Reconciling national testing assessment practices with the CEFR-linked assessment: (how) can it be done?

Sauli Takala

Introduction

Evaluation (or almost a synonymous terms "assessment") is usually regarded as an activity whose purpose is to determine the worth (merits, quality) of objects, performances or activities, programs or systems. All evaluation/assessment needs criteria for what counts as quality, i.e., characteristics/ attributes of merit. In language education – as in all forms of education – curricula and syllabi normally function as such criteria. For this reason, it is usually an important question how close the link is between objectives and evaluation. Tests are an important, though by no means the only, source for making evaluations.

The results of any form of evaluation/assessment are reported using some system of providing such feedback information.

Testing and assessment are usually felt to be difficult and even unpleasant, something that is done because there is an obligation to provide assessments. Testing, assessment and examinations have sometimes been characterized as a task of "duty to society", a duty many would prefer not to have to perform.

There are a great variety of assessment practices depending on the purpose and function of assessment. The spectrum covers a broad range from self-assessment, peer assessment and teacher assessment to assessment for selection and placement, various forms of external assessment and culminating in very high-stakes tests/examinations of various kinds.

There has been systematic work done in testing to develop its theoretical foundations. This has resulted in what are usually called Classical Test Theory (CTT) and Item Response Theory (IRT). For decades, test development has also been a focus of very useful and productive research and development effort. This means that at present we have a good basis for carrying out assessments in a competent manner, and also for evaluating how well they have been carried out. However, continuous social and educational developments mean that new needs emerge, which set new demands for assessment. Perhaps the most prominent such development is the increasing trend to what is called standards-based assessment.

The currently much used term "standard setting" in the field of testing and assessment refers to a decision making process, which seeks to classify the results of test/examination in a limited number of successive levels of achievement (proficiency, mastery, competency). This is, of course, nothing new but but something that teachers have always been doing when they have graded student

Dervin, F. & Suomela-Salmi, E (eds.)
New approaches to assessing language and (Inter) cultural competences in higher education
Frankfurt am Main: Peter Lang.

performances according to some marking system. What is new is that performances are increasingly reported using some form of more explicitly defined scale of a range of performances.

With the development and increasing use of the Common European Framework of Reference for Languages (CEFR 2001) there is a tool which can be used to report language proficiency in a sufficiently transparent and comparable manner and for setting standards.

1. Relating national examination to the CEFR

The Common European Framework of Reference for Languages has a very broad aim. It was developed to provide:

> [...] a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. The description also covers the cultural context in which language is set. The Framework also defines levels of proficiency which allow learners' progress to be measured at each stage of learning and on a life-long basis. (CEFR: 1)

But the CEFR is also specifically concerned with testing and examinations:

> One of the aims of the Framework is to help partners to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications. For this purpose the Descriptive Scheme and the Common Reference Levels have been developed. Between them they provide a conceptual grid which users can exploit to describe their system. (CEFR: 21)

The CEFR is already serving this function flexibly through validated national versions of the European Language Portfolio. By contrast, the mutual recognition of language qualifications awarded by all relevant bodies is a complicated matter. As the Council of Europe's Manual for Relating Examinations to the CEFR (2003/2008) notes

> The language assessment profession in Europe has very different traditions. At the one extreme there are examination providers who operate in the classical tradition of yearly examinations set by a board of experts and marked in relation to an intuitive understanding of the required standard... Then again there are many examinations that focus on the operationalisation of task specifications, with written criteria, marking schemes and examiner training to aid consistency, sometimes including and sometimes excluding some form of pre-

> testing or empirical validation. Finally, at the other extreme, there are highly Centralised examination systems qualifications... National policies, traditions and evaluation cultures as well as the policies, cultures and legitimate interests of language testing and examination bodies are factors that can constrain the common interest of mutual recognition of qualifications. However it is in everybody's best interests that good practices are applied in testing. One of the aims of the Framework is to help partners to describe the levels of proficiency required by existing standards, tests and examinations in order to facilitate comparisons between different systems of qualifications. For this purpose the Descriptive Scheme and the Common Reference Levels have been developed. Between them they provide a conceptual grid which users can exploit to describe their system. (CEFR 2003/2008: 21)

