Relating a Reading Comprehension Test to the CEFR Levels: A Case of Standard Setting in Practice with Focus on Judges and Items

Neus Figueras, Felianka Kaftandjieva*, and Sauli Takala

Abstract: The article addresses some problems and options in setting standards on language tests and examinations. More specifically, it reports on a set of three workshops conducted in the European context where standard setting in language education typically concerns linking tests and examinations to the Council of Europe's *Common European Framework of Reference for Languages* (CEFR) published in 2001. The context of the workshops and the standard-setting procedures are described, and the results and their interpretations are discussed. The focus in the article is on judges (panels) and items, which are considered the most important determinants in valid standard setting (cut scores).

Keywords: CEFR, setting cut-off scores, standards, standard setting

Résumé : L'article adresse des problèmes et des options en relation avec la mise en place de normes d'évaluation dans les épreuves et les examens de langues. Plus précisément, cet article fait un rapport sur trois ateliers menés en contexte européen où, normalement, fixer des normes dans l'enseignement des langues implique lier des épreuves et des examens au *Cadre Européen Commun de Référence* (CECR) publié en 2001 par le Conseil de l'Europe. L'article décrit non seulement le contexte des ateliers et les procédés suivis pour fixer les normes mais discute aussi les résultats et les interprétations. Le point de mire est sur les juges et les points à appliquer, qui sont considérés comme étant les déterminants les plus importants pour fixer des normes valides (« seuils et valeurs limites »).

Mots clés: CECR, établir des seuils et des valeurs limites, normes d'évaluation, fixer des normes d'évaluation

The work presented in this article is to be considered a case study, which aims to contribute to the efforts of those involved in linking language examinations to the *Common European Framework of Reference for*

* Professor Felianka Kaftandjieva passed away in 2009.

© 2013 *The Canadian Modern Language Review/La Revue canadienne des langues vivantes*,
69, 4, (November / novembre), 359–385   doi:10.3138/cmlr.1723.359

Languages (CEFR; Council of Europe, 2001). Although this is mainly a European concern, we hope that sharing with colleagues working in non-European contexts some of the challenges faced in the process of operationalizing both the CEFR levels and some standard-setting procedures will increase their understanding of reports describing the CEFR linkage and inform their decision making with regard to doing similar linkages themselves.

There are several definitions of standard setting, which all reflect the basic notions as described by Cizek and Bunch (2007): "Standard setting is a measurement activity in which a procedure is applied to systematically gather and analyze human judgement for the purpose of deriving one or more cut scores for a test" (p. 338). In the case of the CEFR, the cut scores are set to indicate which level (A1, A2, B1, B2, C1, C2) has been reached by the test taker. When teachers grade their students, they are engaging in standard setting in their own contexts.

This article begins with the context in which the CEFR was produced in the 1990s and published in 2001 (in English, French, and German) by the Council of Europe (CoE). It is a reference tool that draws on decades of intensive European interaction and cooperation to modernize and upgrade modern language education. This is followed by an account of how the CoE has tried to assist member countries and their language education professionals in linking examinations to the CEFR levels in a valid way. The remainder of the article reports on a case study in which a group of international language professionals took part in hands-on training on how to set cut scores on a specially constructed reading comprehension test, using the CEFR as the tool for Performance Level Descriptors (PLDs).

Context of CEFR: development, challenges and response to challenges

The fields of language teaching and language assessment underwent changes in the first decade of the twenty-first century in Europe. The publication of the *Common European Framework of Reference for Languages* by the Council of Europe in 2001 not only had a considerable impact on the design of language policies and curricula, but also on teacher training, classroom materials, and assessment practices. It was in the field of language testing that the impact of the CEFR was felt most strongly, as the reference levels (A1, A2, B1, B2, C1, C2) soon became common currency. Following the frequent requests of testers and testing organizations in Europe, the Council of Europe commissioned the development of a manual for relating examinations to the CEFR levels. The development of the manual followed a seminar in

Helsinki in 2002 at which professionals in the field of language testing agreed that there was a need for research-based and documented argumentation on links to the CEFR. The Council of Europe's manual was published in its final form in 2009, and the literature published over the past decade in relation to the CEFR and the manual is impressive (Byram & Parmenter, 2012; Byrnes, 2007; Figueras & Noijons, 2009; Martyniuk, 2010).

The challenges that have had to be met in these 10 years by European testers are numerous and of different kinds. First and foremost, thorough familiarity with the CEFR – the new standard and source of PLDs – and its broad aims have had to be interpreted within different prevailing testing contexts, as existing test specifications were not based on the CEFR level descriptors. An additional difficulty was caused by the lack of readily available sample items and performances illustrating the CEFR levels that could have facilitated the process of relating existing (or new) examinations to the CEFR. There was broad agreement that such exemplars would clarify the reference standards and help the process of standard setting (i.e., the determination of cut scores that define performance standards). And, last but not least, the absence of a strong tradition of psychometrics in European language testing has been a cause of frustration for many standard setters. The final Council of Europe (2009) manual (published for consultation in 2003 and in its final form in 2009) has been criticized for an alleged lack of clarity and for making excessive demands.

Many European testers and researchers drew on the North American standard-setting literature – "the process of establishing one or more cut scores on examinations . . . distributing examinees' test performances into two or more categories" (Cizek & Bunch, 2007, p. 5) – in the interim. In many European contexts, cut scores and pass marks were decided upon on the basis of tradition and had little relationship, if any, with an explicit performance standard.

