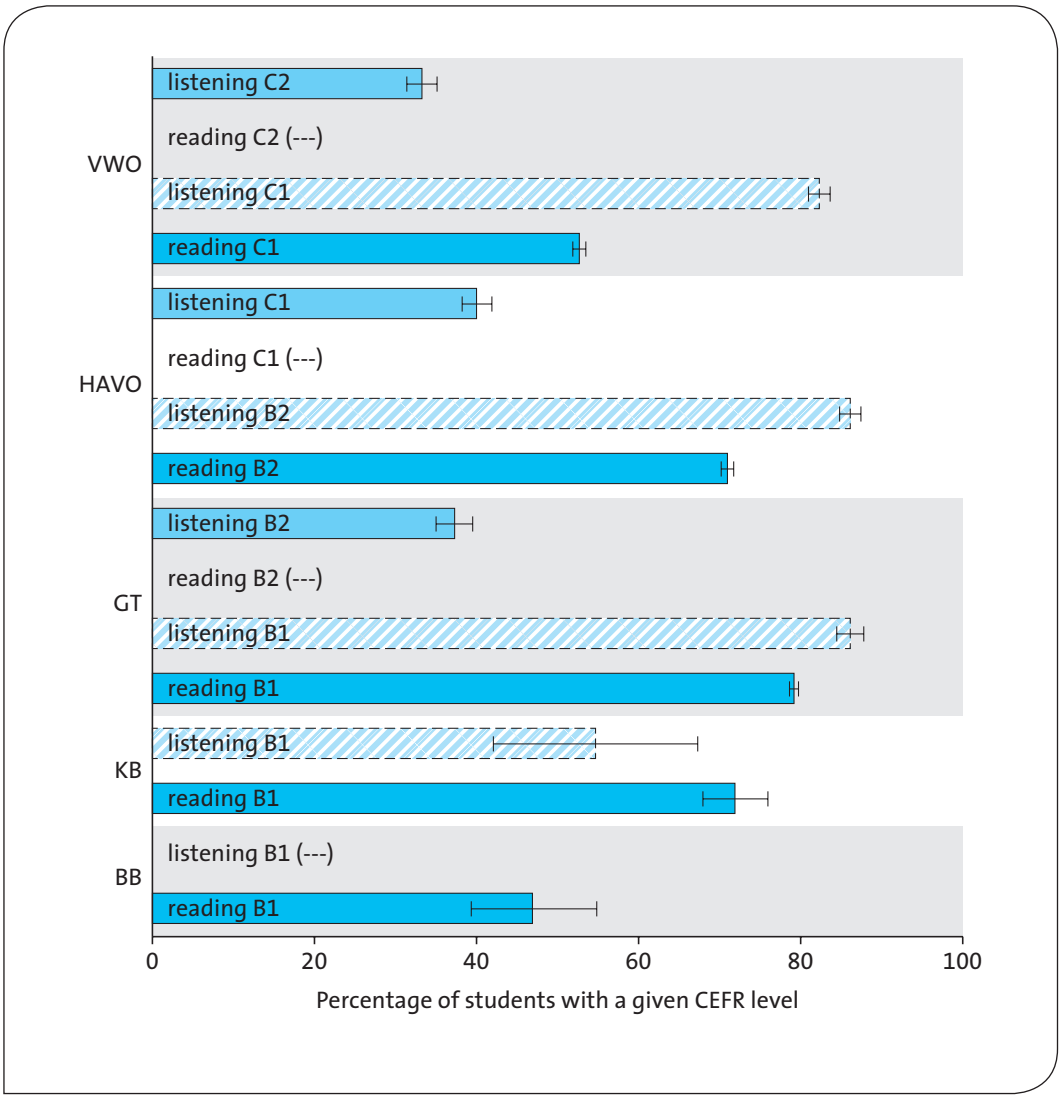


Figure 5.2b Percentage of students with a specific CEFR level in English, by type of education