Following the publication of the CEFR, there were many calls for the Council of Europe (COE) to take an active – even a controlling – role in checking how well examination providers were doing in their efforts to validate the relationship of their examinations to the Common European Framework of Reference. The topic was the theme of a seminar in Helsinki in July 2002. It recommended that assistance (not control) was indeed needed. The Language Policy Division of the Council of Europe in Strasbourg responded by setting up the project[1] to develop a Manual for this purpose. Standard setting (setting cut-off scores) is the key concern of the Manual.

Related to the theme of standard setting, two terms referring to standards are used almost as synonyms to it (Hansche 1998; Hambleton 2001). These two terms are: content standards and performance standards. *Content standards* refer to the curriculum/syllabus/program of study and answer the question: *what* someone should know and be able to do as a result of a specific course of instruction? *Performance standards* on the other hand are "explicit definitions of what students must do to demonstrate proficiency at a specific level on the content standards" (CRESST Assessment Glossary 1999) and answer the question: *how good is good enough*?

Standard setting typically involves setting cut-off points on the scale (such as the CEFR reference levels), and it is easy to see that it is closely linked to performance standards. There is also a relation between standard setting and content standards, since performance standards are always related to some specific content standards. Communication is always about something, about some content.

Cut-offs are most commonly set on the basis of a distribution of (sum) scores. For instance, certain scores in a reading comprehension test are set as cut-offs for different CEFR levels. Sometimes performance standards are presented only as verbal descriptions for different performance categories (e.g.

1 The Authoring Group's composition: Brian North (Chair), Neus Figueras, Piet van Avermaet, Sauli Takala and Norman Verhelst.

Hambleton 2001: 92). Thus in assessing writing and speaking, test-takers can be classified by raters directly into one of the six CEFR performance levels by matching test-taker performance to the verbal descriptors of the corresponding CEFR scale of language proficiency. In the CoE Manual this process is called benchmarking. If this approach is used it is a special case of a standard setting procedure, which does not involve setting cut-offs.[2]

The Manual presents a set of activities, which are useful in being able to relate examinations to the CEFR and collect evidence to substantiate the validity of the claim of linkage[3]. These activities consist of: familiarization with the CEFR, accounting for the content of the examination in relation to the CEFR, training standard setting panelists to rate items/performances in a sufficiently reliable and consistent manner, providing evidence of the sufficient quality of the examination and examining the appropriateness of the cut-offs through various measures of external corroborating evidence. The process is illustrated in the figure below.[4]

The figure illustrates in a concrete fashion how many activities are needed, and also how demanding any project of relating examinations to the CEFR is bound to be.

2. Relating a national examination to the CEFR: some examples from Finland

2.1. Matriculation Examination

There is only one high stakes examination in general education in Finland: the matriculation examination[5] at the end of the upper secondary school (about age 19).

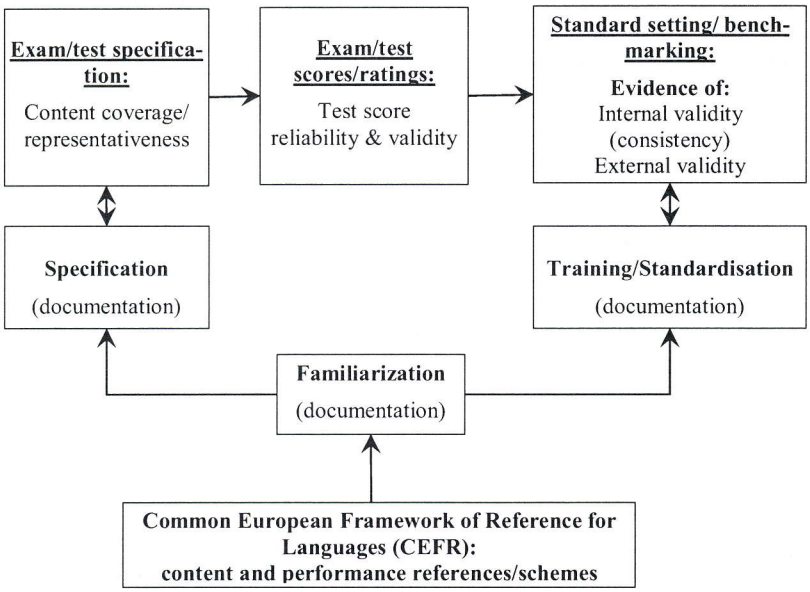
2 Of course performances can be rated on several analytical criteria and the scores can be summed, which makes traditional setting of cut-offs possible.

