What does it mean for a test to be at a particular level?
The need for indicative cut scores

José Noijons
Cito, Dutch National Institute for Educational Measurement

Introduction

Increasingly, language tests in Europe are being linked to the Common European Framework of Reference (CEFR) developed by the Council of Europe (Council of Europe, 2001). This framework distinguishes levels of language competence in terms of descriptors or can-do statements. There have now been a number of studies in which testing agencies claim that their tests are at a particular CEFR-level. In many cases, this has as yet been done through specification of content. Such analysis may successfully show that tests typically cover descriptors that are mentioned in the CEFR at a particular level. However, these studies do not always indicate what score the candidate needs to reach to be given a particular CEFR level. It is one thing to pass a test; it may be another thing to pass the test *at a particular CEFR level*.

In 2004 the Dutch Ministry of Education, Culture and Science commissioned the Dutch Institute for Curriculum Development (SLO) and the Dutch National Institute for Educational Measurement (Cito)

- to establish links between the existing examinations in French, German and English and the CEFR, following the steps as outlined in the preliminary pilot version of the manual[1] published by the Council of Europe (Council of Europe, 2003).
- to study the possibilities of developing more comprehensive CEFR-related examinations in the foreign languages.

The research study as carried out for the Dutch Ministry of Education has been the basis of the present article. In our study we have been confronted with the phenomenon that tests were specified at a particular CEFR level, but that the cut sores for these tests did not necessarily reflect the CEFR level the tests were claimed to be at. An interim report on this study has been published in English (Noijons, 2005).

The linking process

The linking process has been carried out following the procedures as proposed in the Manual and outlined in figure 1 below.

Specification of examination content	Standardisation of judgments	Empirical validation through analysis of test data
⇩	⇩	⇩
Claim of link to the CEFR on the basis of specification	Claim of link to the CEFR on the basis of specification and standardisation	Claim of link to the CEFR - confirmation on the basis of empirical verification

Figure 1 *Visual representation of procedures to relate examination to the CEFR.*

[1] This manual will be referred to in this article as "the Manual".

In the linking process a number of phases have been identified:
Phase 1: *Familiarisation*
Participants in the study are to be made familiar with all aspect of the CEFR.
Phase 2: *Specification*
The content of the examinations under review is to be specified (in terms of the CEFR).
Phase 3: *Standardisation*
Experts are to set standards of minimum competence for each CEFR-level.
Phase 4: *Empirical validation*
Confirmation of a relationship to the CEFR through an independent measure.

Phases 2 and 3 will be discussed in this article.

Phase 2: Specification
As is outlined in the Manual, specification involves mapping the coverage of the examination in relation to 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 we have interpreted the first activity as follows.
(1) A qualitative analysis is to be made of the existing examination syllabus for foreign languages. This analysis is made for each can-do statement in the CEFR using existing specifications. This analysis has been carried out by SLO.

The study pertained to Dutch examinations in reading comprehension in three languages: French, German and English. However, in this article only the examinations in English will be discussed. Thee examinations are at five levels (from high level to low level):

VWO	pre-university
HAVO	higher secondary
GL/TL	lower secondary (combination of general and pre-vocational)
KB	lower secondary: pre-vocational
BB	lower secondary: pre-vocational

Linking attainment targets in the foreign language syllabuses to the CEFR has raised a few issues:

- Attainment targets for all levels of education have been phrased in nearly identical terms, which makes it rather difficult to link targets to a specific CEFR level.
- Differences between levels of education have been expressed through indicators of (performance) levels, rather than through descriptors of language behaviour (as is done in the CEFR).

In the following tables we present how CEFR-levels have been compared to the attainment targets to be found in the Dutch foreign language syllabi. This has been a matter of much interpretation as the formulation of attainment targets differs substantially from the CEFR descriptors.