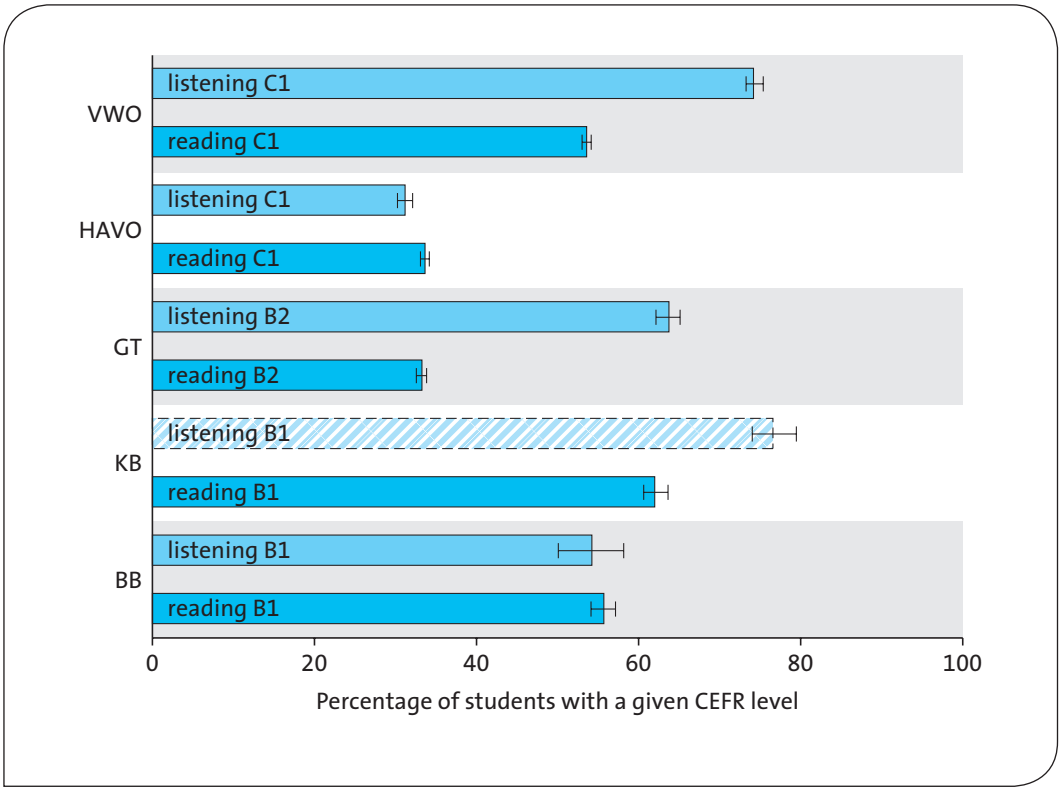
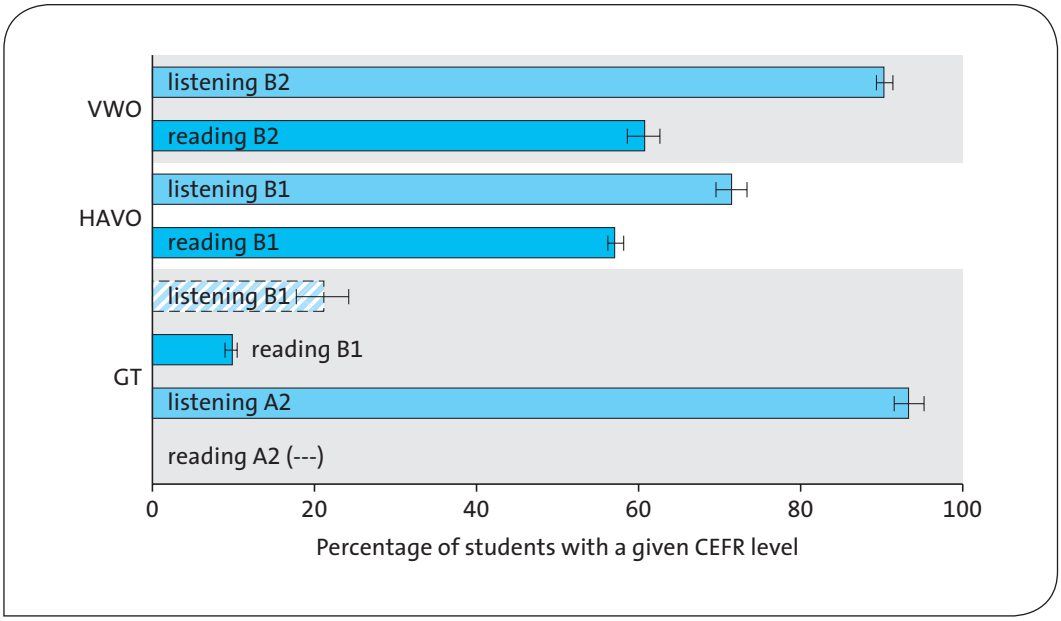