3 There is also a very useful basic introduction to standard setting in Cizek & Bunch (2007).

4 The figure is taken from the revised Manual and it was drafted by the present author.

5 The Matriculation Examination in its present form dates back to 1852.

Figure 1 - Validity evidence of linkage of examination/test results to the CEFR



It is administered twice a year, in spring and in autumn, in all Finnish upper secondary schools, at the same time. A candidate must complete the examination during not more than three consecutive examination periods. The examination can also be completed in one examination period.

The examination consists of at least four tests; one of them, the test in the candidate's mother tongue, is compulsory for all candidates. The candidate then chooses three other compulsory tests from among the following four tests: the test in the second national language, a foreign language test (in one or more languages), the mathematics test, and tests science and humanities subjects. As part of his or her examination, the candidate may include one or more optional tests.

Tests are arranged at two different levels of difficulty in mathematics, the second national language and foreign languages. The levels in mathematics and foreign languages are the advanced course and the basic course, and in the second national language the advanced course and the intermediate course. The candidate may choose which level of each of the above-mentioned subjects he or she takes, regardless of his or her studies at the upper secondary school. The candidate must take a test based on the advanced course in at least one compulsory test.

A candidate receives a matriculation examination certificate following the examination period when all the compulsory tests have been passed. The Ma-

triculation Examination Certificate shows the compulsory and the optional tests passed, together with their levels and grades. The grades (from top to bottom), corresponding credit points and their percentage distribution (approximate normal curve) are as follows:

laudatur (L)	7	5%
eximia cum laude approbatur (E)	6	15%
magna cum laude approbatur (M)	5	20%
cum laude approbatur (C)	4	24%
lubenter approbatur (B)	3	20%
approbatur (A)	2	11%
improbatur (I)	0	5%

The relative shares of grades differ somewhat in various tests and in various examination periods.

2.2. Relating English examination results to the CEFR

2.2.1. Data

The data of the present study consist of two data sets: a) the actual student scores (15000+ students) on the four subtests of the Matriculation Examination, and b) expert ratings.

The English test used in the study comes from the autumn 2001 examination, the advanced course (10 years of English). Its composition (and scoring) was as follows:

a) Listening comprehension: 35 MC items weighted by 2, 70 points plus 5 open-ended questions answered in English, scale 0-2, weighted by 2, 20 points. Maximum score for LC = 90 points.
b) Reading comprehension: 25 MC items, weighted by 2, 50 points plus 5 open-ended questions answered in the mother tongue, scale 0-2, weighted by 2, 20 points. Maximum score for RC = 70 points.
c) Grammar and vocabulary: 40 MC items, no weighting. Maximum score = 40 points.
d) Composition (no weighting): maximum score = 99 points.

Thus the overall maximum score is 299 points.

Fourteen (14) experienced raters carried out a major task:

1. They sorted independently descriptors for different skills to 6 levels as follows: Listening Comprehension 20, Reading Comprehension 20, Writing 25, Grammar 18, and Vocabulary 18. The descriptors were taken from the validated DIALANG scales and YKI scales (Finnish National Foreign Language Certificates). The method was demonstrated with the RC descriptors and the raters got immediate feedback of how their ratings compared with the original scale values. The rest of the task was done independently at home. Our earlier experience suggested that, for enhancing the quality of the data, it would have been better to do all the rating under supervised conditions and with intermittent feedback. However, this was not possible.
2. They rated independently all (110) the Matriculation examination test items (testing the above-mentioned skills) using the CEFR 6-point scales; the instruction told the raters *to indicate for each item at what proficiency level a person would already be able to answer the item correctly.*
3. They rated a sample of 30 written compositions randomly picked but covering the whole range of proficiency, with most compositions representing the middle range of the dimension.

