

Mapping the Dutch Foreign Language State Examinations onto the Common European Framework of Reference

Report of a Cito research project commissioned by
the Dutch Ministry of Education, Culture and Science

José Noijons, Henk Kuijper

October 2006

Contributions to this report by:

Erna van Hest

Ton Heuvelmans

Gunter Maris

Tine Plant

Evelyn Reichard

Digna Samson

Norman Verhelst

© Cito 2006, Arnhem, the Netherlands (*Cito: National Institute of Educational Measurement*).

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, scanning, recording, or otherwise, without the prior permission of the Copyright owner.

Contents

CONTENTS	3
PREFACE	5
INTRODUCTION	7
1 LINKING PROCEDURE	9
2 FAMILIARISATION	11
3 CONTENT SPECIFICATION	13
3.1 INTRODUCTION	13
3.2 CONTENT SPECIFICATION IN RELATION TO THE SCALES FOR COMMUNICATIVE ACTIVITIES.	13
3.2.1 <i>Method</i>	13
3.2.2 <i>Description per global descriptor</i>	14
3.2.3 <i>Description with detailed descriptors</i>	16
3.2.4 <i>Text dimensions: text source, text type, topic and domain</i>	16
3.2.5 <i>Conclusions and recommendations</i>	21
3.3 CONTENT SPECIFICATION IN RELATION TO THE SCALES FOR COMMUNICATIVE COMPETENCE	22
3.3.1 <i>Introduction</i>	22
3.3.2 <i>Method</i>	22
3.3.3 <i>The linguistic and cognitive complexity of texts</i>	24
3.3.4 <i>Task complexity</i>	30
3.3.5 <i>Conclusions texts and tasks</i>	35
4 STANDARDISATION	37
4.1 INTRODUCTION	37
4.2 JUDGEMENT PROCESS	37
4.3 DATA ANALYSIS FOR VALIDATION OF THE STANDARDS	39
4.4 DETERMINING MINIMUM SCORES FOR RELEVANT CEFR-LEVELS IN EACH EXAMINATION.	42
4.4.1 <i>Results standard-setting English</i>	42
4.4.2 <i>Results standard-setting French</i>	48
4.4.3 <i>Results standard-setting German</i>	50
5 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS	55
5.1 SUMMARY	55
5.2 CONCLUSIONS	55
5.3 RECOMMENDATIONS	56
6 REFERENCES	57
ANNEX THE DUTCH EDUCATIONAL SYSTEM	59

Preface

The Common European Framework of Reference for Languages, CEFR, has developed into an important framework for the teaching and learning of foreign languages within Europe. In the Netherlands it has been the basis for a number of innovative developments in the area of foreign language learning, such as the development of Language Portfolios and Language Profiles.

The concept of levels and the concrete description of language abilities in the foreign languages have made the CEFR a very useful framework of reference for the Dutch state examination syllabuses and the state examinations in the foreign languages.

Mapping onto the CEFR can answer the need:

- (1) to make foreign-language teaching in the Netherlands more competence-oriented;
- (2) to make examination requirements for foreign languages in the Netherlands more transparent for both teachers and pupils;
- (3) to make foreign language achievements of Dutch pupils internationally comparable.

Commissioned by the Dutch Ministry of Education and Science in 2005, the Dutch Institute for Curriculum Development (SLO) and the Dutch Institute for Educational Measurement (Cito) have carried out a study into the mapping of examination syllabuses and examinations of foreign languages onto CEFR. The results of that part of the study that was carried out by Cito have been presented in a research report published in January, 2006.

An authorized translation for an international audience of the research report was commissioned by the Dutch Ministry of Education, Culture and Science in June 2006. It is this translation that is presented here.

Dr Erna van Hest
Head of the Department for Examinations in Lower Secondary Education
Cito

Introduction

In 2003 the Dutch Ministry of Education, Culture and Science commissioned SLO and Cito to carry out a linking project which had the following objectives:

- A. to establish links between the existing examinations in French, German and English and the Common European Framework of Reference, CEFR (Council of Europe, 2001), following the steps as outlined in the (preliminary pilot version of the) Manual published by the Council of Europe;
- B. to study the possibilities of developing more comprehensive CEFR-related examinations in the foreign languages.

The project defined a number of goals, resulting in the production of the following:

1. a qualitative analysis of the examination syllabuses and examinations (2003 and 2004) for lower secondary pre-vocational level and for higher secondary/pre-university level French, German and English (carried out by SLO and Cito);
2. a new set of specifications for the examination syllabuses 2007 at the higher levels based on the CEFR for the skills of reading, listening, writing, spoken production and spoken interaction (produced by SLO);
3. a classification of the items in the examinations for French, German and English at five levels within the terms of the CEFR by judges and stakeholders, with a computation of cut-off scores at relevant CEFR levels (carried out by Cito);
4. a psychometric validation of the above-mentioned examinations (carried out by Cito);
5. a set of sample items illustrating those can-do statements at each CEFR-level that are not tested in the present examinations (produced by Cito);
6. prototypes of “CEFR-based examinations of reading” in French, German and English at relevant levels (produced by Cito);
7. inclusion of research data in the syllabuses for the state examinations in the foreign languages that will be produced within the framework of the revision of the examination syllabuses for higher secondary and pre-university level (carried out by Cito in cooperation with SLO);
8. A final report published by Cito in cooperation with SLO.

In 2003 the Council of Europe published a draft Manual for linking language examinations to the CEFR, hereafter named *Manual* (Council of Europe, 2003). In this Manual several methods are described to give a scientific basis to the claim of links of an examination to the CEFR. In this report the activities are described according to the following two methods as worked out in the Manual: (1) linking to (or mapping onto) the CEFR on the basis of specification and (2) linking to the CEFR on the basis of specification AND standardisation. The third method: (3) linking to the CEFR on the basis of empirical validation is being carried out by Cito in 2006.

The research that is reported on here relates to the state examinations of reading comprehension, only, as they have been determined by the State Examination Committee (CEVO). The contents of the *school-based* examinations (of listening, writing and speaking) are the responsibility of the individual schools and may vary to such a degree that general procedures would not have been possible.

The results of this research have been published in two Cito-publications in the Dutch language: Noijons (2006a, 2006b) The results refer only to those parts of the research project that have been conducted by Cito. The results of the SLO research can be found in the examination syllabi published by the State Examination Committee (CEVO, 2006).

1 Linking procedure

The aim of the Manual published by the Council of Europe has been to propose possible procedures for finding evidence to relate examinations to the CEFR. Until then testing institutes would link their tests to the CEFR using methods of their own choice. Some of such research projects have been documented extensively, others less extensively. Such research projects are difficult to compare with each other because of the differences in methods used. Several institutions have different ideas about what CEFR-levels entail and how they can be related to levels of examinations. The Manual offers an excellent opportunity to avoid such problems of comparison.

In the Manual three methods of linking are identified:

1. specification of the contents of the examinations;
2. standardisation of judgements;
3. empirical validation by means of data-analysis of test results.

With the first method the claim of links to the CEFR is done on the basis of specification only. This method (with variations) has been followed by a number of institutes in Europe and outside. The second method may lead to a stronger claim on links, because it is based on specification AND standardisation. The third method may result in claims that are confirmed by empirical verification. Methods 1 and 2 have now been carried out by Cito and are described in this report. Cito is carrying out activities relating to method 3 in the year 2006.




Within the various methods the following phases have been identified:

- *Familiarisation*: making sure that those persons involved in the linking procedure are thoroughly acquainted with the goal, the set-up and the levels of the CEFR (Introduction phase).
- *Specification*: mapping out the extent to which the coverage of the examination syllabuses and the state examinations of reading comprehension can be related to descriptors of the CEFR.
- *Standardisation*: asking judges (language teachers, representatives from politics and industry) through a procedure as described in the Manual to classify a selection of the examination questions that were used in the research project, in terms of CEFR (that is, linking them with CEFR-levels). On that basis it can be determined what score a candidate should get in a particular examination in order to be able to say that he/she masters a level that is relevant to CEFR for that particular examination.
- *Empirical validation*: psychometric validation of the results collected in the standardisation procedure.

In this report the activities that relate to the first three phases are described. These three phases are part of what is called the *internal* validation. The last phase, empirical validation is part of the *external* validation.

In the scheme that follows we demonstrate the relationships between the various phases. This scheme is based on a scheme from the Manual (Figure 1.1). As can be seen, the idea is that each step in the linking process strengthens the claim that an examination is linked to the CEFR.

Figure 1.1 Schematic representation of the steps to be followed for collecting evidence to relate tests and examinations to the CEFR

SPECIFICATION OF EXAMINATION CONTENT	STANDARDISATION OF JUDGEMENTS	EMPIRICAL VALIDATION THROUGH ANALYSIS OF TEST DATA
<p>Internal validity Description and analysis of:</p> <ul style="list-style-type: none"> • general examination content; • process of test development; • marking, grading, results. <p>Test analysis and post-examination review</p>	<p>Familiarisation</p> <ul style="list-style-type: none"> • training in standardised samples of productive skills • training in standardised samples of receptive skills and linguistic competence 	<p>Internal validation</p> <ul style="list-style-type: none"> • Classical test theory • Methods of qualitative analysis • Generalisation theory • Factor analysis • Item response theory
<p>External validity Relate</p>	<p>Benchmarking local performance samples to CEFR levels</p>	<p>External validation</p>
<ul style="list-style-type: none"> • general examination description to CEFR scale; • description of communicative activities tested to CEFR scales; • description of aspects of communicative language competence tested to CEFR scales. 	<ul style="list-style-type: none"> • determination of standards • dissemination and implementation 	<ul style="list-style-type: none"> • correlating to results on tests already calibrated on CEFR • correlating of judgements to CEFR descriptors • anchoring of a test to a test already calibrated on CEFR • anchoring of a test directly to the scales behind the scale values of the CEFR descriptors
		
<p>Claim of link to CEFR on basis of specification</p>	<p>Stronger claim of link to CEFR on basis of specification AND standardisation</p>	<p>Confirmation of claim of link to CEFR on basis of empirical verification</p>

2 Familiarisation

The first phase in the linking process described in the Manual is that of familiarisation. The Manual points out that it is important for all those involved in the linking process to familiarise themselves with the CEFR. This can partly be done through self study, but in the present large scale linking project it was felt to be of prime importance that there was agreement among the different members of the project in their interpretation of CEFR categories and levels. Project members (content specialists) were based at two different institutes (Cito and SLO) and many of their activities would be undertaken at their respective home bases. For this reason a great part of the phase of familiarisation took place in joint sessions in which the representatives of the two institutes discussed their interpretations of the CEFR with each other and in which there was an ambition to come to one single view of the CEFR. All project members (content specialists) have an academic degree in one or more of the foreign languages involved and a first inventory showed that every project member was estimated to function in at least one foreign language at B2-level or higher. The familiarisation process was coordinated by the project member who had gained experience in the development of the diagnostic testing system DIALANG which is based on the CEFR.

The familiarisation process has involved five steps:

1. Familiarisation with the global and specific aims, objectives and functions of the CEFR. For this purpose Chapter 1 of the CEFR has been distributed to all project members (content specialists) for self study.
2. Discussions with reference to the questions that are put at the end of each chapter in the CEFR about the relevance of the chapter in question for the work situation of the user of the CEFR. The discussion was narrowed down to the relevance of the CEFR for the development of curricula, syllabuses and tests.
3. Discussion about the global descriptors of the CEFR and the levels that go with them. Project members (content specialists) made a first, preliminary link of the Dutch education levels to the CEFR. Project members had to do an exercise in which they related the global descriptors to the relevant CEFR-level.
4. Self-assessment of one's own proficiency level for project members in two foreign languages with the help of the self assessment grid (table 2 in the CEFR). The first foreign language was the language that project members had studied (and were qualified to teach in), the second language was a language (of their own choice) that they had gained some proficiency in. These self-assessments were discussed extensively.
5. Sorting more specific individual CEFR descriptors. The descriptors all pertained to reading, the skill that is tested in the state examinations. The descriptors were taken from Appendix C1 of the CEFR, which were developed within the framework of the project DIALANG.

The discussions showed that there was global agreement on the CEFR levels attained in Dutch foreign language education. It must be emphasized here that these estimates were not based on any empirical evidence. The level reached for French was considered to be lower than for German, which in its turn is considered lower than for English. Reading was seen as the skill that is mastered at the highest level by pupils. The lowest level examinations for reading were estimated to be at A2 (English) and between A1 and A2 for German and French. The pre-university examinations for reading were estimated to be at B2 (English), between B1 and B2 for German, and at B1 for French. For writing and speaking it was supposed that pupils at pre-university level would achieve B1 for English and A1-A2 for the other levels of education and with the other languages. Again, these were estimates by project members, their estimates were not based on empirical evidence. However, for the continuation of the project it was important that one could start from a common point of view. All project members estimated their own CEFR-levels. Most of them have university degrees (BA or MA) in one or more foreign languages. They estimated their own CEFR level of proficiency at B2 for reading and listening in the language that they had studied, and at B1-B2 for speaking and writing. Some of them claimed C1 for reading.

3 Content specification

3.1 Introduction

The second phase in the linking process as outlined in the Manual and carried out in the Dutch linking research is called the *specification* phase. The Manual describes what specification in the linking process involves: describing the extent to which an examination covers the categories and levels of the CEFR. The Manual identifies two separate forms of description:

1. a description of the examination in its own right;
2. a content analysis of the examination (in terms of the CEFR).

In the linking study as described here, Cito and SLO have interpreted the first activity as follows. A qualitative analysis is made of the examination syllabus for foreign languages. This analysis is made for each descriptor in the CEFR using the current Dutch examination specifications. This analysis has been carried out by SLO (SLO 2006a, 2006b).

In this chapter the results are reported of the content analysis of the state examinations of reading comprehension in terms of the CEFR. This content analysis of the examinations consists of a description of the examinations through the scales of communicative activities and the scales of communicative competence in the CEFR.

The specification of the examination texts and items has been split up into two parts.

In the first part of the specification phase the focus is on the scales of communicative activities. In this part the examination texts and items have been analysed by means of the four global descriptors and the accompanying detailed descriptors in *Taalprofielen* [Language Profiles] (Liemberg, Meijer, 2004). This is described sections 3.2.2 and 3.2.3. In addition to this, texts have been described on the basis of the following context variables from the CEFR: text source, text type, communicative topics and domain. These results are reported on in section 3.2.4. Section 3.2.5 contains conclusions and recommendations of this first part of the content specification.

The second part of the specification phase consists of a description of the texts and items. This description was made with the help of a description model based on CEFR, the *Dutch Grid* (for a description of the Dutch Grid see section 3.3.2). With the Dutch Grid it is possible to make a close analysis of the linguistic and cognitive complexity of the texts and items. In this part of the specification phase the emphasis is on the scales for communicative competence. This part of the specification phase is reported in section 3.3. In section 3.3.3 the complexity of the texts is discussed. In section 3.3.4 the focus is on the items or tasks in the examinations. Section 3.3.5 contains the conclusions.

3.2 Content specification in relation to the scales for communicative activities.

3.2.1 Method

Within the context of linking the state examinations of reading comprehension in French, German and English to the CEFR, the examinations of the year 2004, for the school types bb (with the exception of French), kb, gl/tl, havo and vwo¹ have been analysed with the help of the four global descriptors and the accompanying detailed descriptors of *Taalprofielen* [Language Profiles] (Liemberg, Meijer, 2004). These global descriptors correspond to the scales of communicative activities for reading in the CEFR. *Taalprofielen* identifies the following global descriptors:

- Reading correspondence
- Reading for orientation
- Reading for information and argument
- Reading instructions

¹ An overview of the Dutch educational system is given in the annex.

These global descriptors in turn are subdivided into detailed descriptors (can-do statements).

In the standard-setting phase (see chapter 4) a minimum CEFR-level has been estimated that would be required to successfully respond to items from each examination for each school type. Based on the estimation of the minimum CEFR-level for an individual item a mean minimum CEFR-level could be computed for each examination. This required minimum CEFR-level is shown in table 3-1 for the examinations of each school type. A score of 1 means A1, a score of 2 means A2, a score of 3 means B1, a score of 4 means B2 and a score of 5 means C1.

Table 3-1 Mean required minimum CEFR level in the examinations of Reading Comprehension

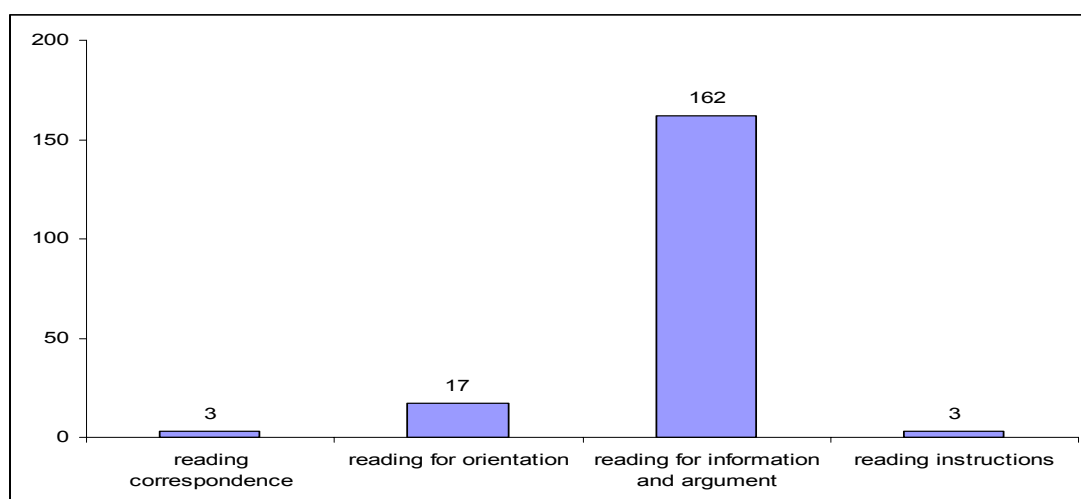
School type ²	English	German	French
bb	1.9	2.3	1.7
kb	2.5	2.7	2.1
gl/tl	2.7	2.6	2.5
havo	3.8	3.8	3.1
vwo	4.3	3.8	3.4

On the basis of the estimations given in table 3-1 the examinations have been analysed with the descriptors for the level or the levels to which the required minimum level corresponded. For reasons of efficiency judges were instructed to assign only one global descriptor to each item. The descriptor that was selected had to be the most appropriate. The next step was to determine to which detailed descriptor each reading item mainly referred to. The assigning of descriptors to the items was carried out by two judges in close consultation.

3.2.2 Description per global descriptor

Figures 3-1 to 3-3 illustrate how the items for the state examinations of reading comprehension in English, French and German are distributed across the global descriptors. They also show for which subsets of the descriptors in the examinations no items or only few items were found.