Table 1 *General level and attainment targets in syllabi for English related to CEFR-levels*

CEFR domains	School type BB	School type BB	School type GL/TL	School type HAVO	School type VWO
	Attainment targets	*Attainment targets*	*Attainment targets*	*Attainment targets*	*Attainment targets*
Overall Reading Comprehension	A2	A2	B1	B2	C1
Reading Correspondence	A2	A2	B1	B2	B2
Reading for Orientation	A2	A2	B1.2-B2.1	B2.1	B2 - C1
Reading for information and argument	A2	A2	B1.2-B2.1	B2.1	C1
Reading instructions	A2	A2	B1	B2	B2
Reception Strategies	A2	B1	B1	B2	B2

(2) The second activity in the specification process pertains to a content analysis of the examinations themselves. Two steps in the examination content analysis have been distinguished.

1. Project members (content specialists) were to analyse the content of the 2004 reading examinations in the foreign languages in terms of the CEFR.
2. On the basis of the entries by the content specialists an overview of content characteristics for an examination at each level in each language was to be made.

Table 2 *Characteristics of texts: discourse types of text and distribution in percentages*

Text types	BB	KB	GL/TL	HAVO	VWO
Descriptive	17	60	14	40	18
Narrative	70	40	82	5	12
Expository	4	-	-	15	6
Argumentative	4	-	4	30	53
Instructive	4	-	-	10	12

Table 3 *Characteristics of items and distribution in percentages*

Items	Characteristics	BB	KB	GL/TL	HAVO	VWO
Response type	MC	67	52	54	48	46
	True/false	-	-	-	-	4
	short answer	33	40	39	35	32
	gap filling	-	7	7	16	19
Behaviour required from students	recognise and retrieve	88	51	39	-	-
	make inferences	11	48	60	100	100
	evaluate	1	1	1		
Type of information in text	explicit	100	94	100	99	79
	implicit	-	6	-	1	21
Content of questions	main idea/gist	21	19	20	28	32
	detail	78	67	67	45	37
	Opinion	1	-	-	6	12
	writer's attitude	-	-	-	5	3
	Conclusion	-	7	3	5	7
	communicative purpose	-	2	3	1	1
	text structure	-	5	7	9	7

Conclusions specification phase

It was noted that the majority of the examination questions was aimed at retrieving explicit and implicit information. However, the CEFR clearly expects readers to be able to do more, and do more complex tasks, certainly at the higher levels. The reason for the relative scarcity of questions tapping other behaviour is that the syllabi, and the examination matrices based on them, require the examinations to focus on retrieving information.

In the specification phase we have found that an increase in exam difficulty (as specified in the Dutch examination system) is reflected in the links with the CEFR levels: the more difficult the examination is, the higher the CEFR level that is associated with such an examination.
The CEFR assumes that as a person's reading proficiency level rises, he or she can read more complex texts (linguistically and cognitively) and can carry out more complex tasks. We have found that an increase in the difficulty of the exams is mainly to be attributed to an increase in the difficulty of the texts, less so to the complexity of the tasks.

Phase 3: Standardisation
As we have seen, the Manual recommends for the standardisation phase that experts set standards of minimum competence for each CEFR-level. The aim has been for judges to determine the minimum CEFR-levels needed by candidates to successfully perform on a given a language test. In other words, to determine cut 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 test.

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 lowest level at which that item should be mastered. There are five baskets, called A1, A2, B1, B2 and C1+, corresponding to the levels that the examination syllabuses aim at (and beyond). If an item is placed in basket B1, this means that according to the judge, a person at level B1 should master the item 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 master the item.

For each item taken from the five examinations and presented to the judges in a random order, the judges were instructed to do the following.

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

Data collection and data analysis
The rating forms have been collected and data entry has taken place at Cito. Data collected included: rater ID, language and rating (1 to 5, corresponding to A1 to C1+) per item.
Data analysis was carried out to validate the accuracy of the standards. The data analysis has comprised two operations:
1. Determining rater agreement and required minimum CEFR-*level* per examination ;
2. Determining minimum *scores* for relevant CEFR-levels on each examination.

(1) Raters sufficiently agree on the minimum CEFR-*level* required for each item to be mastered (see table 4, below). Raters 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. It

must be emphasized here that the items were presented to the judges in random order and that judges were not told from which level examination an individual item was taken.
Data analysis shows that raters are of the opinion that going from low level examinations to high level examinations, for each type of examination an increasing CEFR-level is needed to be able to successfully answer the questions (see figure 3, below).