2.1.2. Results[6]

a) Familiarizing experts with the CEF scales of language proficiency

As in the three other projects referred to in the above, it was considered desirable as a first step to check how well the raters agreed with the original scale and among each other. The correlations between the results of *descriptor sorting task* were as follows: grammar (.941), listening comprehension (.920), vocabulary (.916), writing (.896) and reading comprehension (.868). The level of association is quite high in all skills, and shows that the raters' perception of skill progression was quite close to the original scale. As expected, the absolute agreement between experts' rating of the descriptors and their original scale values, it was much lower (60% in average) and varied in a quite broad interval (24 %-95%).

6 The results reporting here are based on Takala & Kaftandjieva (2002).

This kind of descriptor sorting helps to familiarize experts with the scales, to reveal the discrepancies in their interpretations of the scales, and to increase the level of common understanding and convergence of interpretation.

b) Reliability of ratings

The reliability of rating the items and the compositions was studied by using the alpha coefficient. The results indicated that the rating of the 30 compositions was the most reliable (α = .977, 14 raters). The corresponding figures were .852 for grammar and vocabulary items (13 raters), .737 for listening comprehension (13 raters) and .723 for reading comprehension (13 raters). The high reliability of rating compositions in absolute terms and relative to the rating of the other skills may seem unexpected. Rating written and spoken productions are usually considered problematic in terms of reliability. Yet, it is perhaps not so surprising that the agreement was highest for the composition, since (a) the raters had very extensive experience in rating compositions in the matriculation examination – many years of rating thousands, even tens of thousands of compositions, (b) the longer pieces of writing can be more easily related to the CEF scales than individual comprehension and grammar items, and (c) the raters had no previous experience in rating items. The lower reliabilities for listening and reading comprehension were due, in particular, to 3-5 raters, whose agreement with the rest was quite low.

The mean agreement among raters varied from a fairly low correlation of .631 to a relatively high correlation of .825 (mean .768). The correlation between the raters' assessment and the actual scores was higher, and varied between .667 and .893 (mean .825).

c) Intra-judge consistency

The crucial point in any test-centered standard setting method, as in the present case, is the difficulty the experts usually experience in estimating the empirical difficulty of the items.

The results of this study show that the correlation between the experts' rating of the items and their empirical difficulty is rather low (-26 in average: note that the correlation is understandably negative since an easy item with a high facility index goes with a low CEF level and vice versa), varying between –.45 and –.03, in other words, in this case – as in all other cases – we are facing again the problem of intra-judge inconsistency.

The fact that the reliability of the ratings was comparatively high (between .723 for listening and .852 for grammar) does not contradict the above mentioned conclusion. The high reliability coefficient in this case simply means

that there is internal consistency between raters in their failure to assess the empirical difficulty of the items and to categorize them in a consecutive order.

2.1.3. Standard Setting

The standard setting method applied in this case study can be classified as a test-centered continuum method and can be regarded as a modification and an extension of the classical Angoff yes/no method.[7]

After the aggregation of the individual ratings of the items for every skill, the number of items belonging to a certain level can be detected. Then, one possible way of establishing the cut-off scores is to follow the cumulative frequency distribution of the items. An illustrative example of the described procedure is presented in Table 1.

Table 1 - Setting of cut-off scores – an illustrative example

CEF levels	Number of items per level	Cumulative Frequency	Cut-off scores
A1	3	3	≤ 3
A2	4	7	[4 – 7]
B1	17	24	[8 – 24]
B2	15	39	[25 – 39]
C1	7	46	[40 – 46]
C2	4	50	[47 – 50]

In the way the procedure is described in the above, it looks deceptively easy but requires in fact attention to several points (not discussed here). The actual cut-offs are reported in Table 2.