Figure 5.2c Percentage of students with a specific CEFR level in French, by type of education



6 Summary and conclusions

6 Summary and conclusions

The CEFR describes six levels of language skills that can be used for all languages. These levels are gaining acceptance as standards for the assessment of individual language skills, and are contributing to comparability in the learning, teaching, and evaluation of languages in Europe. The Netherlands Ministry of Education, Culture and Science (OCW) commissioned Cito to conduct an international standard setting study in order to determine the performance that students taking final examinations should deliver on various Dutch examinations and tests in order to demonstrate particular CEFR levels. To achieve this objective, an international conference was organised in September 2013. Using a standard-setting procedure, performance standards for reading and listening comprehension were established for German, English and French. The performance standards have been determined for various types of education, namely for the basic, conceptual, mixed, and theoretical learning routes of pre-vocational secondary education (VMBO); senior general secondary education (HAVO); and university preparatory education (VWO). The performance standards for reading comprehension were determined using the central examinations. Those for listening comprehension were developed using the Cito listening comprehension tests. In all, six performance standards were established for French, and nine for both German and English.

All of the tests for which performance standards were established in this study are standardised tests that conform to a fixed set of criteria concerning base materials, question formats, testing times, test length, test duration, and authorised tools. The descriptors and domains mentioned in the CEFR were not used as guidelines in the construction of the tests used in this study. The exit qualifications for modern foreign languages do not contain any explicit reference to the CEFR. We can nevertheless establish that the tests sufficiently correspond to the CEFR philosophy, as they were all based on a communicative approach. This means that all of these tests can be used to determine whether students have understood the intended messages of the authors and speakers in question. Test data were available for nearly all of the tests used in the study. This made it possible to determine the percentage of students in a given type of education that have achieved the CEFR level for that type of education. The test data also allowed us to determine the reliability of measurements made by examinations and the listening comprehension test. Reliability ranged from .64 to .86. Further investigation is needed in order to determine whether the examinations and listening comprehension tests are sufficiently reliable in assessing the CEFR levels of individual students with the desired precision. The reliability of an examination or listening comprehension test probably has little impact on the actual performance standard, given that the performance standard was not determined according to the assessment of one individual, but according to the assessments of a group of subject-area experts.

In order to establish the performance standards, an international expert panel was established for each language, composed of 16 to 20 subject-area experts in the field of foreign and second language education, as well as the CEFR. Representatives of universities, ministries, testing organisations and schools were invited to serve on an expert panel. The expert panels were asked to determine the minimum scores on a number of tests that student should achieve in order to demonstrate a particular CEFR level. One performance standard was established for each test. A standard setting procedure developed by Cito – the *Data-Driven Direct Consensus (*3DC) method – was used in this process. In contrast to the usual application of the *Direct Consensus* method, the 3DC method also uses empirical data. During the procedure, the CEFR level to be determined is established for each test by subject-area experts, first individually, and then jointly. Subsequently, individual determinations are made for a number of clusters of test items, regarding the score that a borderline candidate should achieve on a given cluster in order

to demonstrate the selected CEFR level. These individual assessments are then presented and discussed in a plenary session. Based on this exchange of arguments, subject-area experts can revise their initial assessments in a second round. The performance standard is established according to the second round of assessments, and is equal to the trimmed mean of the sum of the individual assessments for each cluster. In other words, the choice was made to discard the lowest and highest individual summed scores from the analysis, in order to prevent the performance standard from being overly influenced by extreme assessments.

For French, performance standards were established for CEFR levels A2–B2; for English, CEFR levels B1–C1; and for German, CEFR levels B1 and C2. In comparison with previous standard setting studies conducted in 2006 and 2007 (see Noijons & Kuijper, 2006; Cito, 2007), the ability that the subject-area experts thought a student should possess in order to demonstrate a particular CEFR level was consistently lower. This means that, according to the 2013 standard setting study, the performance of students in the Netherlands can be evaluated with a higher CEFR level than was indicated in the standard setting studies of 2006 and 2007. This applies to all languages and all types of education. The most likely explanation lies in the different composition of the expert panels. In 2006 and 2007, only Dutch subject-area experts participated in the standard setting studies. In 2013, international expert panels were involved. As expected, the performance standards increase according to the type of education for which the test was prepared. Because the listening comprehension tests for each language are calibrated through an item response theory model to a single underlying measurement scale, it is possible to translate the performance standard from one test to another. These comparisons reveal that the subject-area experts were consistent in establishing the performance standards for the listening comprehension tests. The calculated measures of inter-rater agreement support this conclusion. In both the examinations and the listening comprehension tests, the subject-area experts were in strong agreement with each other. This means that the subject-area experts unanimously estimated the performance needed in order to achieve a certain CEFR level.