Table 4 *Rater reliability and rater agreement (English)*

Examination	**Rater reliability (α)**	**Rater agreement (Rho2)**
BB	.78	.74
KB	.92	.90
GLTL	.79	.78
HAVO	.74	.67
VWO	.74	.72

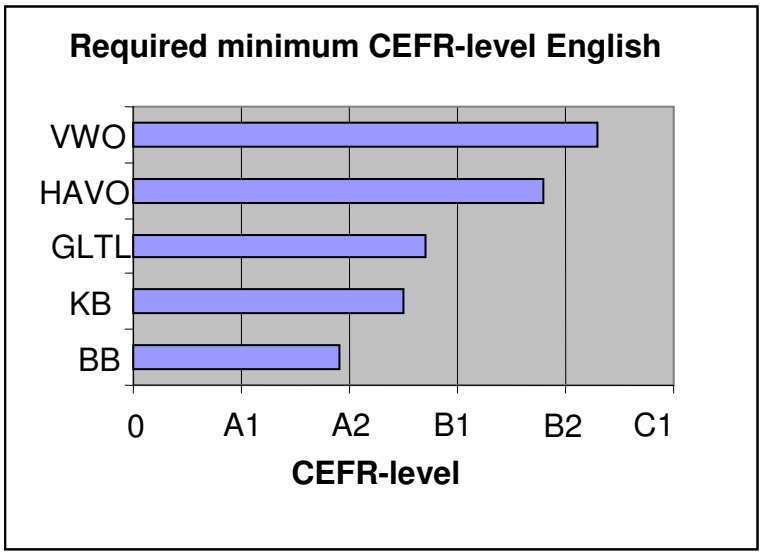

Figure 3 *Required minimum level per examination (English)*

(2) The next step in the data analysis phase has been to determine the minimum *score* on an 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 score for sufficient/insufficient as determined by the State Examination Committee (SEC) would mean in terms of mastery of CEFR-levels.

In the present examination procedures in the Netherlands the SEC determines the cut score for each examination (in each subject). A candidate is said to have passed an examination when he or she has a score that is higher than the cut score. When it is claimed that a Dutch examination in one of the foreign languages is at a particular CEFR-level we need to look at where the SEC cut score is positioned, as that is the only score that has a "civil" effect.

In the following tables we will illustrate where SEC cut scores and relevant CEFR cut scores are to be found in the Dutch examinations for English. We will look at possible differences in CEFR-level between a student that has passed the examination at the cut point and a student

that has passed the examination at the relevant CEFR-level. Also an indication of the score distribution (and consequently of the corresponding CEFR-level) of the sample student population for each examination is given.

In figure 4 we see that the only relevant CEFR level cut score to be computed is that between A1 and A2. We find that the SEC sufficient/insufficient cut score is lower. This means that a student can pass the BB examination without having reached A2 level.