Figure 3-1 Distribution of items in English across the global descriptors taken from *Taalprofielen* [Language Profiles]



² See the Annex for an overview of the Dutch Educational System

Figure 3-2 *Distribution of items in French across the global descriptors taken from Taalprofielen [Language Profiles]*

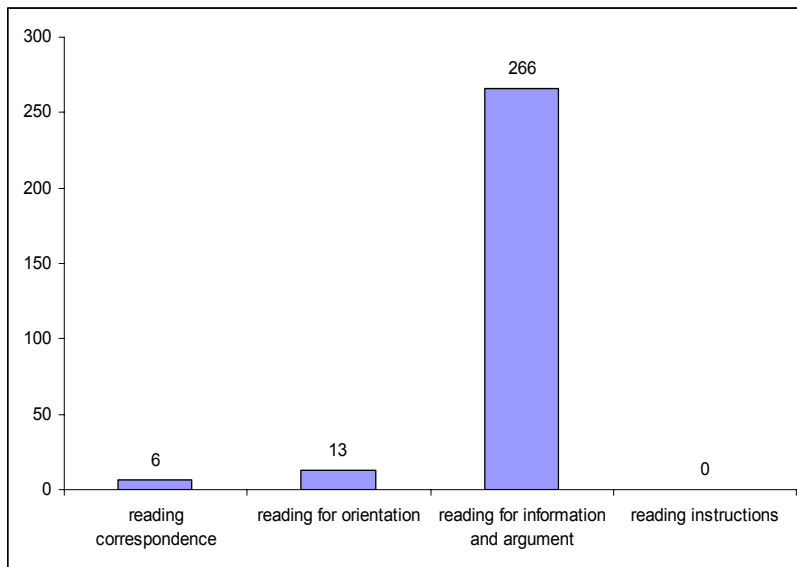
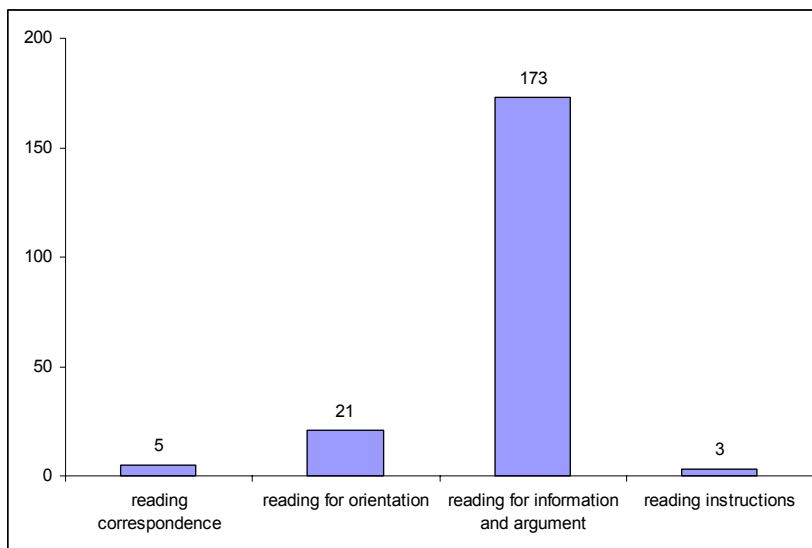


Figure 3-3 *Distribution of items in German across the global descriptors taken from Taalprofielen [Language Profiles]*



It can be seen that for all languages most items refer to the descriptor 'reading for information'. Per descriptor the following remarks can be made.

- **Reading correspondence**
2-3% of the items tap reading correspondence. In the analysis it appeared that at the lower levels these mainly referred to short e-mails or simple letters to the editor. At the higher levels they referred to letters to the editor in which an opinion or arguments were put forward.
- **Reading for orientation**
About 10% of the items tap reading for orientation. Such types of items can be found at all levels.
- **Reading for information**
For all the state examinations of reading comprehension the focus is on this global descriptor. This appears to be the case for about 85% of the items. The variation in items/source material appeared to be wide.

- Reading instructions
In the examinations at the lower levels, items that relate to instructions may occur occasionally, at the higher levels such items hardly occur.

3.2.3 Description with detailed descriptors

Each global descriptor is accompanied by a number of detailed descriptors which describe what a person can do. The items in the examinations have also been classified with the help of these detailed descriptors. The tables going with this classification are only available in the original Dutch version of this report (Cito, 2006a)

The classification tables show that a number of descriptors were not operationalized in the examinations at all, or only occasionally. However, the following points may be taken into consideration.

- The four global descriptors refer to two different categories: (1) *reading instructions* and *reading correspondence* relate to the type of *text* in reading, whereas (2) *reading for information* and *reading for orientation* relate to the reading *objective*. This is a critical problem in arriving at a valid classification.
- If reading correspondence seems to be little operationalized, this need not mean that the skills as described in the detailed descriptors under reading correspondence, are not represented in the examination. For both reading correspondence and reading for information candidates have to be able to understand the main ideas and detailed information in texts. It can be argued that when a candidate demonstrates this skill in reading an article in a magazine – as mentioned with reading for information – he or she will also demonstrate this skill when reading for correspondence. The difference is not in the reading skill, but in the text type.
- The starting point for the assigning of descriptors to items was that to each reading item only one, the most appropriate global descriptor could be assigned. This method rules out the possibility of showing that items can refer to several detailed descriptors, even though this in practice may often be the case (see below).
- The detailed descriptors have sometimes been formulated in such global terms that they could relate to all the items in an examination. See for instance at A2-level the descriptor *can understand short descriptive texts on familiar topics*. This particular descriptor has not been used very often because a more specific descriptor was preferred.
- Only one examination year has been analysed (2004, first session). The description of one examination year may not provide a complete picture of the type of tasks/texts that could be found in the examinations. In other words, if the examination in 2004 does not contain a specific task, this does not mean that such a task will not be present in the examination of another year.
- Two factors may cause some specific descriptors not to be represented: tasks that are not included in the examination syllabus and the examination model, such as - at B2-level - the element speed: *can read letters or e-mails on topic in his own sphere of interest with ease and can grasp the main idea fast*, and tasks that cannot be tested, such as - at B1-level - the factor pleasure: *can read simple texts for pleasure*. Sometimes descriptors cannot be tested because they would go against the testing principle of efficiency, or because they would lead to an unclear marking scheme.

3.2.4 Text dimensions: text source, text type, topic and domain

The CEFR contains a number of dimensions to describe texts. These dimensions have been included in the *Dutch Grid*, a descriptive model for reading- and listening items. For a short description of the Dutch Grid we refer to section 3.3.2.

In the Dutch Grid the following text dimensions are used: text source, text type, communication themes and domain. Text source, communication themes and domain are derived directly from the CEFR and can be looked upon as generic dimensions across all levels. Text type is a classification

taken from DIALANG and it is developed as a globalization of the less systematic description of this dimension in the CEFR. These dimensions can be used to describe and regulate the variety in texts and topics in examinations, dependent on the aim and function of the examination. In this section the state examinations of reading comprehension in English and German are described by means of these four text dimensions. For the examinations of French not enough data were available to be able to carry out this analysis.

Text source

The CEFR lists the following text sources.

Personal	Public	Occupational	Education and Training
Video text	Announcements and messages	Business letters	Authentic texts
Warranty form	Labels	Report/memorandum	Text books
Recipe	Pamphlets, graffiti	Safety instructions	Reference books
Instructional material	Tickets, timetables	Instructional material	Text on blackboard
Novel	Regulations	Regulations	Text on sheet
Magazines	Programmes	Advertising material	Text on screen
Newspaper	Contracts	Tickets etc.	Video text
Junk mail	Menus	Vocational descriptions	Text for practice
Leaflets	Devotional texts	Sign posts	Work or exercise materials
Personal/private letters		Business cards	Articles from magazines
			Abstracts
			Dictionaries

Figures 3-4 and 3-5 show which text sources were the basis for the texts in the state examinations of reading comprehension in German and English.

Figure 3-4 *Text source of the texts in the state examinations of reading comprehension in German as percentage*

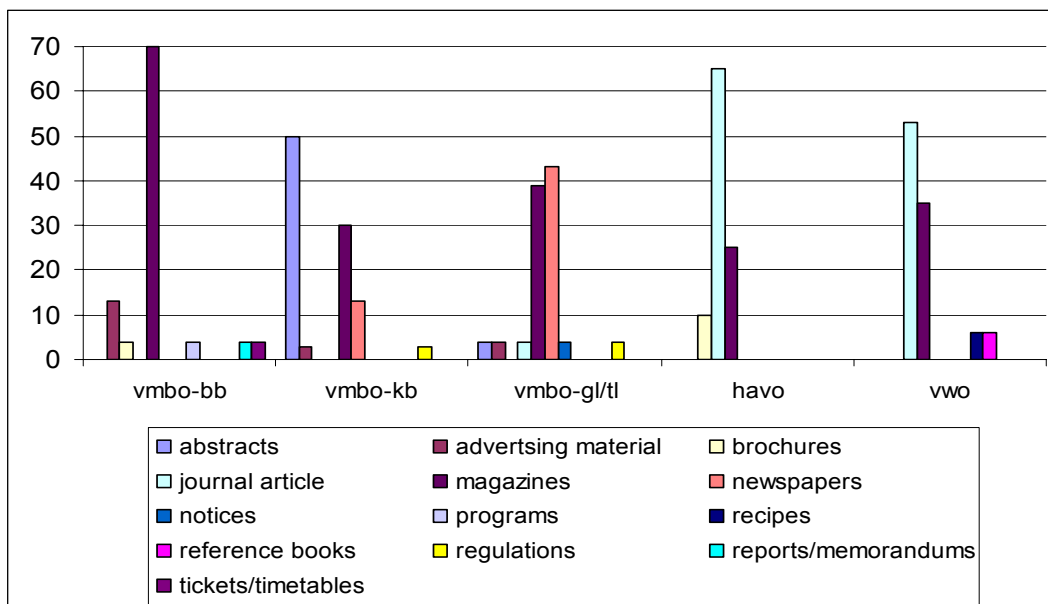
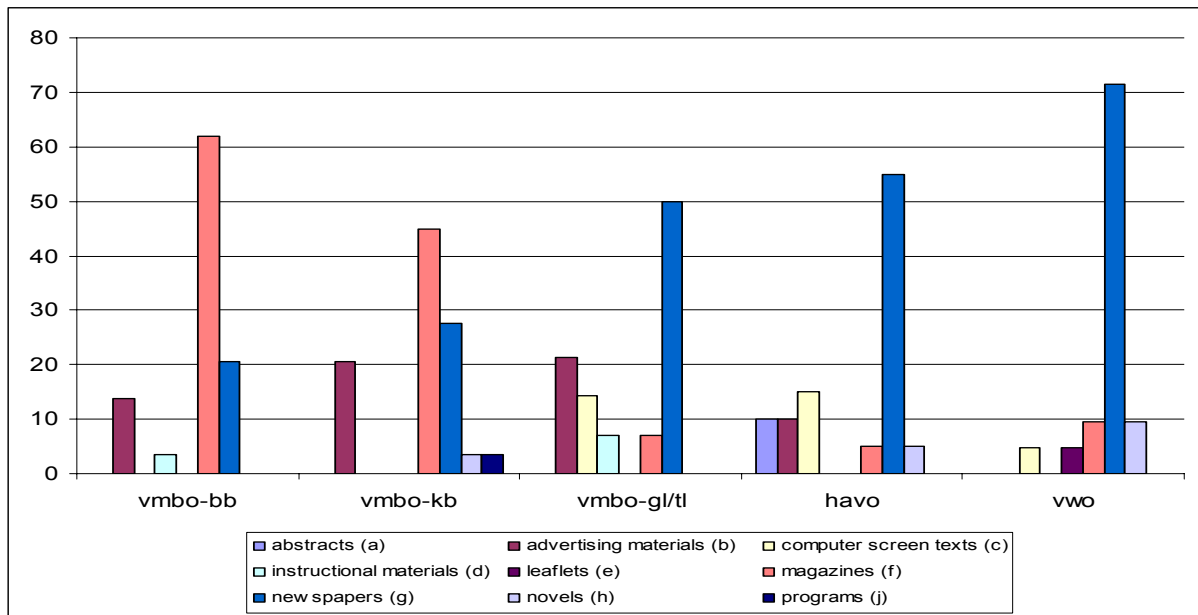


Figure 3-5 *Text source of the texts in the state examinations of reading comprehension in English as percentage*



We see that for German 13 and for English 9 of the text sources mentioned in the CEFR have been used in the respective examinations. The emphasis is on everyday texts from newspapers and magazines in both the English and the German examinations.

Text type

In the reading scales of the CEFR, from the lower levels to the higher levels, a shift from more descriptive text types to more expository and argumentative text types can be seen. In the specification phase the examination texts were classified according to text type in order to be able to see if the same tendency could be seen. Figures 3-6 and 3-7 show the result of this classification for the examination texts in German and English.

Figure 3-6 *Text types per school type state examinations of reading comprehension in German in percentages*

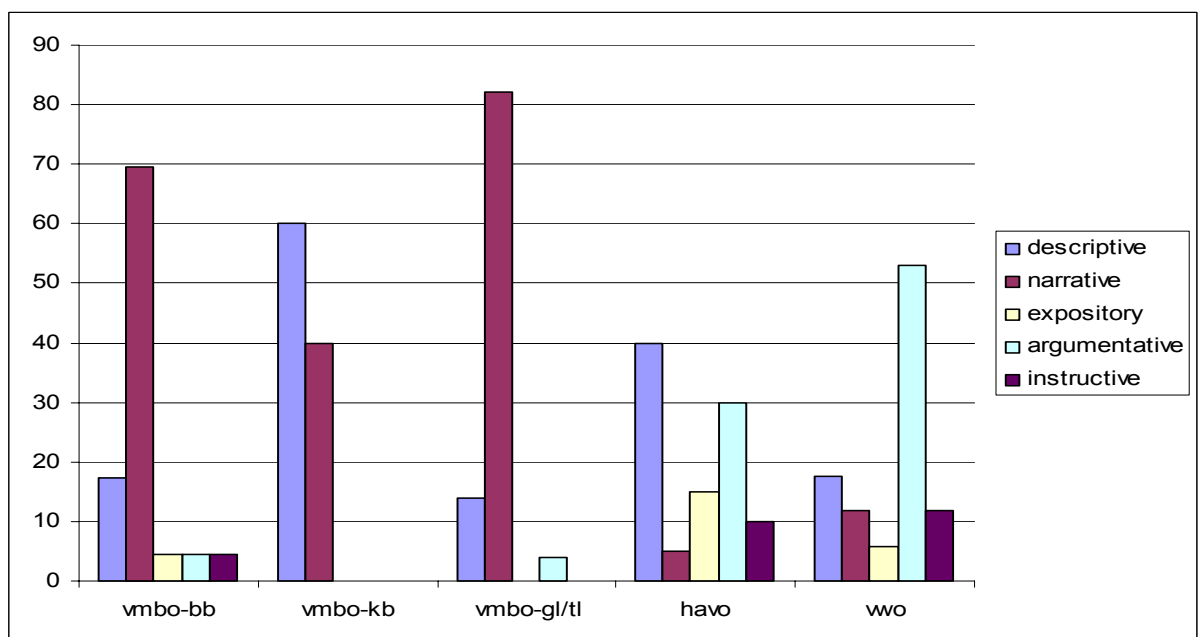
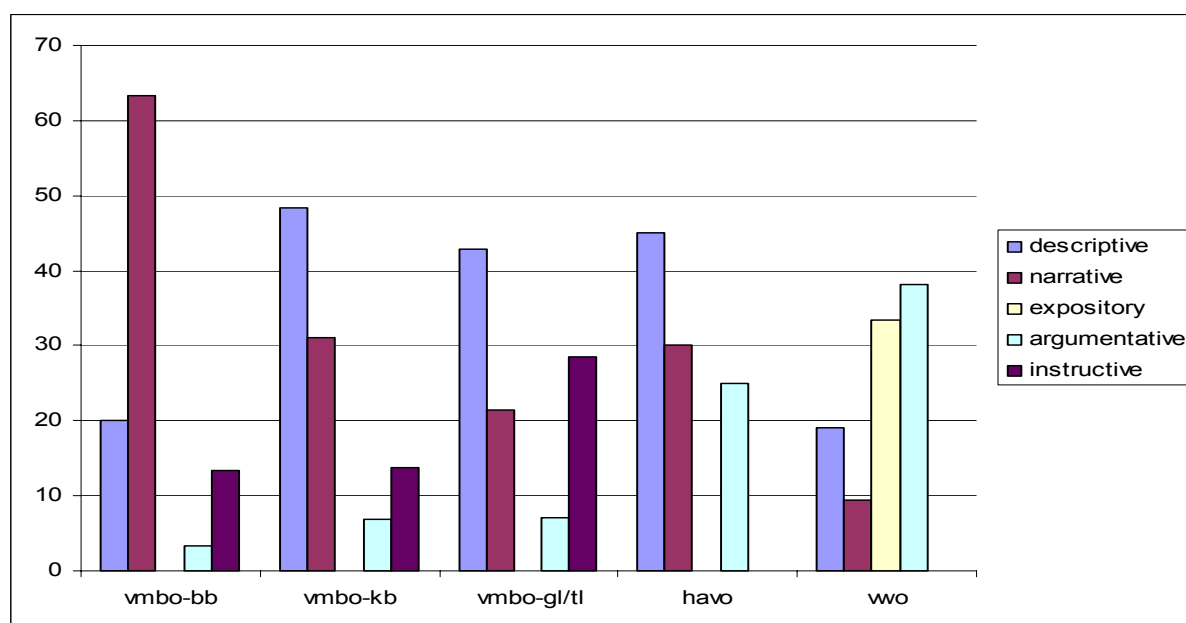


Figure 3-7 Text types in the state examinations of reading comprehension in English in percentage



The examinations for vmbo (bb, kb, gl/tl) are characterized by more descriptive text sources, whereas for havo, and even more so for vwo, expository texts occur more frequently. This is in line with the descriptors in the CEFR.

Communication topics

The CEFR lists the following communication topics.

- | | | |
|----------------------------|-------------------------------|-------------------|
| 1 Personal identification | 5 Travel | 9 Shopping |
| 2 House, home environment | 6 Relations with other people | 10 Food and drink |
| 3 Daily life | 7 Health and body care | 11 Services |
| 4 Free time, entertainment | 8 Education and training | 12 Places |
| | | 13 Language |
| | | 14 Weather |

As with the CEFR-list of text sources, the CEFR does not claim the list of communication topics to be exhaustive and definitive. Also, the categories used are neither mutually exclusive nor prescriptive.