The standard setting study that was conducted has several limitations. First, the recruitment criteria led to some discrepancies in the composition of the expert panels. Second, the CEFR is not interpreted in the same way for every language. The expert panels exhibited differing opinions concerning the highest CEFR level. According to the German expert panel, 30% of students in VWO scored at CEFR level C2 on the listening comprehension test, while the English expert panel did not consider the listening comprehension test for students in VWO suitable for measuring CEFR level C2. In the opinion of the English panel, by definition, tests for students in secondary education do not draw adequately on the knowledge and skills that are required for level C2. It was therefore agreed to establish a performance standard for CEFR level C1 for English in VWO. This outcome does not necessarily discredit the standard setting study, largely because the most important aspect is that there is sufficient support for the outcomes in the areas concerned. The measures of inter-rater agreement demonstrate the existence of such support. Finally, neither the central final examinations nor the listening comprehension tests were specifically constructed according to CEFR guidelines. For example, the tests do not focus on one CEFR level, but on several. In addition, they do not address all domains and descriptors. In some cases, this made it more difficult for the expert panels to establish a performance standard. In the extended plenary discussions, however, it was possible in all cases to arrive at a broadly-supported performance standard.

Several recommendations can be made according to these assessments. First, it would be advisable to examine the extent to which the different languages interpret the CEFR in the same way and whether the descriptors for CEFR level C2 are sufficiently distinctive in relation to the other CEFR levels. This would require the organisation of a standard setting procedure in

which bilingual subject-area experts would establish performance standards for two European languages (e.g. English and German). This would make it possible to reveal language-specific interpretations of the CEFR. Second, it would be advisable to conduct a new standard setting study based on tests that correspond to the CEFR. This would simplify the work of the expert panels and increase the validity of the CEFR measurement. If the tests do not correspond to the CEFR, there is a risk of estimating the CEFR levels of students either too high or too low, because the tests fail to address some domains and descriptors. Finally, the reliability of the central examinations and listening comprehension tests is a concern. The allocation of qualifications is based on the scores that students have achieved on multiple examinations and tests. In the allocation of a CEFR level, the score on one specific test is decisive. This has implications for the way in which the reliability must be assessed. In the first case, the assessments could be based on the full set of tests that is administered. In the second case, the reliability of the individual test must be considered. It is not certain whether the test would be sufficient in all cases to make an accurate determination of the CEFR levels of individual students. Further research is needed in this regard. Nevertheless, the standard setting study has already revealed a highly consistent picture, in which the performance of students in the Netherlands is evaluated with a higher CEFR level than was previously the case.

7 Literature

7 Literature

Angoff, W.H. (1971). Scale, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600). Washington, DC: American Council on Education.

Angoff, W.H. (1988). Validity: An evolving concept. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Erlbaum.

Berk, R.A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*, 137-172.

Berk, R. (1996). Standard setting: the next generation. *Applied Measurement in Education, 9*, 215-235.

Cito (2007). *De koppeling van de Cito kijk- en luistertoetsen moderne vreemde talen aan het Europees Referentiekader.* Arnhem: Cito.

Cito (2009 - 2012). *Examenverslagen 2009, 2010, 2011 en 2012.* www.cito.nl.

Council of Europe (2001). *Common European Framework of References for Languages. Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

CvE. (2010). *Syllabus centraal examen 2012: Moderne vreemde talen havo.* Utrecht: CvE.

CvE. (2010). *Syllabus centraal examen 2012: Moderne vreemde talen vmbo*. Utrecht: CvE.

CvE. (2010). *Syllabus centraal examen 2012: Moderne vreemde talen vwo*. Utrecht: CvE.

Driessen, M., Kleef, A. van & Kleunen, E. van (2012). *Handreiking referentiekader moderne vreemde talen in het mbo.* Ede / 's-Hertogenbosch: Steunpunt taal en rekenen mbo / Cinop.