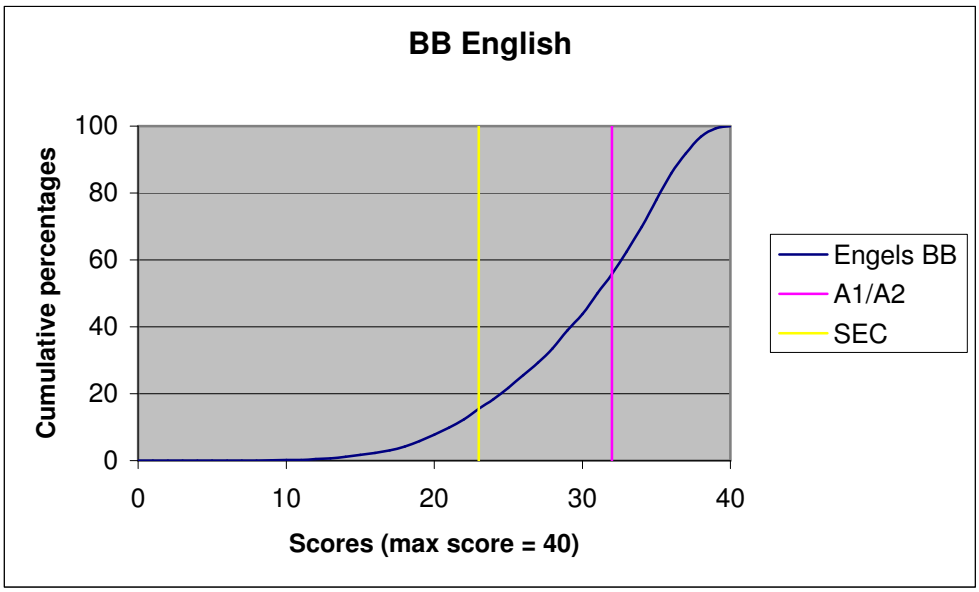

Figure 4 *Distribution of scores and cut scores in BB examination English*

In figure 5 we see that two relevant CEFR level cut scores have been computed: A1/A2 and A2/B1. We find that the SEC sufficient/insufficient cut score is just over the A1/A2 cut score. A small percentage of students reaches B1 level on this examination.

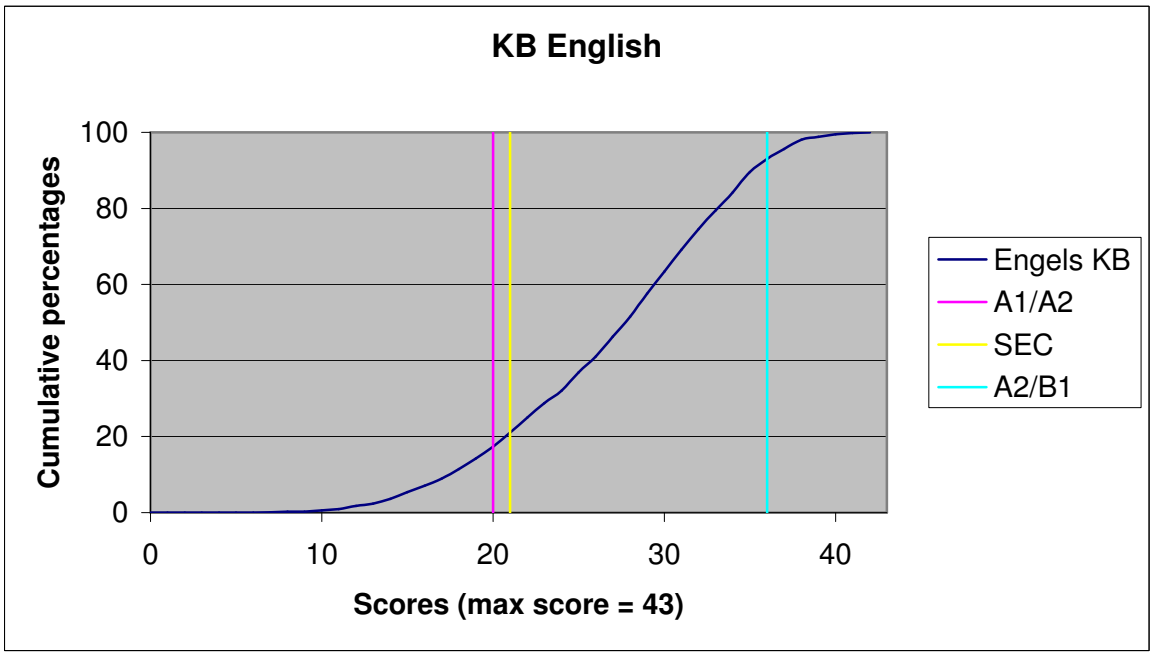

Figure 5 *Distribution of scores and cut scores in KB examination English*

In figure 6 we see that two relevant CEFR level cut scores have been computed: A1/A2 and A2/B1. We find that the SEC sufficient/insufficient cut score is considerably higher than the A1/A2 cut score, but considerably lower than the A2/B1 cut score. A higher percentage of students than in the KB examination has reached B1 level on this examination.