Figures 3-8 and 3-9 show the extent to which these communication topics are dealt with in the state examinations of reading comprehension in German and English.

Figure 3-8 *Communication topics in the state examinations of reading comprehension in German per school type in percentages*

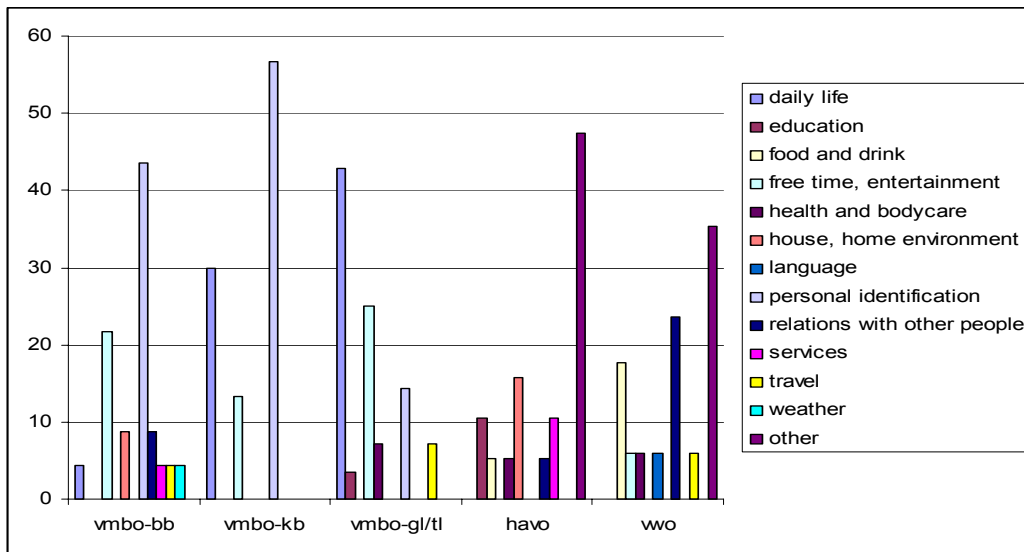
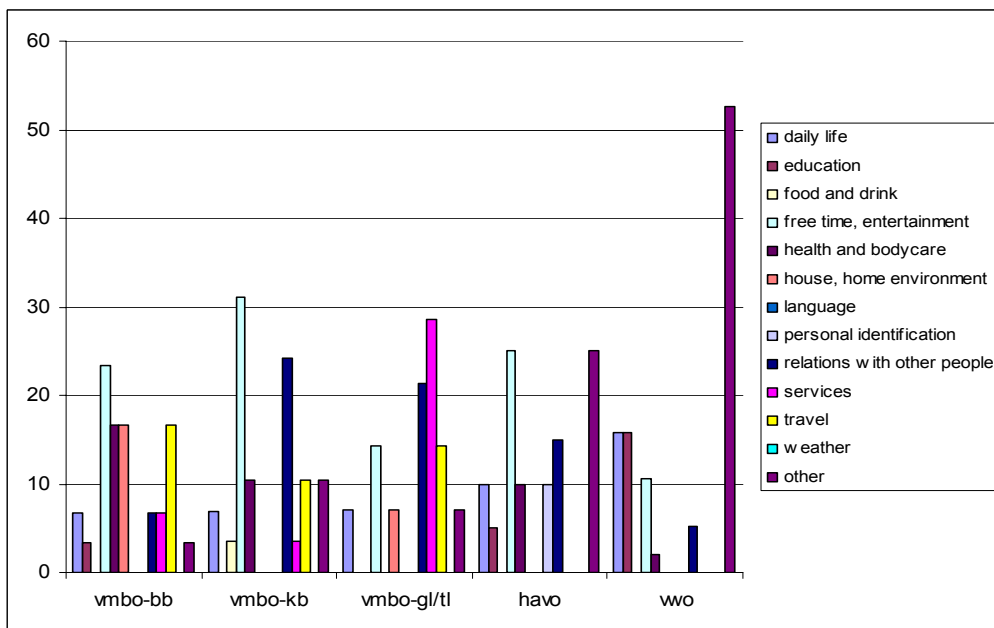


Figure 3-9 *Communication topics in the state examinations of reading comprehension in English per school type in percentages*



It can be seen that nearly all the communication topics of the CEFR appear in the examinations, although not all topics are represented in each examination.

Domains

The CEFR contains four domains:

- personal
- public
- work
- education

From the classification of the texts in these domains it appears that for the most part the examination texts refer to the personal domain and to a lesser extent to the public domain.

3.2.5 Conclusions and recommendations

In sections 3.2.2 and 3.2.3 a description has been given of the specification of the items in the state examinations of reading comprehension in English, French and German on the basis of the global and detailed descriptors in *Taalprofielen* [Language Profiles]. In section 3.2.4 a description has been given of the CEFR-text dimensions, text types, communication topics and domains.

- The distribution of the items across the four global descriptors is not balanced. The items that focus on reading for information appear most often. This has to do with regulations in the examination models which prioritise this type of reading, and with considerations mentioned in section 3.2.3.
- The current examination syllabus and the examination model that is derived from it, are in fact very global as far as the texts and tasks are concerned. The diversity of the source material mentioned in the descriptors in *Taalprofielen* corresponds with the great variety of text materials that can be used for the state examinations. This variety has been confirmed by the fact that the examination texts represent a great variety of what is described in the CEFR for text type, text source and communication topics.
- It is exactly this very wide variety that makes an extensive revision of the current examination syllabuses less opportune. It is for this reason that constructing new examination prototypes has not been deemed advisable.
- Although not all four global descriptors are equally represented in the examinations, it appears that the skills that are described in the detailed descriptors are all represented in the examinations. Reading skills which are described under the global descriptors show a great deal of overlap. When one concludes for instance that *reading correspondence* does not occur very often, this does not mean that the skills as described in the detailed descriptors under reading correspondence are not represented in the examinations. For both *reading correspondence* and *reading for information*, candidates have to be able to understand the main ideas and detailed information in texts. It is likely that when a candidate is able to demonstrate this skill when reading an article from a magazine - as is mentioned with *reading for information* – he or she can also show this skill when reading correspondence. The difference then is not in the reading comprehension, but in the type of text.
- When revising the examination models one could nonetheless consider the necessity of having the missing detailed descriptors appear more explicitly in the examinations. One can think of work-related correspondence, *reading for orientation* and *reading instructions* for instance for the bb and kb examinations. The global descriptor *reading for information* appears most frequently in the examinations. The detailed descriptor *reading of texts for pleasure* that goes with this particular global descriptor does not appear in the state examinations at all.
- Redistribution of the present attainment targets may be considered so that there is more correspondence with the relative importance that is given to each of the descriptors within the CEFR at the various levels. In other words: the present distribution of attainment targets does not reflect the CEFR sufficiently.

3.3 Content specification in relation to the scales for communicative competence

3.3.1 Introduction

The levels in the CEFR represent an increasing degree of language ability. Language learners can deal with increasingly more text types in many more domains and situations, as their proficiency level increases. Their ability to handle linguistically and cognitively more complex texts with an increasing accuracy also increases. This aspect of the language ability is described specifically in the scales for communicative competence and within those even more specifically in the scales for communicative linguistic competences.

In the following sections additional information is given about the texts and items in the state examinations of reading comprehension in German and English in relation to the CEFR. For the examinations in French not enough data were available to carry out such an analysis. However, it can be assumed that the results presented here for German and English in relation to the CEFR will also apply to the state examinations of reading comprehension in French. After all, the examinations in German, English and French are developed following the same examination syllabus and the same examination model.

In section 3.3.2 the method in this phase of the specification process is explained. In sections 3.3.3 and 3.3.4 the results of the analysis of respectively the reading texts and the items are reported. Section 3.3.5 contains a summary and conclusions.

3.3.2 Method

In the CEFR content aspects and levels of language tasks have been described. The language development from a lower to a higher level can be described as the increase in the ability to perform more and more language tasks with text types of an increasing difficulty level.

The question is to what extent the state examinations of reading comprehension contain texts and items that show an increase in difficulty level from vmbo (bb, kb, gl/tl) to vwo as is to be expected from the standard-setting process (see chapter 4). To be able to answer this question a descriptive, CEFR related model is needed to describe texts and items.

A research project into such dimensions in the CEFR has been carried out in the *Dutch Grid-project* (Alderson, 2006). In this European project the CEFR has been analysed for the extent to which it contains clear instructions and guidelines for the description and development of test tasks at various CEFR-levels. It is concluded that the CEFR is a useful instrument, but that clear guidelines for test development and test description at the various levels cannot be found in the CEFR.

The following problems for test construction on the basis of the CEFR have been identified:

1. Terminology problems: are some technical terms synonyms or not?
2. Omissions, when a concept or characteristic needed for test specification simply is not present.
3. Inconsistencies, when a characteristic is mentioned at one level and not at another level, where the same characteristic occurs at two different levels, or when at the same level a characteristic is described differently in different scales.
4. The absence of definitions, when terms are presented, but not defined.

The problem is not so much in the descriptive criteria in the CEFR being absent, but rather in the fact that there is a lack of explicitness, structure, consistency and precision, aspects which are of vital importance for test construction.

On the basis of a thorough analysis of the CEFR the Dutch Grid project has developed a new descriptive model, related to the CEFR, for reading and listening items and texts. The model attempts at describing the relevant dimensions of the CEFR in a more systematic way. This descriptive model is available on the web at www.ling.lancs.ac.uk/CEFRgrid. The Dutch Grid contains descriptive dimensions for texts and items in reading and listening. The text dimensions can be subdivided into a content category and a category referring to cognitive and linguistic complexity. Table 3-2 gives a survey of these dimensions. For each dimension it is shown if the descriptive structure is derived directly from the CEFR, or is an adaptation of the CEFR or is derived from a different taxonomy.

Table 3-2 *Descriptive dimensions of the Dutch Grid for reading*

Text dimensions	
Content	Source
Text source	Directly from CEFR
Text type	DIALANG
Domain	Directly from CEFR
Topic	Directly from CEFR
Cognitive and linguistic complexity	
Abstraction level	Adapted on the basis of CEFR
Vocabulary	Adapted on the basis of CEFR
Grammatical complexity	Adapted on the basis of CEFR
Text length	Adapted on the basis of CEFR
CEFR-level estimate	Directly from CEFR
Reading item dimensions	
Question type	Adapted on the basis of CEFR
Operations	Adapted on the basis of CEFR
CEFR-level estimate	Directly from CEFR

In the linking study the Dutch Grid has been used to provide a first description of the items of the examinations of reading comprehension. This description has also been used for the selection of the examination items for the standardisation (see chapter 4).

The analyses are based on the state examinations of reading comprehension in German and English from vmbo-bb up to and including vwo of the years 2003 and 2004. For each of the two languages the analyses have been carried out by 2 persons. One person did the analyses for the three vmbo-examinations, the other person for the havo- and vwo- examinations. In the manual of the Dutch Grid it is recommended to have each text and item analysed by more than one person and to discuss the findings. Considering the great number of texts and items (see Table 3-3) this plan could not be carried out for financial and organisational constraints. Nonetheless we consider the results to be sufficiently valid to report them here. These results can be considered as indicative. Clear trends can be recognized and the information is useful for further construction of the examinations.

Table 3-3 *Number of texts and items analysed*

Examination	Number of texts		Number of items	
	German	English	German	English
vmbo-bb	23	30	72	69
vmbo-kb	30	29	84	79
vmbo-gl/tl	28	14	85	37
havo	20	20	85	82
vwo	17	21	85	89
total	118	114	411	356
grand total	232		767	

3.3.3 The linguistic and cognitive complexity of texts

In this section an overview is given of the linguistic and cognitive complexity of the examination texts on the basis of the degree of abstraction, the vocabulary, the grammar, and the length of the texts.

Degree of abstraction of texts

The descriptions of the CEFR-levels show that the degree of abstraction of texts increases from level A1 to level C2. From the results of the standard-setting it has been concluded that the difficulty levels of the state examinations of reading comprehension in German and English increase from A2 to B2/C1. If the examinations are to reflect this structure, then such an increase in abstraction should appear in the examinations. Figures 3-10 and 3-11 give an indication of the degree of abstraction of the texts in the state examinations of reading comprehension in German and English.

Figure 3-10 *Degree of abstraction of the examination texts German in percentages*

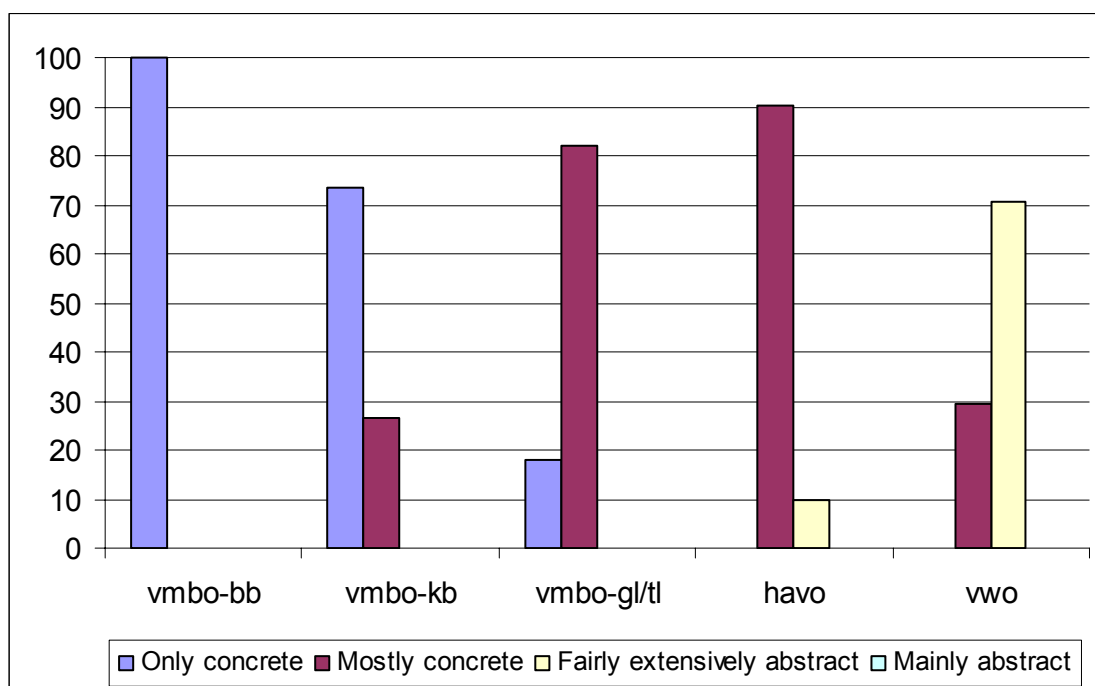
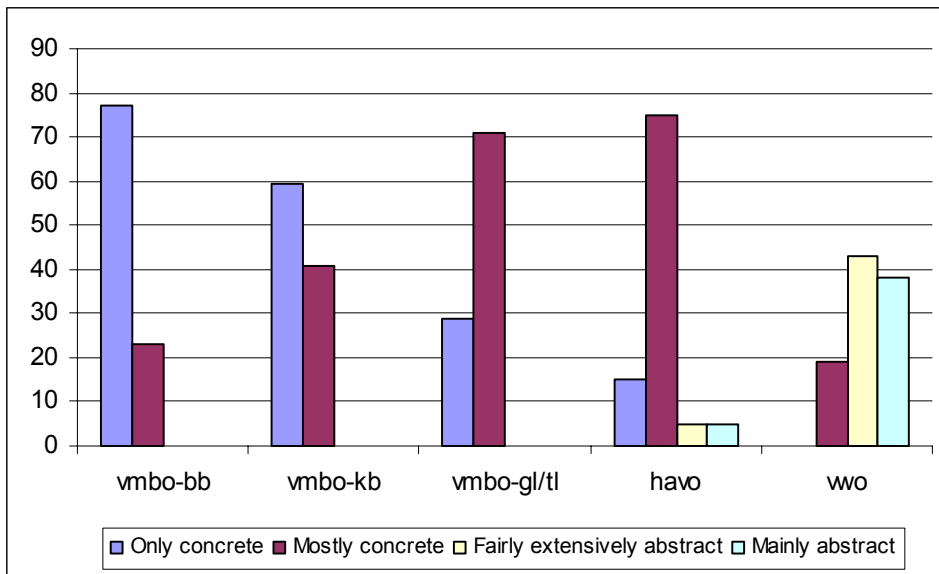


Figure 3-11 Degree of abstraction of the examination texts for English in percentages



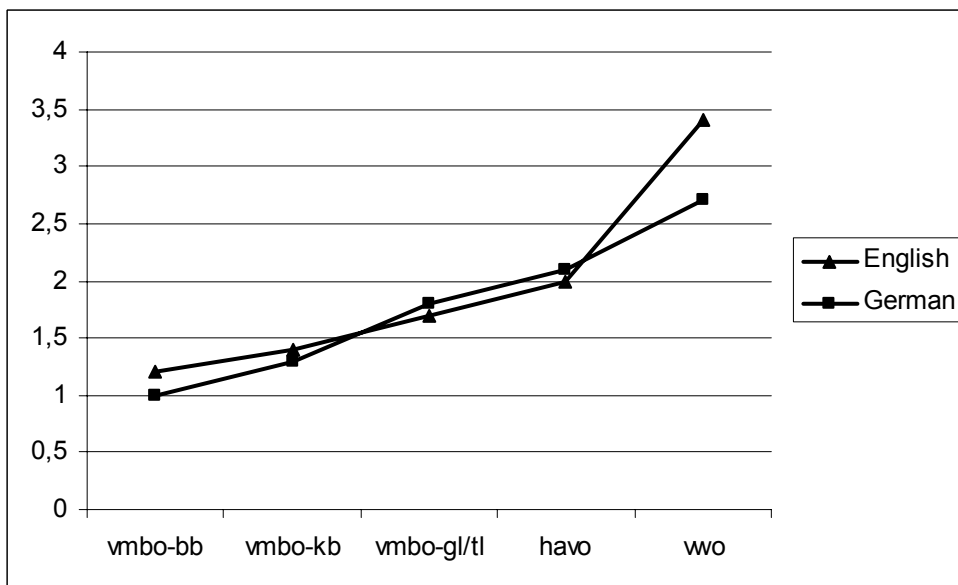
From vmbo to vwo, the percentage of texts with a concrete content decreases, and there is an increase in texts with a more abstract content.

When we assign the following values to the degree of abstraction, we can compute an abstraction score for each school type:

- Only concrete 1
- Mostly concrete 2
- Fairly extensive abstract 3
- Mainly abstract 4

Figure 3-12 shows these abstraction scores for the school types.

Figure 3-12 Abstraction scores of the examination texts reading comprehension English and German



The examinations in English and German appear to reflect the development described in the CEFR. The abstraction of the examination texts increases from vmbo to vwo.