Embretson, S. E. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum Publishers.

Evers, A., Lucassen, W., Meijer, R. & Sijtsma, K. (2010). *COTAN Beoordelingssysteem voor de kwaliteit van tests*. Amsterdam, NIP/COTAN.

Finn, R.H. (1970). A note on estimating the reliability of categorical data. *Educational and Psychological Measurement, 30*, 71-76.

Goodwin, L.D. (1999). Relations between observed item difficulty levels and Angoff minimum passing levels for a group of borderline candidates. *Applied Measurement in Education, 12*, 13-28.

Gower, J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics 27*, 623-637.

Gwet, K.L. (2010). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg, MD: Advanced Analytics LLC.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1992). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R.K., Jaeger, R.M., Plake, B.S. & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*, 355-366.

Jaeger, R.M. (1989). Certification of student competence. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485-514). New York: American Council on Education and Macmillan.

Kaftandjieva, F. (2004). Standard setting. In *Reference Supplement to the preliminary pilot version of the Manual for Relating Language examinations to the CEF, Section B* (DGIV/EDU/LANG(2004) 13). Strasbourg, France: Language Policy Division.

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Le Breton, J. M., & Sentor, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11*, 815-852.

Nederlandse Taalunie (2008). *Gemeenschappelijk Europees Referentiekader voor Moderne Vreemde Talen. Leren, Onderwijzen, Beoordelen*. Den Haag.

Mitzel, H.C., Lewis, D.M., Patz, R.J., & Green, D.R. (2001). The bookmark procedure: Psychological perspectives. In G.J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Erlbaum.

Noijons, J. & Kuijper, H. (2006). *De koppeling van de centrale examens leesvaardigheid moderne vreemde talen aan het Europees Referentiekader.* Arnhem: Cito.

R Development Core Team (2011). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420-428.

Sim, J & Wright, C.C (2005). The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirement. *Physical Therapy, 85*, 257-268.

Sireci, S.G., Hambleton, R.K., & Pitoniak, M.J. (2004). Setting passing scores on licensure exams using direct consensus. *CLEAR Exam Review, 15*, 21-25.

Tinsley, H.E.A., & Weiss, D.J. (2000). Interrater reliability and agreement. In H.E.A. Tinsley & S.D. Brown (Eds.), *Handbook of applied multivariate statistics and mathematical modeling* (pp. 95–124). New York: Academic Press.

Til, A. van, Beeker, A., Fasglio, D. & Trimbos, B. (2011). *Toetsen en beoordelen met het ERK*. Cito/ SLO: Arnhem/Enschede.

Uebersax, J.S. (1992). A review of modeling approaches for the analysis of observer agreement. *Investigative Radiology, 17*, 738-743.

Van der Linden, W.J. & Hambleton, R.K. (1997). *Handbook of Modern Item Response Theory*. New York: Springer.

Verhelst, N.D. (1993). Itemresponstheorie. In: T.J.H.M. Eggen & P.F. Sanders (red.). *Psychometrie in de praktijk.* (pp. 83-178). Arnhem: Cito.

Verhelst, N.D. (2009). Profielanalyse met item response theory. Arnhem: Cito.

Verhelst, N.D., & Glas, C.A.W. (1995). The one-parameter logistic model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models. Foundations, recent developments, and applications*. New York: Springer-Verlag.

Verhelst N.D., Glas, C.A.W., & Verstralen, H.H.F.M. (1995). *OPLM: One-Parameter Logistic Model. Computer program and manual*. Arnhem: Cito.

Appendix

Appendix

Participants

Table 1 Name, country of origin and employer of the participants in the expert panel for German

Name	Country of origin	Employer
Beunk, René	Netherlands	Candea College Duiven
Boers-Müller, Anne-Marie	Netherlands	SLO
Bormann-Knoll, Simone	Germany	Hamburger Volkshochschule
Dengler, Stefanie	Germany	Goethe Institute
Fasoglio, Daniela	Netherlands	SLO
Ferbezar, Ina	Slovenia	Universität Ljubljana
Gippner, Gabriele	Germany	Humboldt Universität Berlin
Kecker, Gabriele	Germany	TestDaF-Institut
Kunkel-Razum, Kathrin	Germany	Dudenverlag
Kuri, Sonja	Italy	Universität Udine
Mitteregger, Brigitte	Austria	ÖSD (Österreichisches Sprachdiplom Deutsch)
Roche, Jörg	Germany	LMU (Ludwig-Maximilians-Universität München)
Rodenbücher, Ingrid	Germany	Ministerium für Schule und Weiterbildung, Nordrhein-Westfalen
Schellens, Britta	United Kingdom	Goethe Institute London
Studer, Thomas	Switzerland	Universität Fribourg
Touraki, Katerina	Greece	Goethe Institute Athens
Wertenschlag, Lukas	Switzerland	Universität Freiburg
Zeidler, Beate	Germany	Telc GmbH (The European Language Certificates)