7 For an authoritative discussion of standard setting methods and issues related to them consult Section B in the CoE Reference Supplement to the Manual written by Dr Felianka Kaftandjieva. The author has discovered through her extensive research that the method compares favourably with other methods.

Table 2 - Cut-off scores for grammar, listening comprehension and reading comprehension (listening and reading scores are weighted scores)

CEFR level	Grammar (max. 40 points)	Listening (max. 90 points)	Reading (max. 70 points)	Writing/essay (max. 99 points)
	Cut-off scores	Cut-off scores	Cut-off scores	
A1	-	-	-	Below 40
A2	2 or below	6 or below	10 or below	40 – 58
B1	3 – 18	7 – 40	11 – 20	59 – 80
B2	19 – 25	41 – 74	21 – 32	81 – 94
C1	26 – 35	75 – 84	33 – 62	95 – 97
C2	36 – 40	85 - 90	63 - 70	98 - 99

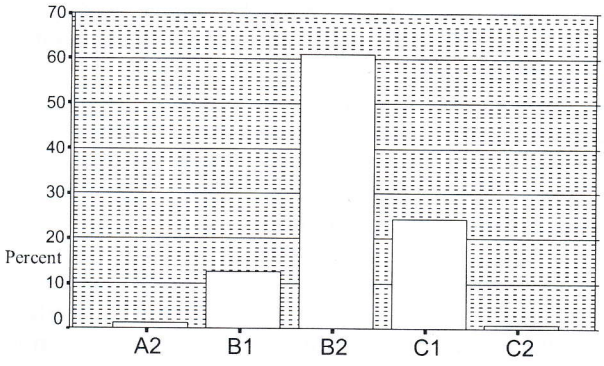
The next step after obtaining the CEFR skill-specific scores was to move toward a final aggregated level of language proficiency.

The results of the matriculation examination converted to the CEF scale of language proficiency are presented in Fig. 2.

As expected, the majority of graduates who took the Matriculation exam in the advanced English test are on level B2 or above[8]. The other hypothesis, ie. that for someone who is on level A2 or lower, it is rather impossible to pass the matriculation exam, was confirmed too.

8 Other languages do not reach the same level, but the results are not discussed here. See Tuokko (2007) for results in English concerning the end of the comprehensive school (age 15-16).

Figure 2 - Frequency distribution of the Matriculation results (n = 150370)



The link between the CEF scale and the Matriculation grades can be seen in the following figure.

As the figure indicates, the failed grade (improbatur) corresponds to the level "below B1" on the CEFR scale, the lowest pass grade (approbatur) corresponds to mid-B1, lubenter approbatur corresponds to B1/B1+, cum laude approbatur corresponds to mid-B2, magna cum laude corresponds to top-B2, eximia cum laude corresponds to low/mid C1 and laudatur corresponds to C1/C1+.

We have now arrived at a stage where we can suggest an answer to the question in the rubric of the article: Reconciling national testing assessment practices with the CEFR-linked assessment: (how) can it be done? If a national assessment system is not already closely related to the CEFR proficiency levels, a promising way seems to be to provide a conversion table, which relates the two scales to each other.

If there is an interest in not only claiming but in fact achieving increased transparency of examinations and tests, the following points need to be considered (cf. Figueras et al. 2005):

1. European performance benchmarks need to be developed for different languages. These will consist of samples of spoken and written performances, plus calibrated test items for reading and listening.
2. Examinations (and certificates) interested in the linkage need to be related to the CEFR through the application of procedures like those described in the Manual (other procedures

can be used but their validity needs to be shown). The linkage needs to be documented and reported in detail.

3. A European, or indeed international, Examination Chart may be produced showing the comparability of different examinations (i.e. the validated link between examination grades and CEFR levels). An example of what such a chart might look like the figure below.
4. The development of cooperative training networks and CEFR-linking training packages need to be discussed.
5. It should be discussed whether training in the assessment of language learning in relation to the CEFR ought to become a regular part of teacher education.