Vocabulary

According to the CEFR, reading texts are characterized by a more extensive vocabulary as the CEFR-level increases. If the examinations are to reflect this increase, this tendency should be visible in the examinations from vmbo to vwo. Figures 3-13 en 3-14 show the nature of the vocabulary in the examination texts.

Figure 3-13 *Vocabulary of the examination texts German in percentages*

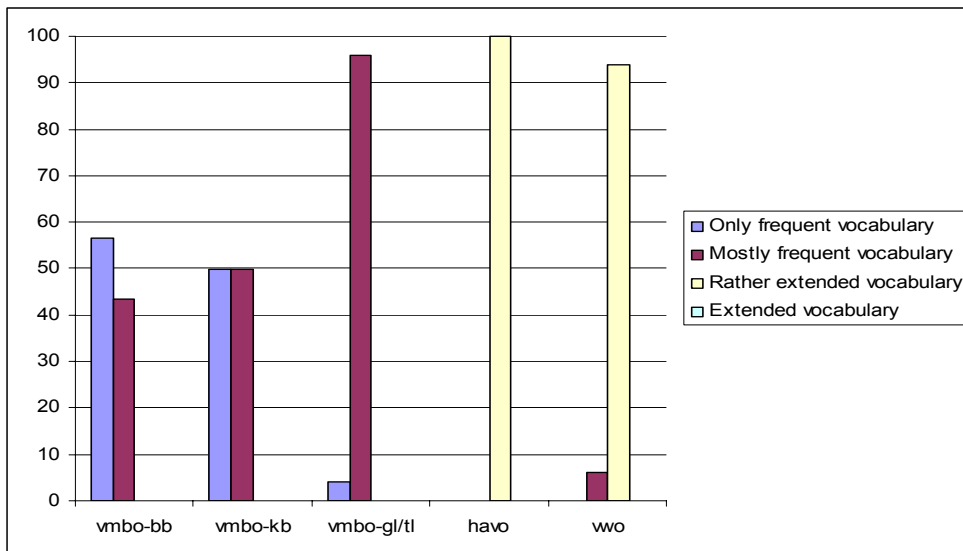
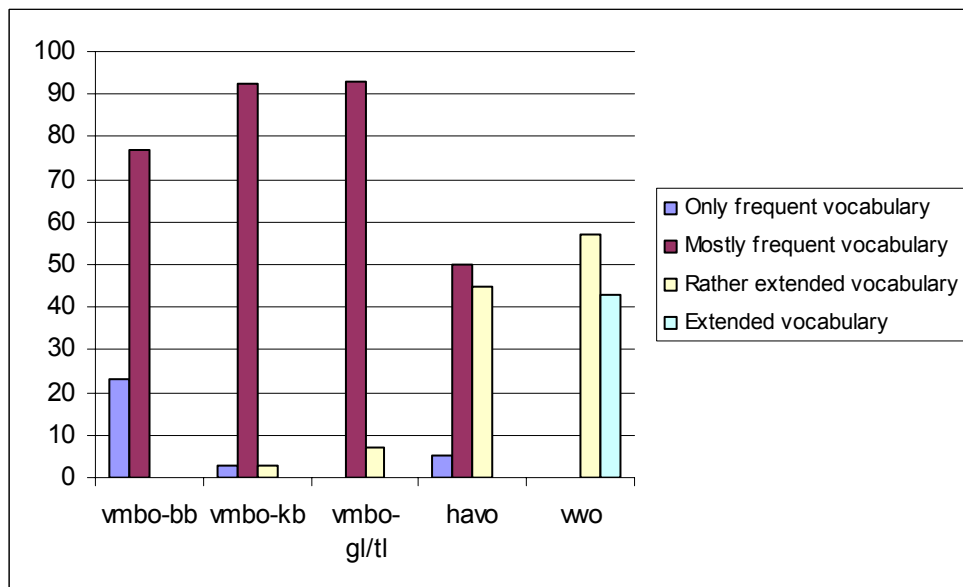


Figure 3-14 *Vocabulary of the examination texts English in percentages*



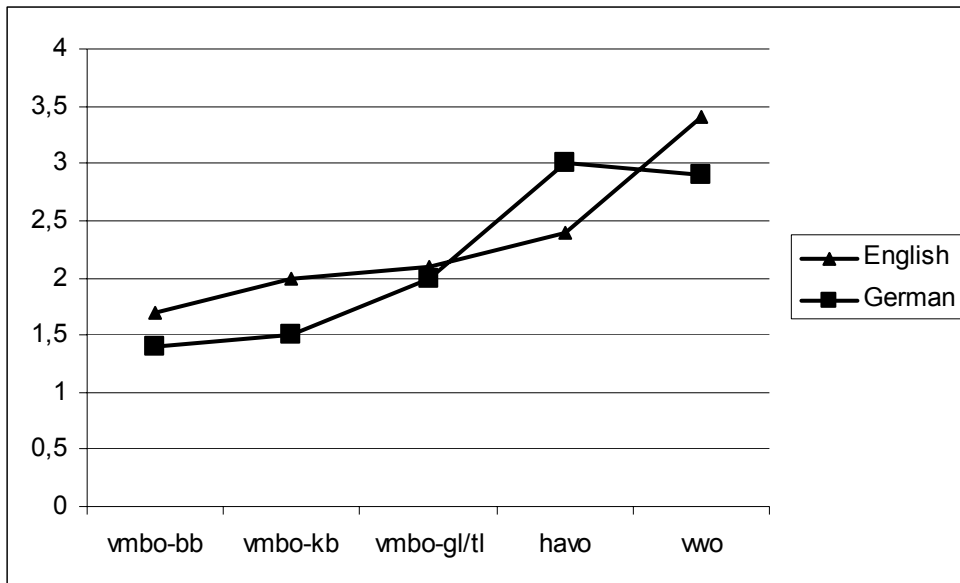
We see that the expected tendency described in the CEFR is reflected in the examinations. The examinations contain an increasing use of less frequent vocabulary from vmbo to vwo.

When we assign the following values to the vocabulary used we can compute a vocabulary score for each school type.

- Only frequent vocabulary 1
- Mostly frequent vocabulary 2
- Rather extended vocabulary 3
- Extended vocabulary 4

Figure 3-15 shows these vocabulary scores for the school types.

Figure 3-15 *Vocabulary scores of examination texts reading English and German*



The examinations appear to reflect the CEFR. Vmbo-examinations tend to contain frequently used words, whereas the vocabulary in havo and vwo ranges from rather extended to extended. For German the difference between havo and vwo is less prominent.

Grammar

The CEFR presupposes an increase in grammatical complexity of the reading texts as the CEFR-level increases.

Figures 3-16 and 3-17 show the grammatical complexity of the reading texts in the state examinations.

Figure 3-16 Grammatical complexity of the examination texts for German in percentages

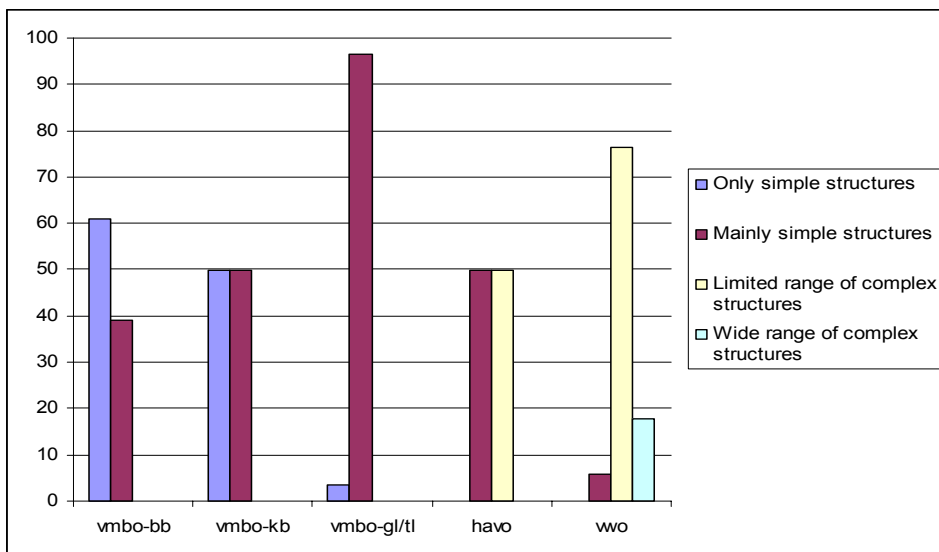
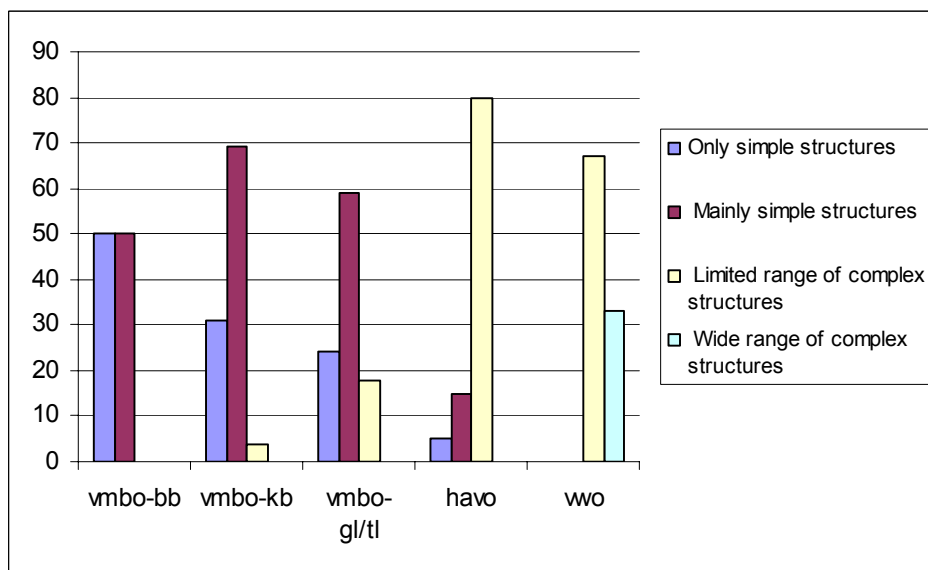


Figure 3-17 Grammatical complexity of the examination texts for English in percentages

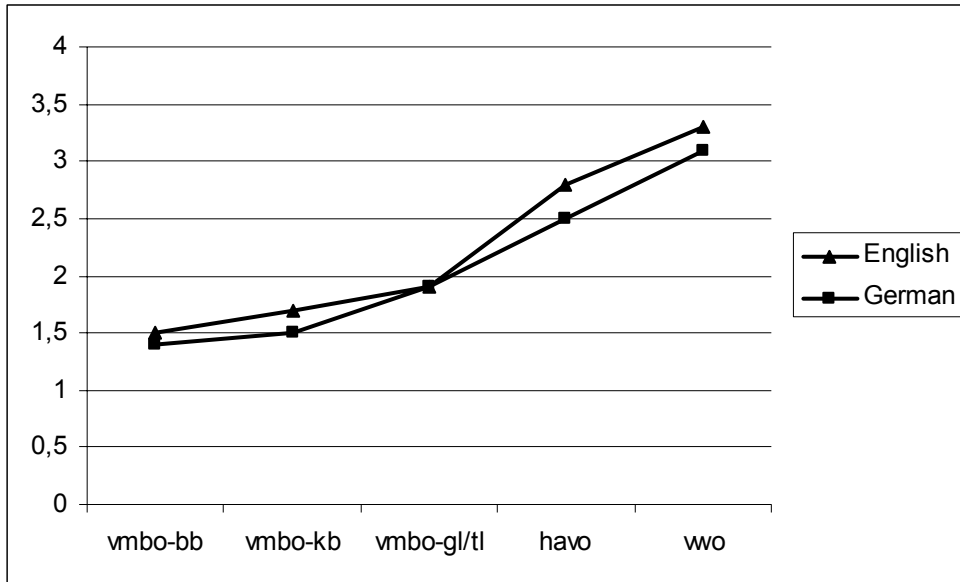


It can be seen that the expected tendency in the CEFR is also found in the examinations. The examination texts contain an increasing complexity of grammatical structures from vmbo to wvo. When we assign the following values to the grammatical structure, we can compute a grammar score for the examinations of each school type.

- Only simple structures 1
- Mainly simple structures 2
- Limited range of complex structures 3
- Wide range of complex structures 4

Figure 3-18 shows these grammar scores for the state examinations of reading comprehension English and German of the different school types.

Figure 3-18 Grammar scores of examination texts reading English and German



The examinations appear to reflect the CEFR. Vmbo-examinations tend to contain simple grammatical structures, vwo-texts tend to contain more complex structures, while the havo-texts assume a middle position.

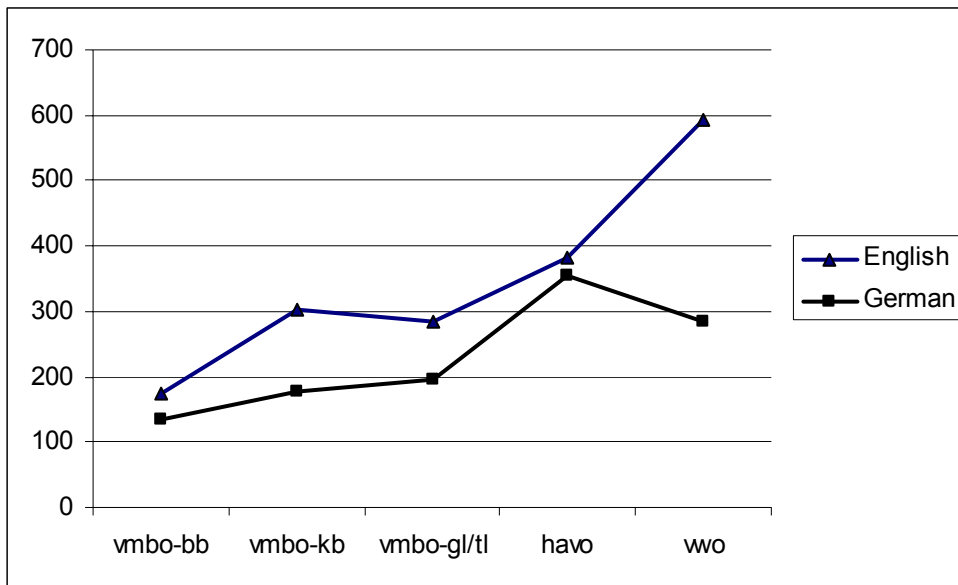
Text length

Another dimension in the CEFR, which is closely related to an increasing reading skill, is the length of texts that a candidate should be able to understand. Tables 3-4 and figure 3-19 show the average text lengths of the examination texts.

Table 3-4 Mean text length in number of words per school type

	vmbo-bb		vmbo-kb		vmbo-gl/tl		havo		vwo	
	English	German	English	German	English	German	English	German	English.	German
Mean	174	133	304	177	285	197	381	354	593	284
sd	130	73	221.7	127	246.5	155	237.4	169	254.8	168
Number of texts	30	22	29	15	14	27	20	20	21	17

Figure 3-19 Mean text length in number of words per school type



The examination texts in the state examinations of reading comprehension become longer from vmbo to vwo. Vwo and havo tend to have longer texts than the vmbo-examinations. It is remarkable that for German the vwo-examinations tend to contain somewhat shorter texts than the havo-examinations.

3.3.4 Task complexity

It appears from the last section that the increase of linguistic and cognitive complexity in the scales for communicative competence of the CEFR in the state examinations of reading comprehension are being reflected by the cognitive and linguistic characteristics of the examination texts. However, not just the texts, but also the tasks that readers have to perform with a text should show an increase in complexity according to the CEFR and *Taalprofielen*. The Dutch Grid also provides a descriptive framework for task complexity which is based on the CEFR. The reading items can be described within that framework according to the dimensions of question type and operations.

Question types

The Dutch Grid identifies the following question types. The CEFR does not describe a relation between levels and question types.

Answer type	Question type
Selected response	Multiple choice question
	True-False question
	Combination question
	Sequencing question
	Citing
Short structured response	Short answer question
	Cloze question
	Gap-filling question
	Complete a question
	Complete a summary
Extended constructed response	Essay
	Summary
	Justify in own words
	Other

Figures 3-20 and 3-21 show the distribution in percentages of question types in the state examinations of reading comprehensions German and English.

Figure 3-20 Question types in the examinations of reading of German in percentages

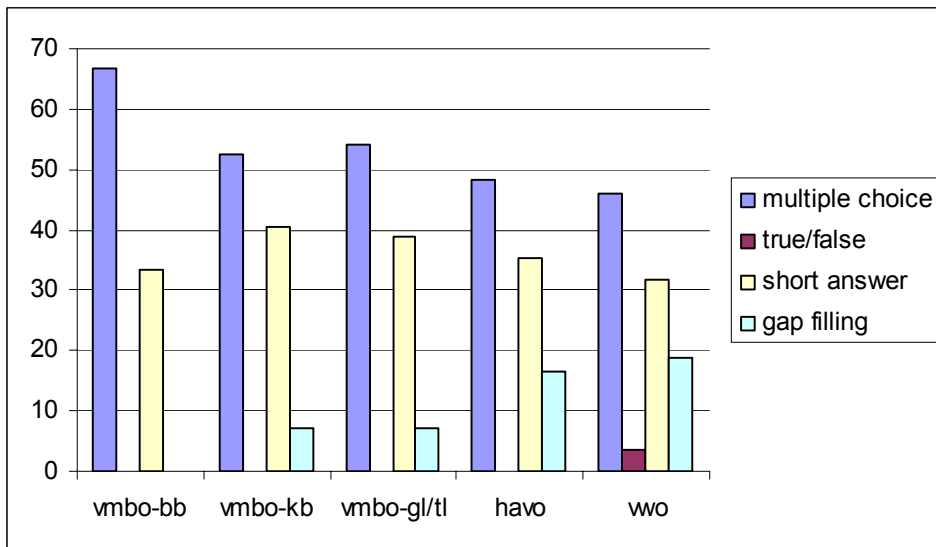
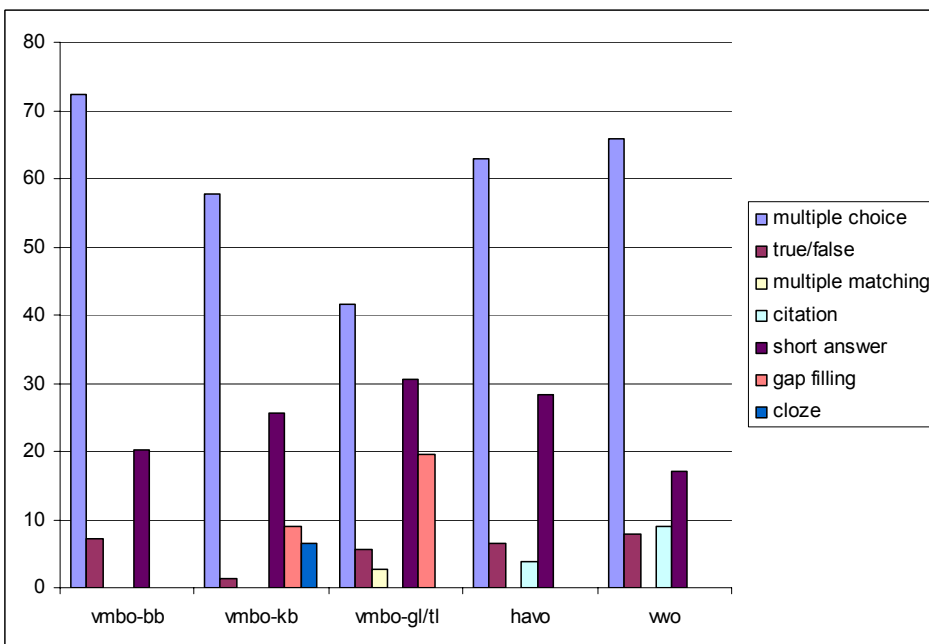


Figure 3-21 Question types in the examinations of reading of English in percentages



In the examinations a variation in closed and open short answers can be seen. There is no clear relation between school type and question type. However, no such relationship is presupposed in the CEFR.