Table 2 Name, country of origin and employer of the participants in the expert panel for English

Name	Country of origin	Employer
Beeker, Anne	Netherlands	SLO
Bos, Jonna	Netherlands	Isendoorn college / Cito
Budreikiene, Irena	Lithuania	Utena Adolfas Sapoka Gimnasium
Denies, Katrijn	Belgium	KU Leuven, Research Unit in Educational Effectiveness and Evaluation
Etxeandia, John	Spain	Basque Education Department
Froehlich, Veronika	Germany	University of Education Heidelberg
Helness, Hildegunn L.	Norway	University of Bergen
Holt, Peter	Turkey	Sabanci University, Istanbul
Huhta, Ari	Finland	University of Jyväskylä, Centre for Applied Language Studies (CALS)
Kollias, Charalambos	Greece	Hellenic American University
Kvasova, Olga	Ukraine	Taras Shevchenko, National University of Kiev
Lammervo, Tiina	Finland	University of Jyväskylä
Märcz, Robert	Hungary	Foreign Language Centre, University of Pécs
Moe, Eli	Norway	University of Bergen and Vox
Pascoal, José	Portugal	Assessment Centre for Portuguese, University of Lisbon
Rini, Danilo	Italy	University for Foreigners, Perugia
Romera, Josu	Spain	Escuelas Oficiales de Idiomas (Basque Government Spain)
Spoettl, Carol	Austria	University of Innsbruck, Institut für Anglistik
Tsagari, Dina	Cyprus	Department of English Studies

Table 3 Name, country of origin and employer of the participants in the expert panel for French

Name	Country of origin	Employer
Aler, Trees	Netherlands	VLLT (Vereniging van Leraren in Levende Talen)
Amar, Faezeh	France	IRFFLE (Institut de Recherche et de Formation en Français Langue étrangère) de l'Université de Nantes
Baraona, Geneviève	France	Institut National des Langues et Civilisations orientales
Basterra, Maite	Spain	Escuelas Oficiales de Idiomas (Basque Government-Spain)
Beltran, Laurence	France	Université d'Avignon
Bickel, Marguerite	France	Editions Didier
Brems, Maria	Belgium	AKOV (Flemish Agency for Quality Assurance in Education and Training)
Folny, Vincent	France	CIEP (Centre international d'études pédagogiques)
Härmälä, Marita	Finland	The Finnish National Board of Education / University of Jyväskylä, Centre for Applied Language Studies (CALS)
Hoppe, Christelle	France	IRFFLE (Institut de Recherche et de Formation en Français Langue étrangère) de l'Université de Nantes
Jong, Kim de	Netherlands	SLO
Kancellary Delage, Catherine	France	Université de Bordeaux
Koecher, Liliane	France	Institut International d'Études Françaises, Université de Strasbourg
Mertens, Jürgen	Germany	Université des Sciences de l'Éducation Ludwigsburg
O'Leary, Christine	United Kingdom	Sheffield Hallam University
Segeat Mistretta, Marina	France	Schola Mediterranea
Senges, Sylvie	France	Université de Bordeaux
Thomaes-Jauréguiberry, Dominique	Netherlands	Montessori College Eindhoven/Fontys University of Applied Sciences, Teacher Training Programme in French, Tilburg
Wauthion, Michel	Netherlands	Institut français des Pays-Bas / Ambassade de France aux Pays-Bas
Weiler, Theresa	Austria	BIFIE (Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens)

Cito

Secondary education

Performance Standards for the CEFR in Dutch secondary education

An international standard setting study

Cito
Amsterdamseweg 13
P.O. Box 1034
6801 MG Arnhem
The Netherlands
T +31 26 352 11 11
www.cito.nl

Cover photo: Ron Steemers