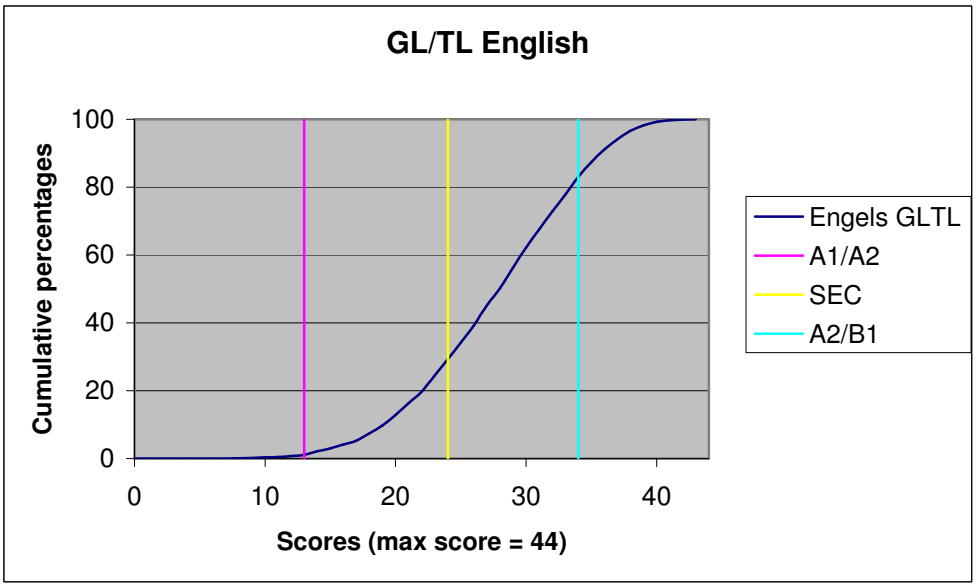

Figure 6 *Distribution of scores and cut scores in GL/TL examination English*

In figure 7 we see that two CEFR level cut scores have been computed: A2/B1 and B1/B2. From the internal validation process and from the judgments of raters we have concluded that this examination is aimed at students in the B1 to B2 range. We find the SEC sufficient/insufficient score to reflect this. For students to pass this examination, they need to have a score that is considerably higher than at the A2/B1 cut point. However, students can pass this examination without having reached B2 level. Only a very small percentage (ca 5%) reaches B2 level on this examination.

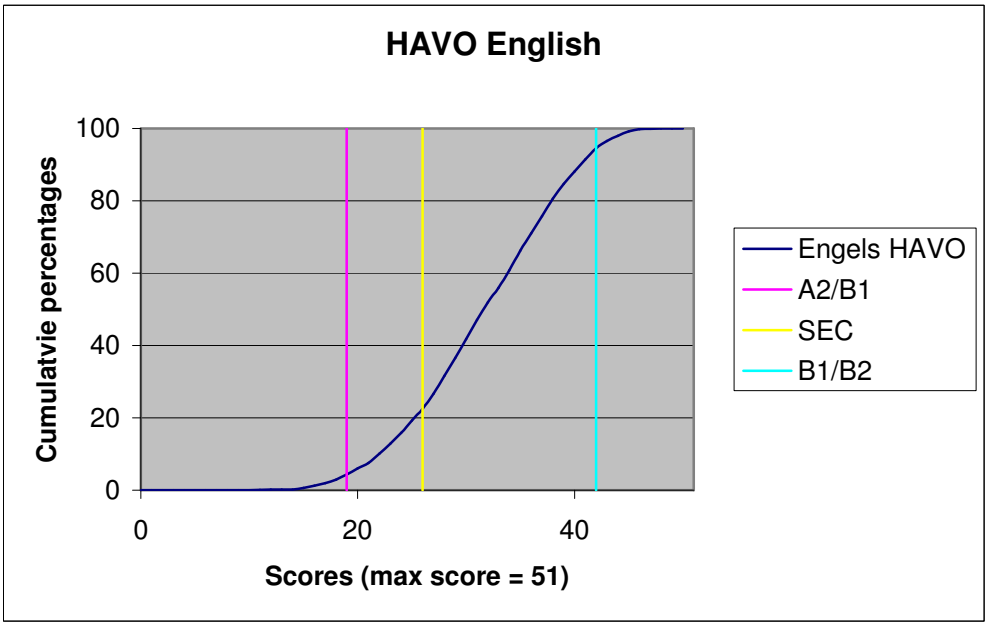

Figure 7 *Distribution of scores and cut scores in HAVO examination English*

In figure 8 we see that two CEFR level cut scores have been computed: A2/B1 and B1/B2. From the internal validation process and from the judgments of raters we have concluded that this examination is aimed at students in the B2 range. We find that the SEC sufficient/insufficient score does not reflect this. Students can pass this examination with a score that is considerably lower than what experts expect at B2 level. However, a considerable percentage of students (ca 35%) does reach B2 level on this examination.