Operations

The CEFR presupposes that a reader at higher CEFR-levels should be able to perform a more extensive repertoire of reading operations.

The descriptive framework of the Dutch Grid for the reading operations contains three dimensions:

- The task dimension identifies three types of operations: recognize, making inferences and evaluating
- The explicitness dimension describes whether the information that is asked for is to be found in the text explicitly or implicitly
- The content dimension describes what it is that is being asked for in the question.

Below these dimensions as distinguished in the Dutch Grid are illustrated. The three dimensions are independent of each other and can occur in all combinations.

Task dimension	Explicitness dimension	Content dimension
Recognise and retrieve	Explicit	Main ideas; gist /broad outlines Details Opinion Attitude of the author
Make inferences	Implicit	Conclusions Communicative objective Text structure/ relations between text parts
Evaluate		

In this section the items in the state examinations of reading comprehension are described in the light of these task dimensions.

Task dimensions

From the CEFR scales it can be seen that, as the CEFR-level increases, there is an increase in reading items aimed at making inferences and evaluations on the basis of the text. At the same time there is a decrease in items which aim at measuring direct recognition of information. Figures 3-22 and 3-23 show an overview of task dimensions in examination tasks.

Figure 3-22 *Task dimensions of the reading items for German in percentages*

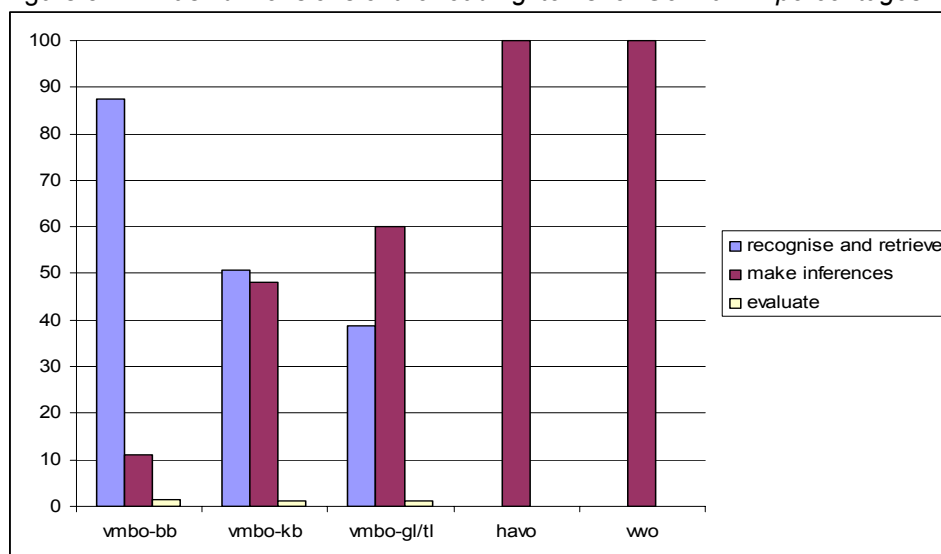
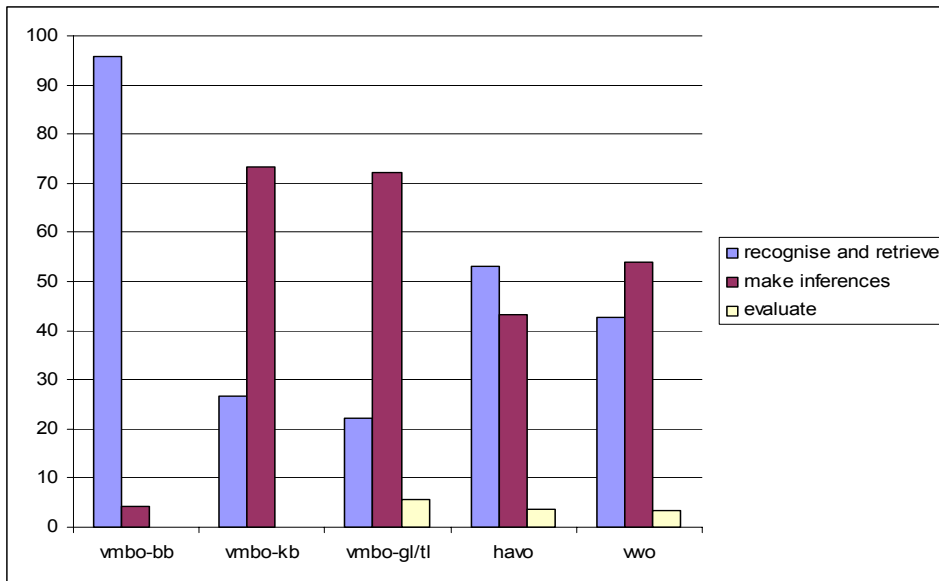


Figure 3-23 Task dimensions of the reading items for English in percentages



The tendency as described in the CEFR is clearly reflected in the examinations for German. For the vmbo-examinations more recognition is asked for, whereas the items in the havo- and wvo-examinations tap the making of inferences. For English the pattern is more diffuse.

The explicitness dimension

Figures 3-24 and 3-25 show to what extent reading items refer to explicit or implicit information.

Figure 3-24 Nature of the information that the items German refer to in percentages

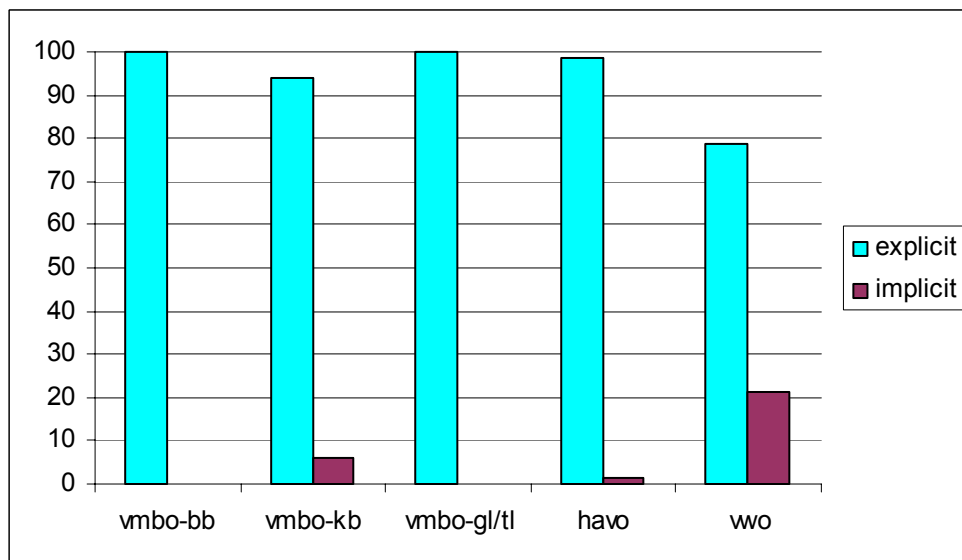
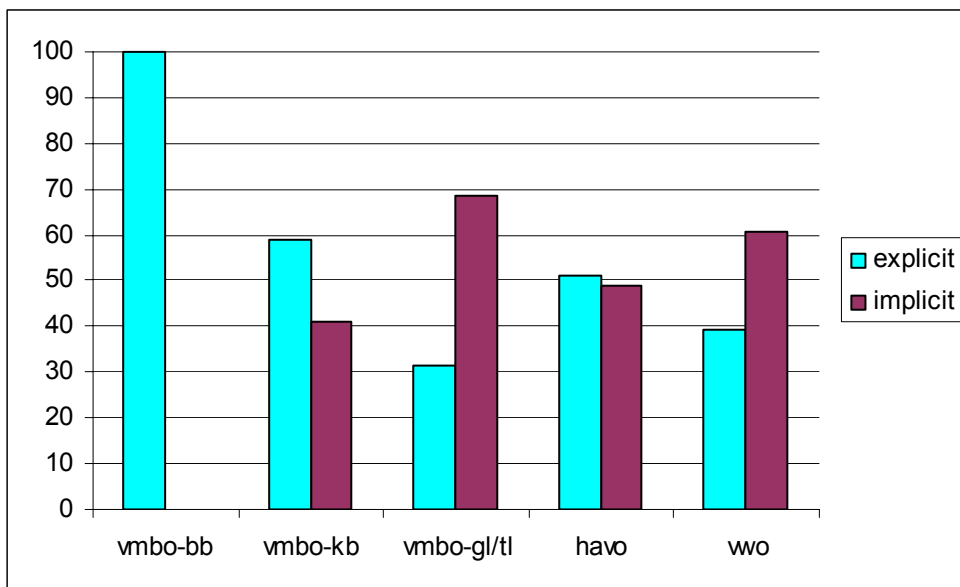


Figure 3-25 Nature of the information that the items English refer to in percentages



For all the examinations in German the information is mostly or exclusively found mentioned explicitly in the text. For English the information that is asked for is mostly implicit, with the exception of vmbo-bb.

Content dimension

The content to which reading items refer to is described for the various levels in the CEFR. However, they are not consistent and a clear relation with the levels cannot be established beforehand. An implicit relation is present, because in the ‘can-do statements’ this content description is very often (but not always) linked to terms that indicate the simplicity or the complexity of the texts to be read. The only clear, though implicit, presupposition in the CEFR is that as the level increases, the variety increases of what someone should be able to understand from a text; also, the texts will become more complex, both linguistically and cognitively.

Figures 3-26 and 3-27 show the content of the operations in the reading items for the various school types.

Figure 3-26 Content of the operations to which the items German refer in percentages

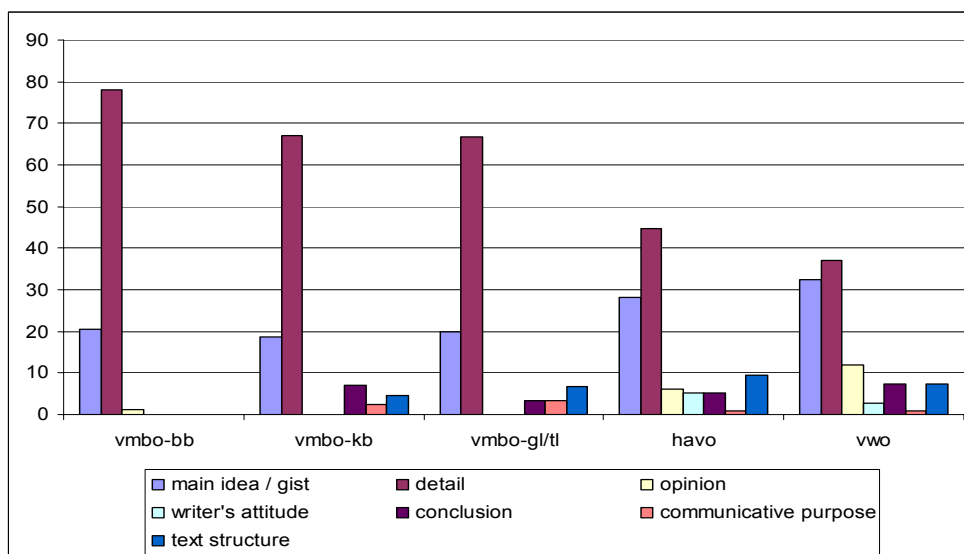
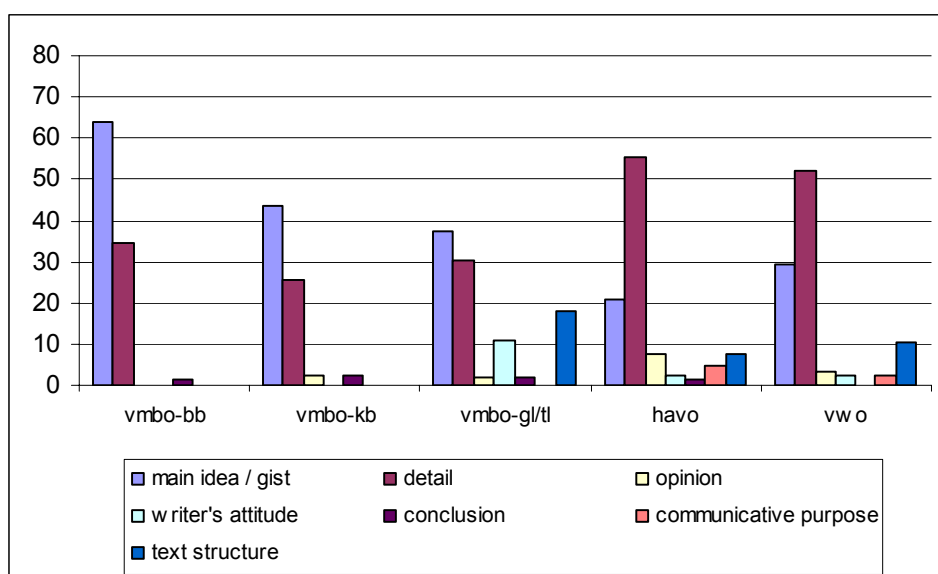


Figure 3-27 Content of the operations to which the items English refer in percentages



In both German and English the examinations show an increase in the diversity of the content of the operations from vmbo to vwo, which corresponds to the implicit purpose of the level descriptions in the CEFR. Also worth noting is the decrease from vmbo-bb to vwo in the examinations of German as to the percentage of questions about details and the slight increase in questions about main ideas. This trend seems to go the other way in the examinations in English.

3.3.5 Conclusions texts and tasks

The standard-setting for the state examinations of reading comprehension in the foreign languages shows that from vmbo-bb to vwo the state examinations of reading comprehension refer to increasingly higher CEFR-levels (see chapter 4).

The CEFR presupposes that higher levels will show an increase of the linguistic and cognitive complexity of reading texts which language learners should be able to understand. This presupposition also goes for the reading tasks which they have to be able to perform. This development is described specifically in the scales for communicative competences.

The content description carried out for the examinations of reading comprehension in German and English shows that this increase of linguistic and cognitive complexity in the examinations is specifically found at text level. The texts become more complex grammatically; the abstractness of the texts increases and the vocabulary becomes increasingly more extensive and more varied. The descriptions of the items indicate that as the level of the examinations is higher, examinees should be able to perform a wider variety of operations.

4 Standardisation

4.1 Introduction

In this chapter the method and the results of the standardisation procedure are reported upon. Section 4.2 contains a description of the judgement process and section 4.3 reports on the data analyses carried out to determine the quality of the standardisation procedure. The results of the standard-setting for English, French and German are presented in section 4.4.

The Manual emphasizes the necessity of reaching a *common understanding* of the meaning of CEFR levels, notably by language professionals in other sectors, regions and countries (in this project: others than the judges at Cito and SLO), before starting the standardisation process. The Manual distinguishes four stages in this phase:

1. Familiarisation: a process that is similar to the one described in phase 1. No further activities have been carried out in this area, except for the process of standard-setting (see also below at point 4)
2. Training with standardized samples of performance for *productive* skills. This has been outside of the scope of this project that has been aimed at linking examinations of *reading comprehension*.
3. Benchmarking of performances. Again, this has been outside of the scope of this project.
4. Standard-setting. The Manual distinguishes two main phases in this process:
 - The judgement process
 - Data analysis for validation of the standards.

4.2 Judgement process

A short overview of the various steps in this process is given below.

- *Definition of goals for the decision procedure*

The aim of the judgement process is that judges determine the minimum CEFR-level needed by a candidate to successfully perform on a given language test. In other words, to determine cut-off scores for each examination at which a candidate can be said to have acquired a CEFR-level that is relevant to the aim of the specific examination.

The standard-setting algorithm that was used is described here briefly. The data are collected by the so-called basket procedure. A judge is asked to put each item into a labelled basket corresponding to the minimum CEFR-level that is needed to carry out the task in the item. There are five baskets, called A1, A2, B1, B2 and C1+, corresponding to the levels that the examination syllabuses aim at (and beyond). C1+ refers to the levels C1 and C2. If an item is placed in basket B1, this means that according to the judge, a person at level B1 should be able to carry out the task correctly and by implication mastery is assumed at all higher levels (persons at levels B2 and higher). It cannot be expected, however, that a person at level A2 (or lower) will be able to carry out the task correctly. This method of standard-setting has been developed for the project DIALANG (Dialang, 2002). The method was used because it has been shown to be manageable and to yield reliable and useful results. Moreover, several members of the linking project were familiar with this method.
- *Selection of reading items*

Ideally all the items in the state examinations of reading comprehension in French, German and English under review should be judged during the standard-setting. However, this would have meant that for each language judges would have had to rate a total of circa 250 reading items (to go with over 50 texts). It is clear that this would have been too strenuous a task to perform during one session.

It was therefore decided to create representative samples of the five examinations per language. On average examinations contain 45-50 items of which one-third or circa 15 items per examination were selected for the standard-setting procedure. For the sampling the test matrices of the examinations were used.

Sampling criteria have been:

- Type of text (descriptive, argumentative etc);
- Text length;
- Type of reading item (multiple choice, open-ended);
- Behaviour/Operations (reproduction, inference, prediction);
- Content questions (detail, gist).

The reading items thus selected were then put in a random order so that judges would have to decide for each individual item what minimum level was required to answer the question correctly. As a matter of fact, the *texts* (rather than the items) accompanied by one or more items were put in a random order as it would not have made sense to reproduce the same text a number of times to go with each item accompanying that text. Also, we had to present the items coupled with a text in the (logical) order in which they were included in the examination.