However, this scenario, which promotes the quality of relating tests, examinations and certificates to the CEFR may well be challenged by another scenario. As Figueras et al. (2005) note, it is possible that no credible system will be developed to independently validate the claims that examinations are adequately linked to the CEFR. Groups may emerge and proclaim that they have validated the participating exams' linkage to the CEFR without providing any or sufficient evidence to support these claims.

Table 3 - An illustration of a possible European chart for language examinations

CEFR level	Country A	Country B		Country C			
	Exam a	Exam b	Exam c	Exam d	Exam e	Exam f	Exam g
C2							
C1							
B2							
B1							
A2							
A1							

Linkage to the CEFR may in some contexts be required and thus deemed to have taken place without the provision of the resources necessary for an adequate linking project. Finally, and most significantly, it is possible that the aim of promoting local competence building may not be realized as the linking is out-

sourced to a small group of external consultants with minimal local involvement.

The author believes that the linking reported in this article is carried out in a thoughtful manner and it has been reported, as is required for transparency (Takala & Kaftandjieva 2002). Thus it would be possible to place the advanced English matriculation grades on the chart. It seems obvious that such a comparative chart, onto which validated examinations can be placed, is possible to make: this would greatly enhance transparency and comparability.

Bibliography

Cizek, G.J. & M.B. Bunch. 2007. *Standard setting: A Guide to Establishing and Evaluating Performance Standards on Tests.* Thousand Oaks, CA: Sage.

Figuearas, N., North, B., Takala, S., Verhelst, N. & P. Van Aevermat. 2005. Relating Examinations to the Common European Framework: A Manual. *Language Testing*, 22 (3), 261-276.

Hambleton, R.K. 2001. Setting Performance Standards on Educational Assessments and Criteria for Evaluating the Process. In: G.J. Cizek (ed.). *Setting Performance Standards: Concepts, Methods, and Perspectives.* Mahwah, N.J.: Erlbaum, 89-116.

Kaftandjieva, F. & S. Takala. 2002. Council of Europe Scales of Language Proficiency: A validation study. In: *Common European Framework of Reference. Case studies.* Strasbourg, Council of Europe, 106-129. (pdf available by request from sjtakala@hotmail.com).

Kaftandjieva, F. & S. Takala. 2003. Development and Validation of Scales of Language Proficiency. In: W. Vagle (ed.). *Vurdering av språkferdighet, NTNU. Trondheim*, 31-38 (pdf available from Takala: sjtakala@hotmail.com).

Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEF). Council of Europe, Language Policy. DGIV/EDU/LANG (2003) 5. (under revision)

Takala, S. & F. Kaftandjieva. 2002. *Relating the Finnish Matriculation Examination English Test Results to the CEF Scales*. Helsinki Seminar, June 31-July 2, 2002 (available by request from Takala: sjtakala@hotmail.com).

Tuokko, E. 2007. *Mille tasolle peruskoulun englannin opiskelussa päästään? [What level do pupils reach in English at the end of the comprehensive school?]*. Jyväskylä. Jyväskylä Studies in Humanities 69.

Webography

CRESST Assessment Glossary. 1999. From CRESST – National Center for Research on Evaluation, Standards, and Student. Testing Web site: http://www.cse.ucla.edu/CRESST/pages/glossary.htm, visited on 12 December 2003

Hansche, L. 1998. *Handbook for the Development of Performance Standards: Meeting the Requirements of Title I*. Washington, DC: US Department of Education and the Council of Chief State School Officers. From SCASS CAS Publications and Products Web site: http://www.ccsso.org/projects/SCASS/Projects/Comprehensive_Assessment_Systems_for_ESEA_Title_I/Publications_and_Products/, visited on 23 October 2003.

Kaftandjieva, F. 2004. *Standard setting. Section B in Reference Supplement to the Manual for relating language examinations to the CEFR*. Strasbourg: Council of Europe (available at: http://www.coe.int/T/DG4/Linguistic/Default-en-asp).