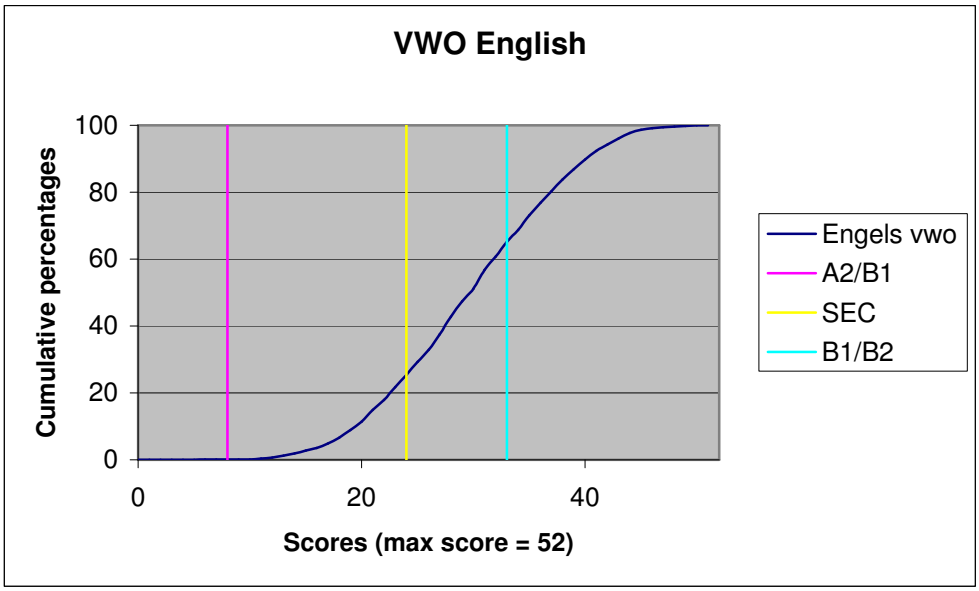

Figure 8 *Distribution of scores and cut scores in VWO examination English*

General conclusions

At Cito, the Dutch National Institute for Educational Measurement, research has been carried out on the levels of existing foreign language examinations produced by Cito. Two important phases in this research have been specification of examination content and standard-setting. It has been found that for most examinations it is possible to say: within this range of scores the test taker is at this particular CEFR-level and within another score range on the same examination, the test taker is at a higher CEFR-level. However, the existing pass/fail score (the sufficient/insufficient score) has little or no relationship with these levels and may be anywhere in the possible score range of the examination. In fact, sufficient/insufficient decisions in Dutch foreign language examinations do not generally coincide with cut scores for specific CEFR-levels.

The claim that testing agencies make: that their examinations are at a particular CEFR-level may be based on specification of examination content only. If this is the case, we do not really know what it means for a student to have "passed" the examination in terms of the CEFR. We do not know what CEFR-level the student has reached.

If scores on examinations are related to CEFR-levels through standard setting, it is important to see to what extent the specified content of the examination relates to the CEFR-level the student is given. If a student has a (very) high score on a test and is (therefore) placed at a high CEFR level, this may not be a valid judgement. If the majority of tasks in the test relates to a relatively low level, we can only say that the student is good at performing these relatively low-level tasks. We would need to set an examination at a higher CEFR-level (content-wise) to be able to place students at a higher level. Within the CEFR framework it is difficult to accept that one test with a variety of tasks at a number of CEFR-levels is able to place students at various levels depending on their scores.

References

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.

Noijons, J. (ed.) (2005), *Mapping the Dutch Foreign Language State Examinations onto the Common European Framework*; Interim Report. Cito, Arnhem.