In this way 77 items for English, 79 items for German and 81 items for French were selected.

- *Selection of judges*

In order to raise the validity of the standard-setting judges were recruited from the teaching profession at secondary schools (at which the students are trained to take the foreign language examinations) and at institutes for higher and university education where such teachers are trained. These lecturers from higher education also train teachers for primary schools, which is relevant in the case of English, because in primary schools English will be taught at the lowest level (if A1). Other judges were recruited from the business world and from private language institutes. A last group of judges included a politician and members of the State Examination Committees (CEVO).

There were a number of reasons why project members have been excluded from the standard-setting. Some project members had been responsible for constructing the items to be judged. Although it would have been useful to know how they would relate items to CEFR-levels, their familiarity with the items might yield biased results.

- *Training of judges*

Judges were trained in much the same way as in the Familiarisation phase of the project, to get familiar with CEFR categories and levels. This training was given by the same person who had trained project members. A lively discussion on the relevance of the CEFR for various purposes developed. The judges were then given a number of texts with items to judge following the basket procedure described above. For each item in the training session the following question was put to the judges:

- Please indicate for each item which level (A1, A2, B1, B2 or C1+) is minimally required to carry out the task correctly. (*Circle for each item the number in the column with the answer of your choice*).

Text	Tasks	Level				
		A1	A2	B1	B2	C1+
1	1	1	2	3	4	5

The training of the judges at this stage took place in three separate language groups (French, German and English) and was led by project members. During the discussions in the language groups the judges reached agreement on the required minimum level for each of the four items selected for training. These items were not included in the actual standard-setting.

- *The judgment sessions*

After the training the judges were given sets of texts and items in the randomised order described above. Judges took between two and three hours to rate all the items using a rating form as described above. There have been no complaints of this task having been too strenuous. As a

matter of fact, many of the judges expressed their willingness to take part in future standard-setting sessions for listening, speaking and writing tasks.

- *Data collection procedures*

The rating forms have been collected and the data have been transferred to optical reading forms. Data collected included: rater ID, language and ratings (1 to 5, corresponding to A1 to C1+) per item.

4.3 Data analysis for validation of the standards

The next phase in the standard-setting procedure has been the data analysis to validate the accuracy of the standards. The data analysis comprises two operations:

1. *Determining rater agreement*
2. *Determining minimum scores for relevant CEFR-levels on each examination*

Determining rater agreement

Rater agreement and rater reliability have been computed. Then results are given for each language for the total number of items that have been rated. Thereafter items have been regrouped and rater reliability (Cronbach's α), rater agreement and the mean required minimum level (1=A1, 2=A2, 3=B1, 4=B2) across judges and items for each examination are shown.

French reading items

Number of judges: 10

All items as judged in random order

Estimated variance components

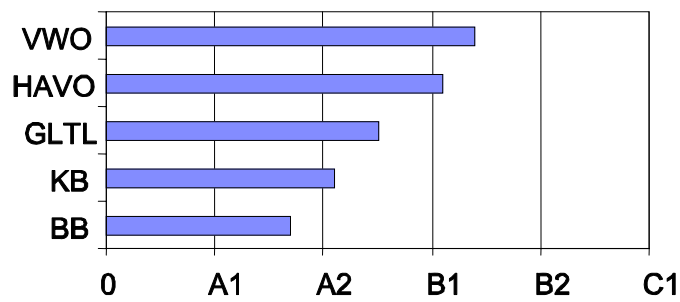
Reading items (p):	0.5093	56%
Judges (b):	0.0679	7%
Residue (pb,e):	0.3281	36%
Rater reliability (α):	0.9395	
Rater agreement (Rho2):	0.9279	

Table 4.1 *Rater reliability, rater agreement and required minimum level per examination (French)*

Examination	N reading items	Rater reliability (α)	Rater agreement (Rho2)	Mean required minimum level
bb	15	.66	.54	1,7
kb	14	.90	.87	2,1
gl/tl	17	.78	.70	2,5
havo	19	.80	.73	3,1
vwo	15	.71	.64	3,4

The values that are given in the last column (mean required minimum level) correspond with the CEFR levels 1=A1, 2=A2, 3=B1, 4=B2, 5=C1. The mean required minimum level is also shown in the next figure.

Figure 4.1 *Required minimum level per examination (French)*



German reading items

N judges: 12

All items as judged in random order

Estimated variance components

Reading items (p): 0.4894 58%

Judges (b): 0.0876 10%

Residue (pb,e): 0.2659 32%

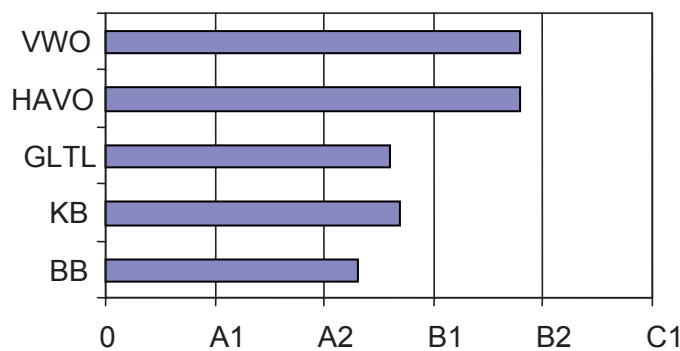
Rater reliability (α): 0.9567

Rater agreement (Rho2): 0.9432

Table 4.2 *Rater reliability, rater agreement and required minimum level per examination (German)*

Examination	N items	Rater reliability (α)	Rater agreement (Rho2)	Mean required minimum level
bb	16	.89	.80	2,3
kb	14	.42	.34	2,7
gl/tl	15	.85	.83	2,6
havo	17	.79	.72	3,8
vwo	16	.88	.83	3,8

Figure 4.2 *Required minimum level per examination (German)*



English reading items

Number of judges: 15

All items as judged in random order

Estimated variance components

Reading items (p): 0.9108 70%

Judges (b): 0.0384 3%

Residue (pb,e): 0.3429 27%

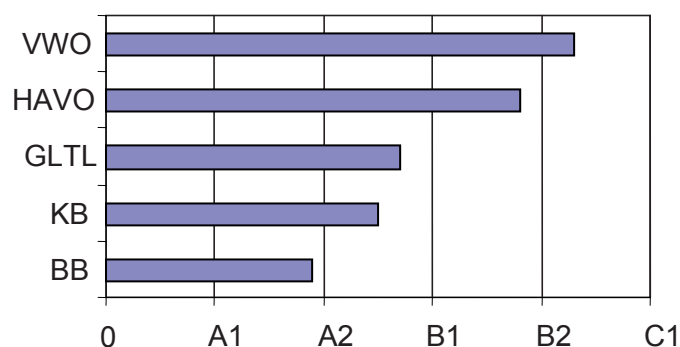
Rater reliability (α): 0.9755

Rater agreement (Rho2): 0.9728

Table 4.3 Rater reliability, rater agreement and required minimum level per examination (English)

Examination	N items	Rater reliability (α)	Rater agreement (Rho2)	Mean required minimum level
bb	16	.78	.74	1,9
kb	13	.92	.90	2,5
gl/tl	12	.79	.78	2,7
havo	18	.74	.67	3,8
vwo	17	.74	.72	4,3

Figure 4.3 Required minimum level per examination (English)



Discussion analysis of rater agreement

- Rater agreement for all the items presented to judges was over .90 for all three languages. This would indicate that judges agree sufficiently on the minimum CEFR-level required for each item to be mastered.
- Judges have placed the items taken from the lowest level examination (bb) at the lower end of the CEFR scales and they have placed items taken from the higher level examinations (havo and vwo) at the higher end of the CEFR-scale. This difference in level corresponds with the range in the Dutch examination levels where bb is the lowest level and vwo is the highest.
- External judges seem to agree with the item writers on the level of difficulty of sets of items.
- Data analysis shows that for French and English, judges are of the opinion that a higher CEFR-level is needed to be able to successfully answer the questions if moving up through the examinations from low level to high level.
- For German, judgements suggest that CEFR-levels required for mastery of the items at a higher level are in a less consistent order of increasing difficulty.

4.4 Determining minimum scores for relevant CEFR-levels in each examination.

The next step in the data analysis phase has been the determining of minimum scores on a state examination needed by a student to be able to claim that he or she is at a relevant CEFR level. Also, we would like to know what the actual cut-off score as determined by the State Examination Committee (CEVO) would mean in terms of mastery of CEFR-levels.

For this purpose we have extrapolated the data found in the “basket” procedure to the actual examinations taken by students in 2004.

The method that has been followed to link the above-mentioned scores and cut-off scores to CEFR-levels is being illustrated below for English. After that the most important conclusions for French and German will follow.

4.4.1 Results standard-setting English

In the next sections an overview of the findings with the state examinations reading English will follow: the scores that go with the cut-off scores sufficient/insufficient as set by CEVO and the scores that go with the relevant cut-off scores with CEFR-levels. We will indicate how the judgements of CEVO of what is a sufficient performance match with the judgements of relevant CEFR-levels.

In table 4.4 an overview is given of the number of examinees whose responses have been analysed as part of quality control procedures for the examinations and to determine cut-off scores and reliability estimates (Cronbach’s alpha) with unweighted and weighted scores. Weights are used in Item Response Theory (IRT) analyses that are at the basis of standard-setting procedures.

Table 4.4 *Examinations in English: number of respondents, reliability estimates with unweighted (unw.) and weighted (w.) scores.*

Level	N	A (unw.)	α (w.)
vwo	1858	0.823	0.838
havo	2036	0.782	0.807
gl/tl	2683	0.744	0.783
kb	2131	0.834	0.849
bb	2250	0.841	0.855

The basic counting in the standard-setting procedures has been the average (across judges) cumulative number of reading items that has been put in each basket, starting with the lowest basket A1. A fictitious example is given in table 4.5. From the table, we know that on average 3.2 items were put in basket B1 and 1.9 items (on average) were put in a lower level basket. So (on average) 5.1 items should be mastered at level B1. In the standard-setting this number is interpreted as a *minimum requirement for B1*. And therefore the cut-off score A2/B1 is positioned at 5/6.

Table 4.5 *An example of the outcomes of the standard-setting procedure*

Level	Frequency	Cumulative frequency
A1	0.8	0.8
A2	1.1	1.9
B1	3.2	5.1
B2	8.0	13.1
C1	1.9	15

When items are calibrated using Item Response Theory (IRT), different items may be given different weights. When these weights are taken into account one might replace the column frequency in table 4.5 by a column with the average weight of the items put into each basket and then cumulate these weights across levels. These cumulative weights can then be interpreted as the minimum weighted score on a test consisting of all the items used in the standard setting. If the calibration is reliable, these cumulative weighted scores can be converted into a measure on the latent trait. An example is given in table 4.6.

Table 4.6 *Results for vwo Examination in English*

Level	Theta	Unweighted score	Weighted score
A1	-0.980	0.00	0.00
A2	-0.939	0.07	0.20
B1	-0.478	1.60	5.73
B2	-0.182	9.07	32.20
C1	1.277	15.00	54.00

A sample of 15 items taken from the vwo examination was rated by 15 judges. The maximum unweighted score therefore is 15 and the maximum weighted score (based on an analysis of the whole examination vwo) for these 15 items is 54. When we consider level B2, the cut-off score for unweighted scores at B1/B2 is 9/10; for weighted scores it is 32/33 and on the theta scale the cut-off is -0.182. For the A2/B1 cut-off point, the theta value is -0.478. This operation can be used to apply the standard setting to larger tests than the sample of items that has been presented to the judges.

We discuss two methods.

- (1) In the calibration, weighted scores can be transformed into theta-estimates. For the vwo examination in English we find the following results (see table 4.7). This suggests that the cut-off score for B1/B2 is 126/127. For A2/B1 we could choose either 28/29 or 29/30 (since we have results with three decimals). The disadvantage of this procedure is that it requires the use of weighted scores, which may be impractical in real applications: test scores of pupils are in fact reported in unweighted *total* scores

Table 4.7 *Correspondence weighted score – theta vwo examination in English*

Weighted score	Theta estimate
28	-0.490
29	-0.478
30	-0.468
...	...
126	0.181
127	0.189

- (2) We can, however, also estimate theta on the basis of unweighted scores. The results of applying this method are given in table 4.8. For B1/B2 the cut-off score is at 33/34, whereas for A2/B1 the cut-off score is at 8/9.

Table 4.8 Correspondence unweighted score – theta vwo Examination in English

Unweighted score	Theta estimate
8	-0.490
9	-0.478
10	-0.468
...	...
33	0.181
34	0.189

Some remarks on the minimum scores needed for a relevant CEFR-level on each examination

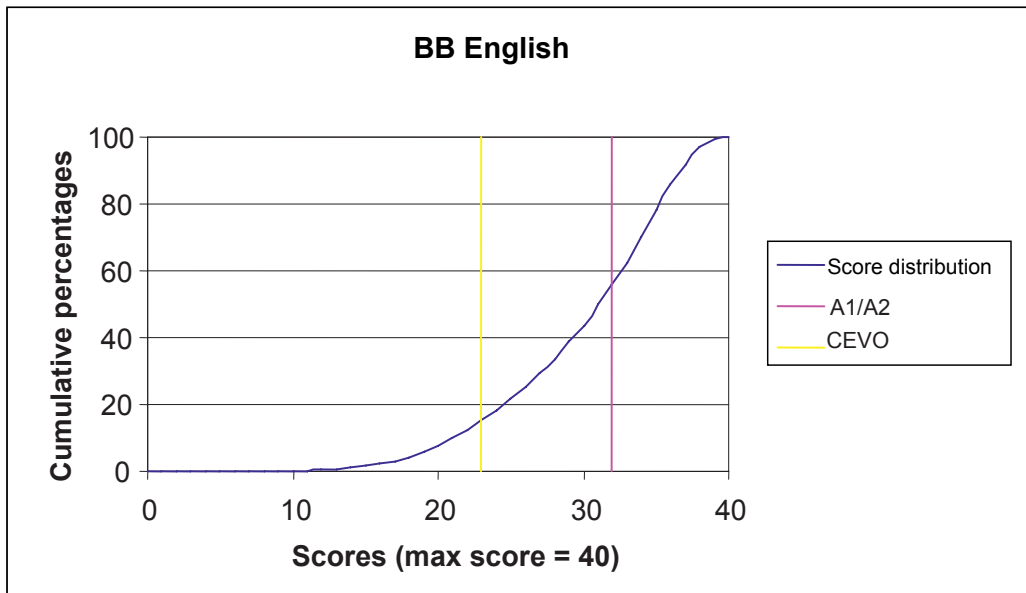
- In the preceding sections we have discussed the determination of cut-off scores for *relevant* levels. This issue may need some further elaboration. In the rater-agreement analyses for the vwo examination in English, to take an example, we saw that judges considered the minimum CEFR level needed to master the sample of items taken from the vwo examination was judged to be between B1 and B2. From other sources in the internal validation study we also know that the examination was *aimed at* a level that corresponds to B2. We therefore believe that the most relevant CEFR-score is between B1 and B2. It is at this score that we indicate what minimum score the candidate needs to prove that he functions at B2.
It is also possible to compute cut-off scores at A2/B1 and at B2/C1 for the vwo examination. However, in the case of a cut-off score A2/B2 (*does the candidate function at the B1 level?*) we have more suitable instruments at our disposal, namely examinations that are geared towards a lower level than the vwo examination. If we wish to determine if a candidate is at C1 level (at the cut-off score B2/C1), then we might be able to compute the cut-off score needed for C1 in the vwo examination. However, from a validity point of view this is a debatable procedure as judges have indicated that in the present vwo examination there are (considerably) fewer items at C1 level than at B2 level. We would run the risk of claiming *that a person functions at C1 level because he has mastered (nearly) all B2 items in a test, whereas in fact we should be showing that that person has mastered a sizable number of C1 items as well as some B2 items*). It is therefore important to choose the right instrument with sufficient items at a particular CEFR-level to be able to give a reliable estimate on CEFR-level mastery.
- In the present examination procedures in the Netherlands the State Examination Committee (CEVO) determines the cut-off score for each examination. When it is claimed that an examination is at a particular CEFR-level, we need to look at where the CEVO cut-off score is positioned. At this point in time it has not yet been possible for either students or schools to claim that they are at a higher level than the CEFR-level that corresponds to the CEVO cut-off score. If a candidate has managed to achieve a higher score than the scores that go with the cut-off score, one can of course conclude that his/her CEFR-level will be higher than the level that goes with the cut-off score.
- The judgements of on the difficulty of the items measured can be validated empirically by comparing judgements with the answers of the candidates to the items. A complicating issue is that the examination results are sometimes difficult to compare because the examinations do not contain common items. Empirical validation can show if differences in judgements of a random sample of items by expert judges may result in that the estimated required level of mastery for a particular CEFR-level in one examinations is different from the required mastery of the same level in another examination.

In the following figures we will illustrate where CEVO cut-off scores and relevant CEFR cut-off scores are to be found in the Dutch examinations for English. As can be seen, the difference between the cut-off point with the CEVO cut-off score (pass/fail) and the cut-off score that goes with the relevant CEFR-level of an examination, as concluded from the standard-setting, may be considerable. We also

give an indication of the score distribution (and consequently of the corresponding CEFR-levels) of the sample student population for each examination.

In figure 4.4 below, the distribution of scores and cut-off scores is given for the bb examination in English.

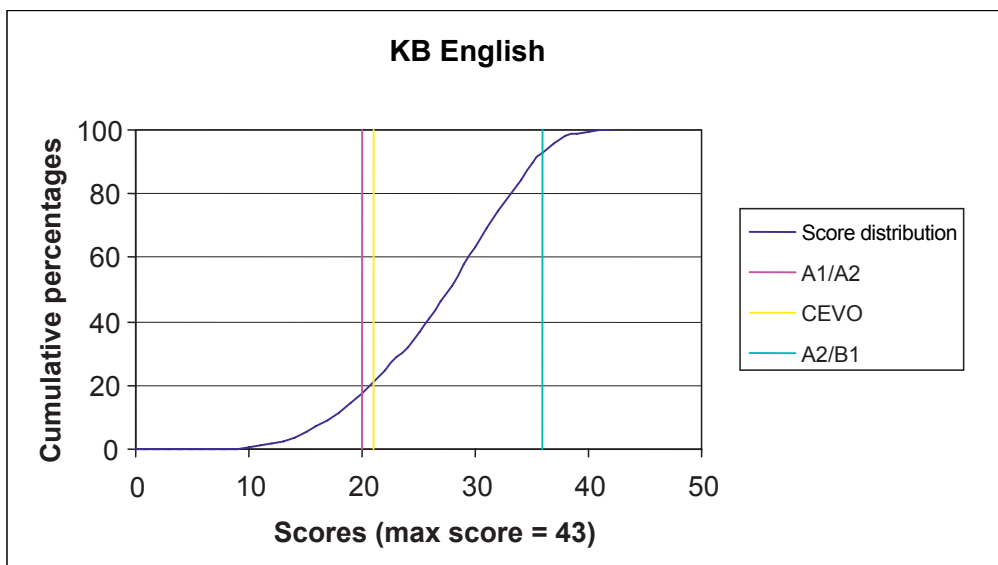
Figure 4.4 *Distribution of scores and cut-off scores bb examination in English*



We see that the relevant CEFR level cut-off score with the examination bb is between A1 and A2. That is to say that the scores to the right of the cut-off score line, so scores of 32 and higher, belong to the A2 level. Scores lower than 32 indicate a level of A1. We find/see that the CEVO cut-off score is lower than level A2. This means that a student can pass the bb examination without having reached the relevant CEFR A2 level.

In figure 4.5 below we see the distribution of scores and cut-off scores is given for the KB examination in English.

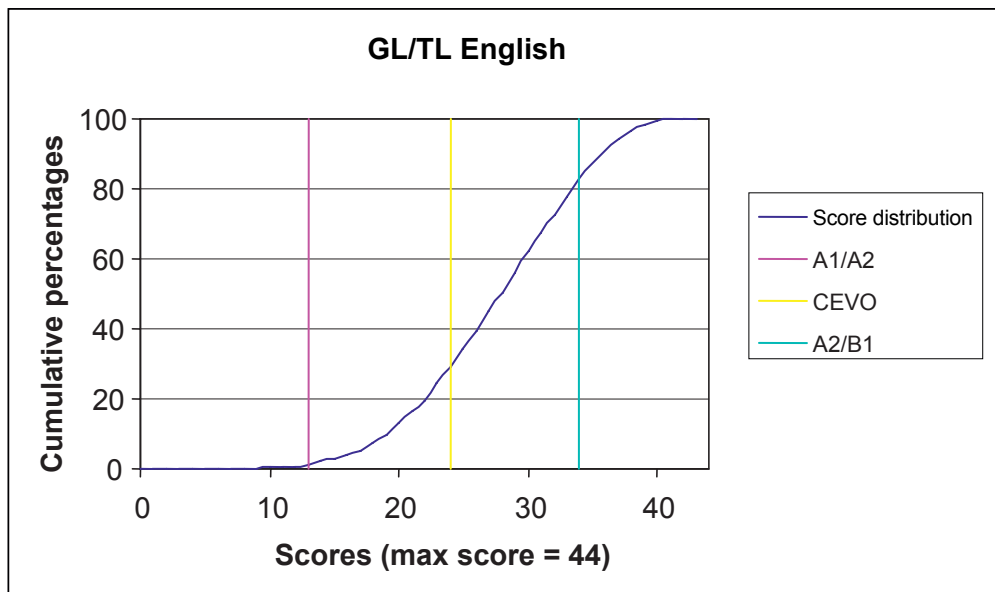
Figure 4.5 *Distribution of scores and cut-off scores KB examination in English*



We see that cut-off scores for two relevant CEFR-levels have been computed: A1/A2 and A2/B1. That is to say that the scores to the right of the cut-off score line A1/A2, so scores of 20 and higher, belong to the A2 level. Scores higher than 36 indicate a level of B1. We find that the CEVO cut-off score is just a little over the A1/A2 cut-off score. A small percentage of students reach B1 level for this examination.

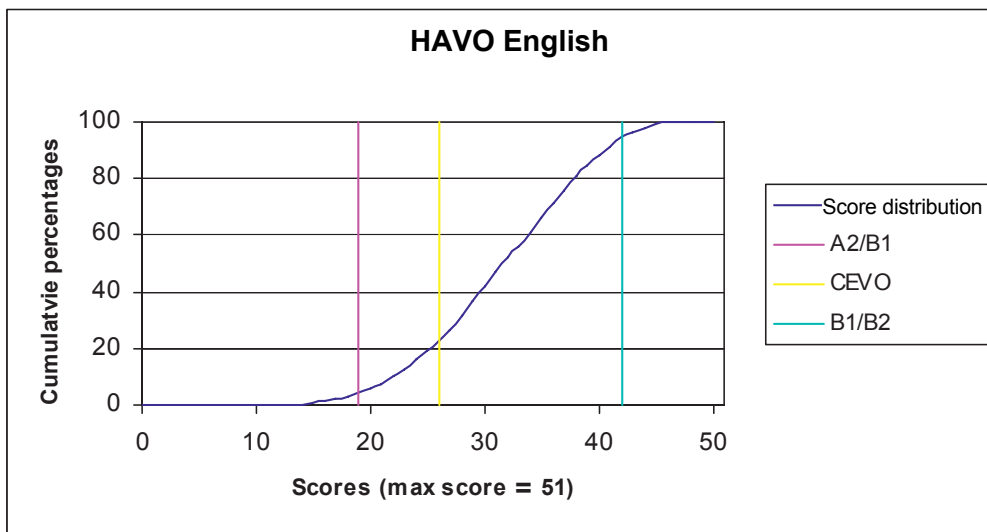
In figure 4.6 below we see the distribution of scores and cut-off scores is given for the gl/tl examination in English.

Figure 4.6 *Distribution of scores and cut-off scores gl/tl examination in English*



We see that two cut-off scores for relevant CEFR-levels have been computed: A1/A2 and A2/B1. That is to say that the scores to the right of the cut-off score line A1/A2, so scores of 24 and higher, belong to the A2 level. If candidates get scores higher than 34 then this indicates a level of B1. We find that the CEVO cut-off score is considerably higher than the A1/A2 cut-off score, but considerably lower than the A2/B1 cut-off score. A higher percentage of students than in the KB examination have reached B1 level for this examination.

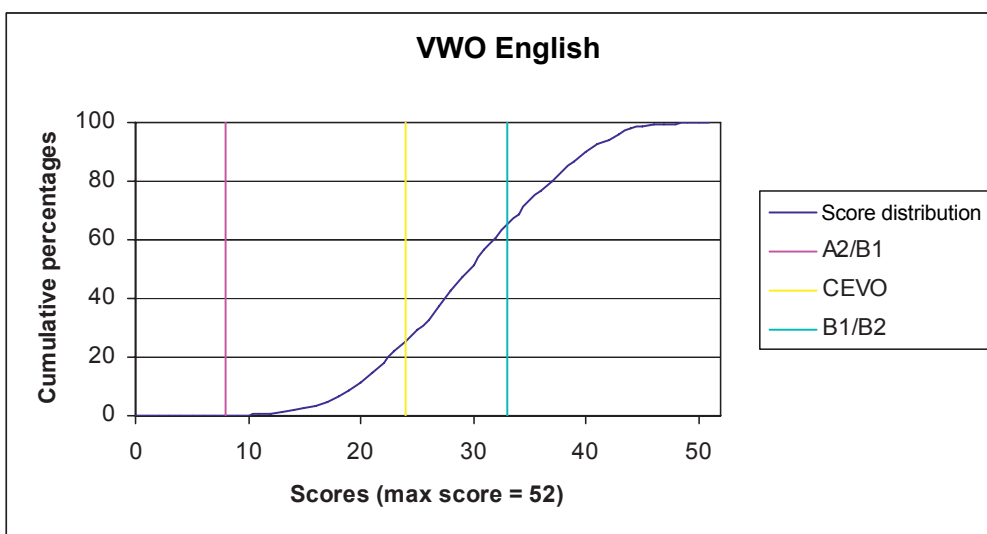
Figure 4.7 *Distribution of scores and cut-off scores have examination in English*



In figure 4.7 we see that two cut-off scores for CEFR-level have been computed: A2/B1 and B1/B2. That is to say that the scores to the right of the cut-off score line A2/B1, so scores of 19 and higher, belong to the B1 level. If candidates get scores higher than 42 then this indicates a level of B2. From the internal validation process and from the judgments of Judges we have concluded that this examination is aimed at students in the B1 to B2 range. We find the CEVO sufficient/insufficient score to reflect this. For students to get a sufficient mark on this examination, they need to have a score that is considerably higher than the one at the A2/B1 cut-off point. However, students can pass this examination without having reached B2 level. Even though only a small percentage of candidates (ca 5%) reach B2 level for this examination, in all fairness one can say that candidates can prove to have reached level B2 in this examination. After all, with a score of 42 out of 51 items one can claim level B2. The fact that only 5% really reaches this level B2 does not reduce the opportunity to demonstrate B2 competency.

In figure 4.8 below we see the distribution of scores and cut-off scores is given for the vwo examination in English.

Figure 4.8 *Distribution of scores and cut-off scores vwo examination in English*



We see that cut-off scores with two CEFR-levels have been computed: A2/B1 and B1/B2. That is to say that the scores to the right of the cut-off score line A2/B1, so scores of 8 and higher, belong to the

B1 level. If candidates get scores higher than 33 this indicates a level of B2. On the basis of the internal validation process and from the estimates of judges we have concluded that this examination is aimed at students in the B2 to C1 range. We find the CEVO cut-off score does not support this. Students can pass this examination with a score that is considerably lower than the score that judges expect at B2 level. However, a considerable percentage of students (circa 35%) do in fact reach B2 level on this examination. It needs to be said that it is not possible to prove with a high score on this examination that one has reached the C1 level: there were not enough items the C1+ basket.

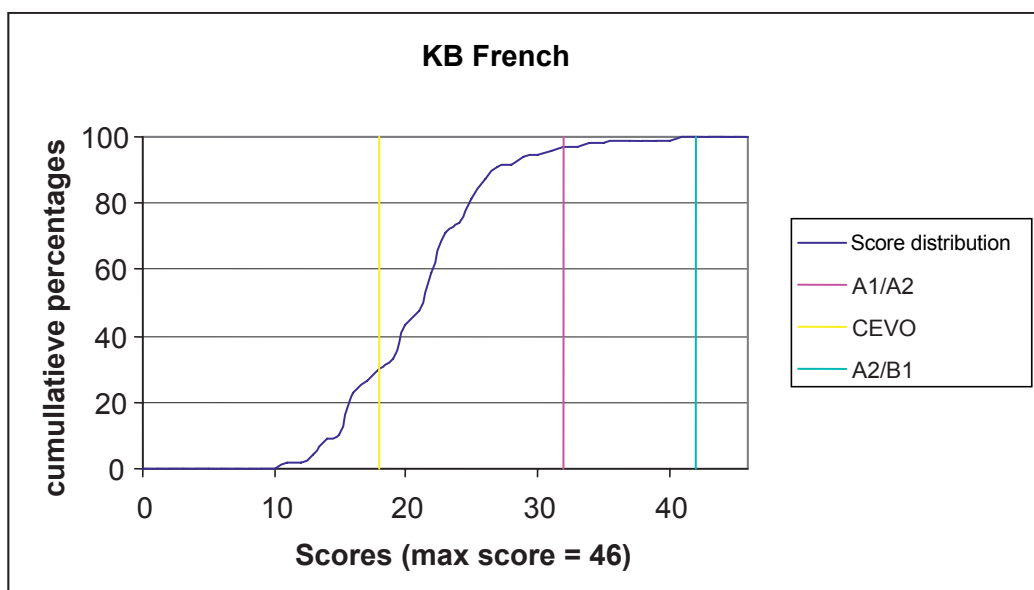
4.4.2 Results standard-setting French

In the sections below an overview is given of the findings for the state examinations of reading comprehension in French: the scores that go with the cut-off scores sufficient/insufficient, as determined by CEVO and the scores that go with the relevant cut-off scores for CEFR-levels. We will illustrate to what extent the judgements of CEVO what is a sufficient performance correspond with the estimates during the standard-setting for relevant CEFR-levels.

The kb examination in French is different from other reading examinations in French and in the other foreign languages in that it is a computer-based test. The sample was not large enough to carry out reliable IRT-analyses on that sample.

In figure 4.9 below the distribution of scores and cut-off scores is given for the KB-examination in French.

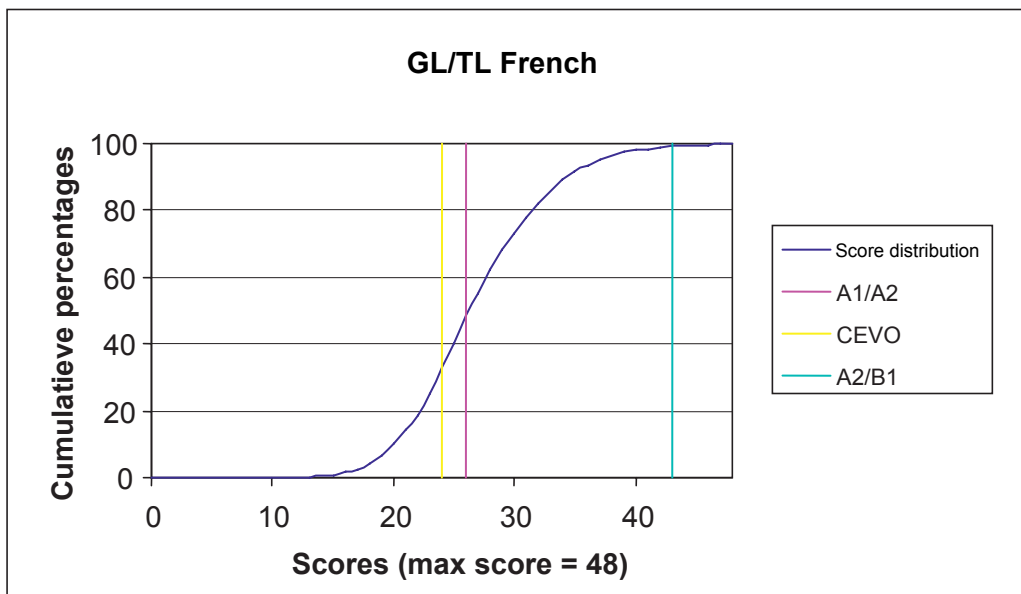
Figure 4.9 *Distribution of scores and cut-off scores KB examination in French*



The jagged line of the distribution of the scores in this figure is caused by the small number of examination candidates that took the kb examination. In this figure we see that two relevant CEFR cut-off scores have been computed. The CEVO cut-off point is in between these two points. The judges estimate that the minimum level for the kb examination in French is positioned just beyond A2 (see figure 4.1). It is clear that the CEVO cut-off score is positioned considerably lower. This means that candidates get a pass on this examination even if they have a level that is lower than the level determined as minimum CEFR-level during the standard-setting.

In figure 4.10 below the distribution of scores and cut-off scores is given for the gl/tl French.

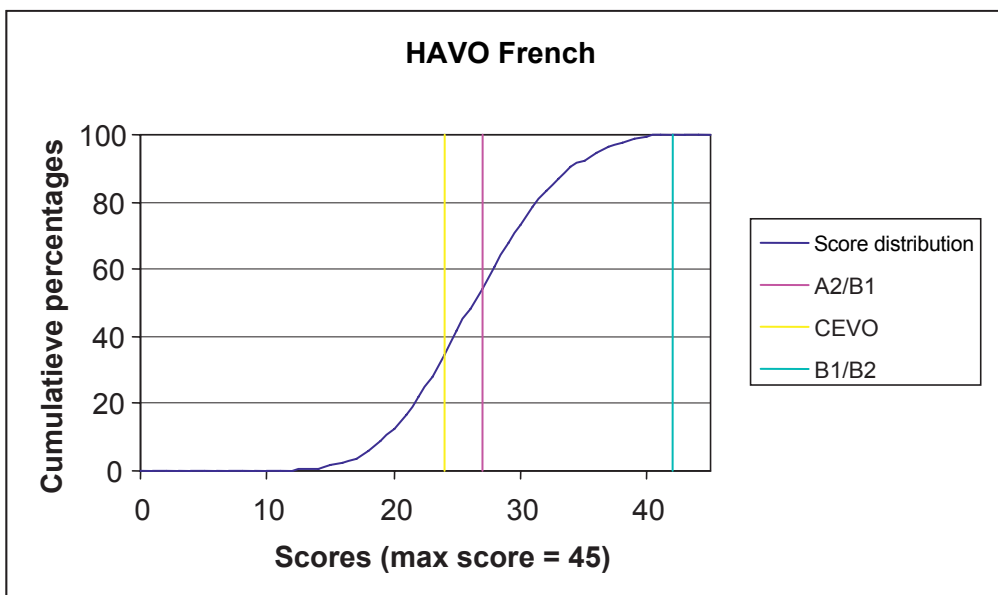
Figure 4.10 *Distribution of scores and cut-off scores gl/tl examination in French*



In this figure we see that two cut-off scores for CEFR-levels have been computed. The CEVO cut-off score is positioned almost at the same point as the lower cut-off score of A1/A2. However, in the standard-setting the judges estimated that the minimum level for the gl/tl examination in French is halfway between A2 and B1 (see figure 4.1). It is clear that the CEVO cut-off score is considerably lower than in fact any other relevant CEFR cut-off score (as estimated during the standard-setting). So again in this case candidates can pass the examination even though they have acquired a level that is lower than any minimum CEFR-level estimated during the standard-setting.

In figure 4.11 below the distribution of scores and cut-off scores is given for the haveo examination in French.

Figure 4.11 *Distribution of scores and cut-off scores haveo examination in French*

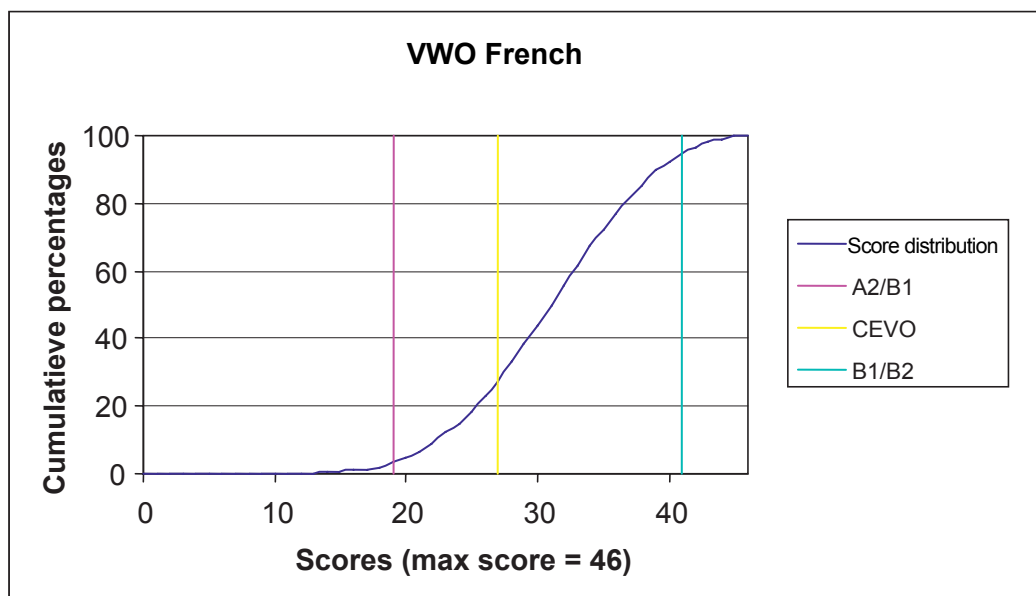


In this figure 4.11 we see that two relevant CEFR cut-off scores have been computed, at A2/B1 and at B1/B2. The CEVO cut-off score is positioned close to the lower cut-off score at A2/B1. The judges estimate with the standard-setting that the minimum level for the haveo examination in French is just

past B1 (see figure 4.1). The CEVO cut-off score is positioned just a little under the most relevant CEFR cut-off score (as estimated during the standard-setting). Although not a single candidate does achieve the B2 level on this examination, candidates on this examination could have proved that they had acquired the B2 level, if they had gained a score of 42 on 45 items

In figure 4.12 below the distribution of scores and cut-off scores is given for the vwo examination in French.

Figure 4.12 *Distribution of scores and cut-off scores vwo examination in French*



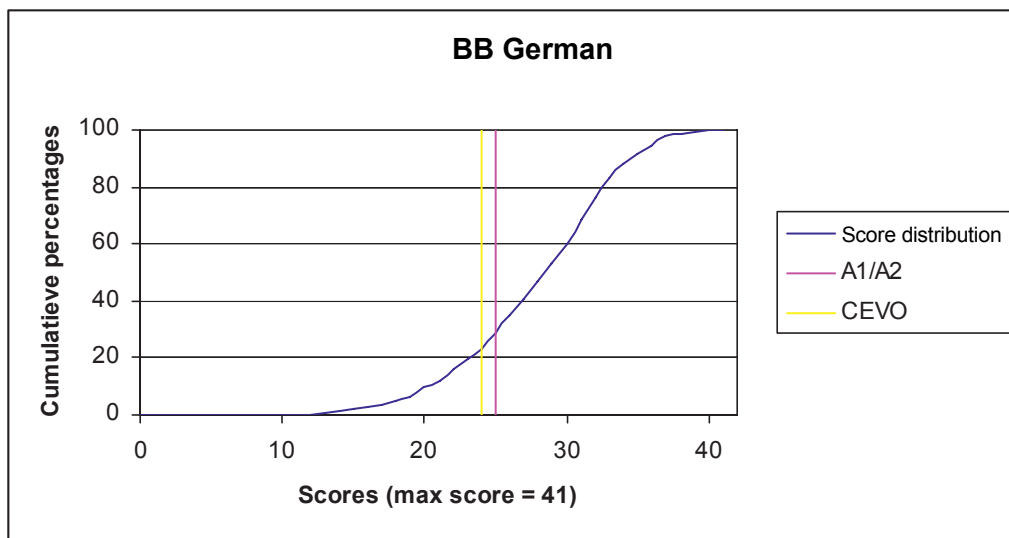
In figure 4.12 we see that two relevant CEFR cut-off scores have been computed, at the level A2/B1 and at the level B1/B2. The CEVO cut-off score is positioned in between those two levels. The judges estimated at the standard-setting that the minimum-level for the vwo examination in French is positioned just past B1 (see figure 4.1) CEVO expects of candidates on the vwo-level French a minimum-level that seems to correspond with the level that is estimated by the judges at the standard –setting as the minimum-level for the vwo examination.

4.4.3 Results standard-setting German

In the sections below an overview is given of the findings with the state examinations of reading comprehension in German: the scores that go with the cut-off scores sufficient/insufficient, as determined by CEVO and the scores that go with the relevant cut-off scores for CEFR-levels.

In figure 4.13 below the distribution of scores and cut-off scores is given for the bb examination in German.

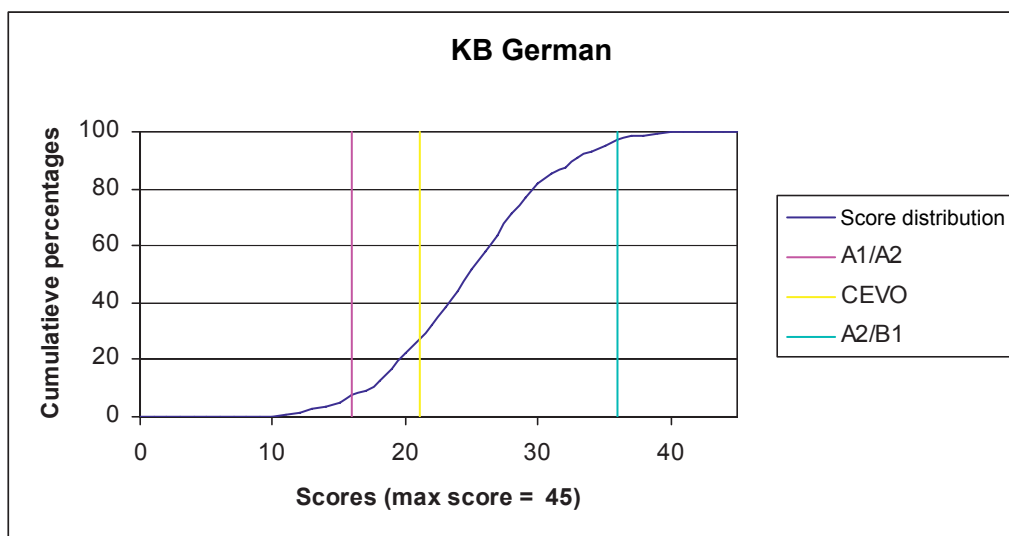
Figure 4-13 *Distribution of scores and cut-off scores bb examination in German*



In this figure we see that the CEVO cut-off score coincides almost with the relevant CEFR cut-off score A1/A2. Any candidate that passed the test, also reached the relevant A2-level.

In figure 4.14 below the distribution of scores and cut-off scores is given for the KB examination in German.

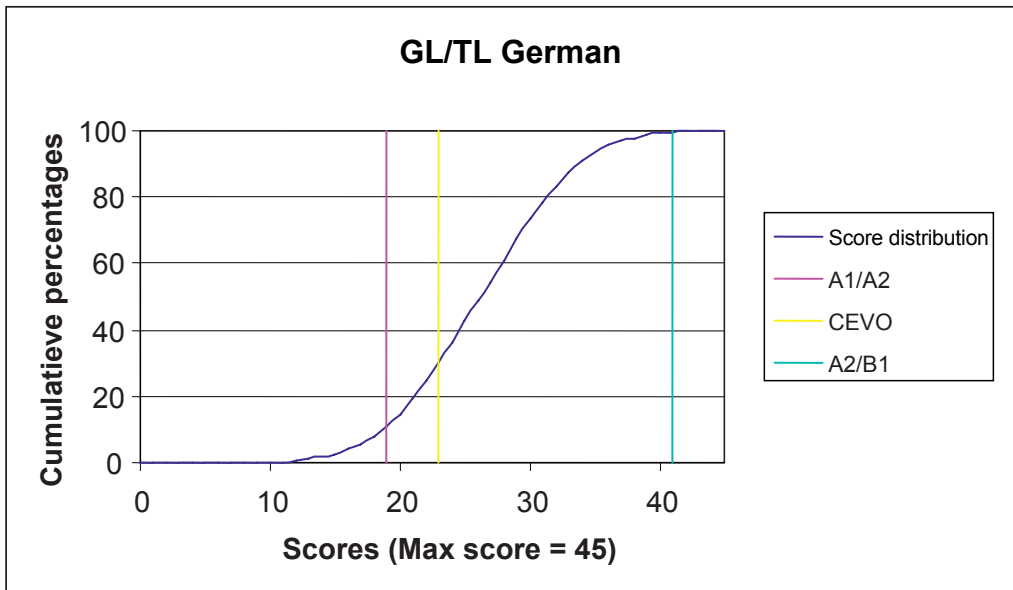
Figure 4-14 *Distribution of scores and cut-off scores KB examination in German*



In this figure we see that two relevant CEFR cut-off scores have been computed. The CEVO cut-off point is in between these two points. The judges estimate that the minimum level for the KB examination in German is positioned between A2 and B1 (see figure 4.2). It is clear that the CEVO cut-off score is positioned considerably lower.

In figure 4.15 below the distribution of scores and cut-off scores is given for the gl/tl examination in German.

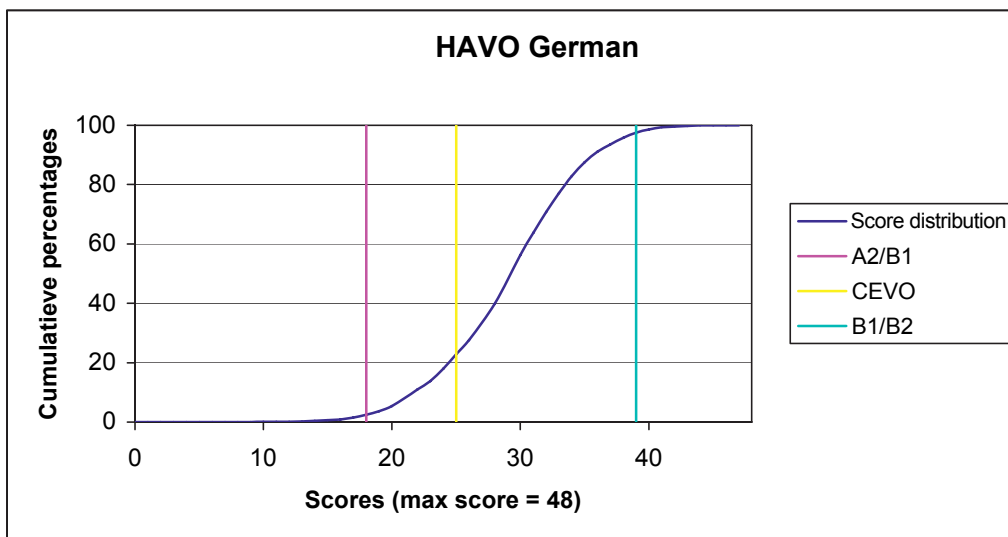
Figure 4.15 *Distribution of scores and cut-off scores gl/tl examination in German*



In this figure 4.15 we see that two relevant CEFR cut-off scores have been computed, at A1/A2 and at A2/B1. The CEVO cut-off score is positioned in between, close to the cut-off score at A1/A2. The required minimum level for the gl/tl examination is positioned between A2 and B1 (see figure 4.2). It is clear that CEVO and the judges agree with one another in their judgement that this examination should indicate if a candidate can be placed at A2 level. It must be noted here that none of the candidates on this examination acquires the B1 level.

In figure 4.16 below the distribution of scores and cut-off scores is given for the havo examination in German.

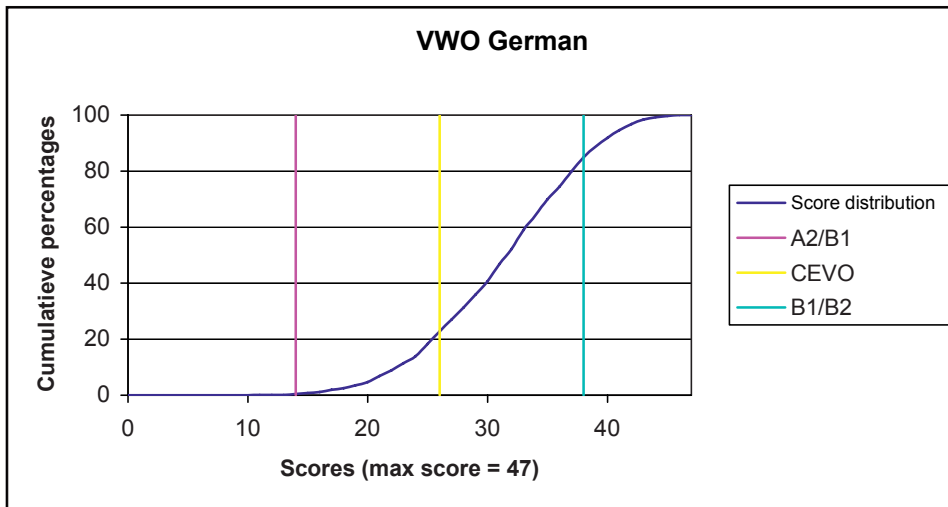
Figure 4.16 *Distribution of scores and cut-off scores havo examination in German*



In this figure we see that two relevant CEFR cut-off scores have been computed, at A2/B1 and at B1/B2. The CEVO cut-off score is positioned in between. The judges estimate that the minimum level for the havo examination is positioned in between B1 and B2 (see figure 4.2). It is clear that the CEVO cut-off score is positioned considerably lower. Only a very small percentage of candidates achieve the B2 level in the havo-examination.

In figure 4.17 below, the distribution of scores and cut-off scores is given for the vwo examination in German.

Figure 4.17 *Distribution of scores and cut-off scores vwo examination in German*



In this figure we see that two relevant CEFR cut-off scores have been computed, at A2/B1 and at B1/B2. The CEVO cut-off score is positioned exactly in-between. The judges estimate that the minimum level for the vwo examination is positioned in between B1 and B2 (see figure 4.2). It is clear that the CEVO cut-off score is positioned considerably lower. Only a small percentage of candidates achieve the B2 level in the vwo examination.

5 Summary, conclusions and recommendations

Commissioned by the Dutch Ministry of Education, Culture and Science, SLO and Cito have carried out a project which had the following goals:

- Linking the existing examination syllabuses and state examinations of reading comprehension in French, German and English to the CEFR according to the steps described in a manual for linking examinations to the CEFR published by the Council of Europe in 2003;
- Exploring possibilities of producing more complete or broader CEFR-related state examinations of reading comprehension in the foreign languages.

5.1 Summary

In this report an account is given of the content specification and the standardisation of the state examinations of reading comprehension in the foreign languages French, German and English as carried out by Cito. For this the procedures as described in the draft manual published by the Council of Europe have been followed. To arrive at valid specification and standardisation, project members first had to get familiar with the CEFR in the familiarisation phase. And later, when external judges were asked to determine the minimum CEFR-level of the items that occur in the reading comprehension examinations, these judges also had to get familiar with the CEFR.

Project members performed a content analysis of the examinations of reading comprehension during the specification process. The issue has been to find out to what extent examinations of reading comprehension at all secondary school levels did indeed contain texts and items that show an increase in CEFR-level. To be able to answer this question a CEFR-related descriptive model was used with which the texts and items have been described.

It proved possible to indicate for each of the examinations of reading comprehension for the various school types what the mean minimum-reading comprehension level of a candidate should be in terms of the CEFR in order to be able to answer the questions in an examination correctly. In addition to this we have been able to indicate for each of the examinations used (with the exception of bb-level French) what the minimum scores should be to reach the relevant CEFR-levels. We have compared these minimum scores per examination with the State Examination Committee's cut-off scores.

5.2 Conclusions

Below we give an overview of the main conclusions that we have been able to draw in this study. We should like to point out here that our task has been to link the current state examinations of reading comprehension in French, German and English to the CEFR. We leave it to others to draw conclusions as to the content of and required levels in the examination syllabi and the state examinations of reading comprehension.

1. Following all the proposed steps in the draft manual turned out to be a time consuming and costly process.
2. The linking process is a good way of critically reviewing and evaluating the content and statistical characteristics of the examinations in question.
3. From the content specification it appears that the emphasis in the state examinations of reading comprehension is on the global descriptor 'reading for information'. When the specific descriptors of the CEFR are taken into account, it is found that these specific descriptors are sufficiently represented in the examinations. It may be considered to include a few more texts for reading for correspondence and reading of instructions in the examinations, especially at the lower levels.
4. The examinations contain a variation in text sources, text types and topics as mentioned in the CEFR.

5. The examination texts, from low level to high level examinations, reflect the increase in linguistic and cognitive complexity, as supposed in the CEFR.
6. The variation in reading tasks which the candidates have to perform increases from low level to high level examinations.
7. The phases of specification and standardisation could best be carried out in both a national and an international context.
8. In many cases the State Examination Committee's cut-off scores (sufficient/insufficient) do not coincide with the cut-off points as estimated by judges for the relevant CEFR-levels.

5.3 Recommendations

- Further research will need to be carried out to find out if examinations exist that have been calibrated in the way the draft manual proposes. For a number of tests it is claimed that they have been linked to the CEFR in a valid and reliable way. However, we have the impression that often this is not done in the way the manual proposes.
- Foreign test development organisations will have to be contacted for empirical validation. A number of these organisations have already shown an interest in particular forms of cooperation with empirical validation.
- An empirical validation (external validation) will have to be carried out in order to be able to fully validate links with the CEFR. The Dutch Ministry of Education, Culture and Science has provided funds for this activity.
- For 2006 a proposal for a comparable linking project for Arabic, Spanish and Turkish was made to the Dutch Ministry of Education, Culture and Science. This activity is now also being carried out by Cito, funded by the Ministry.
- A popular version of the present report with the most important findings should be published. This publication will promote familiarity with the CEFR among all those involved, especially teachers and students. This version has since been published.

6 References

Alderson, J. Charles (2006), Neus Figueras, Henk Kuijper, Günter Nold, Sauli Takala, Claire Tardieu 'Analysing Tests of Reading and Listening in Relation to the Common European Framework of Reference: The Experience of the Dutch CEFR Construct Project.' *Language Assessment Quarterly*. 3 (1), 3-30.

CEVO (2006), *Syllabus voor het Centraal Examen Moderne Vreemde Talen*, Utrecht.

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

Council of Europe (2003), Language Policy Division, *Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching and Assessment. Manual Preliminary Pilot Version*. Council of Europe, Strasbourg.

DIALANG (2002), *Diagnostic tests in 14 languages on the internet*: www.dialang.org.

Liemberg, E. (2004) & D. Meijer, *Taalprofielen*. NaB-MVT, Enschede.

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

Noijons, José (2006b) & Henk Kuijper, *Leesvaardigheidsexamens moderne vreemde talen in Europees verband*, Cito, Arnhem.

ANNEX The Dutch Educational System

Short overview of the Dutch Educational System

Primary education

basisschool compulsory from age 5

Secondary education:

vwo pre-university education (6 years, ages 12-18)
havo senior general education (5 years, ages 12-17)
vmbo lower secondary & pre-vocational education (4 years, ages 12-16)

- theoretical track (tl)
- combined track (gl/tl)
- advanced vocational track (kb)
- basic vocational track (bb)

Vocational education

mbo secondary vocational training
bve combination of secondary vocational education and training and adult education.

Higher education

hbo higher vocational education (four years, after havo/vwo)
universiteit university (4 years, after vwo)