In contrast to the European scene, research and development work on standard setting had been discussed and reported in North America since the 1960s. European researchers embarked on their own research in the beginning of the twenty-first century – in most cases in foreign language education. As language education in multilingual Europe has always been a high priority in educational policy, it was hardly a coincidence that the research started in foreign language education. Undertakings such as the Dutch CEFR construct project (Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu, 2006) and a project to explore the possibility of creating a European item bank for testing reading and listening (the EBAFLS project; http://www.cito.com/

research_and_development/participation_international_research/ebafls.aspx) helped identify the challenges in operationalizing the CEFR descriptors into test specifications and test items. In addition, projects such as the ones reported at the ALTE (Association of Language Testers in Europe) meeting in Cambridge in November 2007 and at the EALTA (European Association for Language Testing and Assessment) Colloquium in Athens in May 2008 (Figueras & Noijons, 2009) served the purpose of sharing problems and possible solutions in standard-setting endeavours.

The Council of Europe itself was also very active in building up researchers' competence in standard setting through the development and dissemination of materials that could help users of the manual access sample items and performances. DVDs with oral performances rated in relation to CEFR levels were published along with the manual in various languages, and a CD was issued in 2006 with reading and listening items provided by institutions that had released CEFR-related items. This CD is available from the Language Policy Division (DECS-LANG@coe.int). The Council of Europe (2004) also published a Reference Supplement to the preliminary pilot version of the manual for relating language examinations to the CEFR that contained several chapters (currently eight) providing a more in-depth discussion of methodological issues. The revised version of the manual (2009) came out in a research climate that was quite different from the one when the preliminary pilot version was published in 2003. The new version of the manual (Council of Europe, 2009) incorporated feedback received from readers and users, as well as the conclusions of the debates during the Intergovernmental Forum held in Strasbourg in February 2007 (available at http://www.coe.int/T/DG4/Linguistic/Default_en.asp), including the new recommendation of the Committee of Ministers to its member states on the use of the Council of Europe's *Common European Framework of Reference for Languages* (CEFR) and the promotion of plurilingualism: CM/Rec(2008)7E.

In compliance with this recommendation, the revised Council of Europe (2009) manual stresses the vital importance of the need for examination providers to write up reports that document the use of procedures, discuss decisions taken, and provide evidence for the claims being made for the examination. In addition, it presents 10 standard-setting methods – many of which were not covered in the earlier version. It is well known that judges have some difficulty in judging, for instance, the probability of test takers being able to answer an item correctly at different CEFR levels. This same difficulty arises in all standard-setting work, and it has led some researchers to look for methods that do not require judges to make cognitively demanding judgements. While this is one legitimate approach, we do not think that it is the only reasonable one.

It should be noted that most of the standard-setting methods are developed for large-scale testing and may require extensive resources and sophisticated methods. Yet most assessment is done by teachers and, thus, deals with small groups. In such situations, it is essential that teachers can relate tests (e.g., texts and items) to the CEFR level reasonably well. We believe that teachers as well as panel judges can improve their ability to judge items through focused training and through receiving feedback. Test developers should also improve their ability to target tests and items at particular levels, which can be done by getting feedback on how their preliminary level estimates match the empirical results. As the workshops were arranged to provide this training, the focus of this article will be on exploring various aspects of the role of judgements in standard setting. Accordingly we

- report on the preparatory work carried out before and during three standard-setting workshops organized to collect the necessary data,
- discuss the challenges encountered in simultaneously operationalizing the different CEFR levels (A2–C1 range) in a scale,
- justify and problematize decisions taken to be able to assign items to levels, and
- present the results of the judgement sessions at the workshops using the Basket procedure as recommended in the Council of Europe Manual (2009, p. 91).The Basket procedure is basically an item-descriptor matching method. We compare the results of using that procedure with those of three other methods (Angoff method, contrasting groups method, and borderline groups method).

The test booklet referred to in this article, and used for the project, is publicly available, as is the corresponding statistical information (by requesting it from the authors). The reading comprehension test items can be used as exemplars of A2–C1 CEFR levels, and also for replication purposes in other standard-setting seminars. We look forward to hearing about any such use of the test.

Method

Participants

The main purpose of the project was to provide a concrete case study in a European context in which standard-setting procedures were applied in a principled manner, and to make available illustrative reading comprehension items at some CEFR levels that were documented in a transparent fashion and empirically validated. Haertel and Lorie (2004) have noted that

> [t]he problem of standard setting is sometimes viewed as no more than the problem of choosing a cut score, with scant attention to the performance standard. It should be clear, however, that a standards-based score interpretation is not defensible unless the cut score and the performance standard correspond to one another. (p. 2)

To fulfil this stated objective, a first workshop was organized in Barcelona before the EALTA conference in Sitges in May 2007. The success of the Barcelona workshop, mostly due to the novelty of the topic of standard setting in Europe and also to the need for training, led to the organization – on request – of two additional workshops, one in Turku (Finland) in 2007 and one in Budapest (Hungary) in 2008. A more limited (mainly dissemination) workshop was also held in Siena (Italy) in 2011.

The focus in this case study is on judges and how they rate items. We strongly agree with Reckase (2010) who states that

> [a] second thing I know is that *test items are complicated*. This might not be as obvious as the complexities of people because we tend not to study items as much as we do the people around us. I consider test items as being somewhat like equivalent to little poems. They are a constrained literary form that requires careful choice of words and clear communication in a limited space. It would be better to identify people who have demonstrated good item writing skills, rather than expect that with minimal training to do this creative job. (2004, p. 4)

As mentioned above, the data reported here come from the three first workshops. A general overview of steps taken is given, and the activities carried out are described in more detail in the following sections. They illustrate the key steps in a test development cycle and the key steps in standard setting, as listed in the 2009 version of the CoE manual and in Cizek and Bunch (2007, pp. 35–37).

The exam booklet: development, trialling and analysis

For reasons of practicality and ease of replicability, the researchers decided to produce a test that would meet the requirements of the project in terms of ownership (the test would belong to the project); quality (it would be valid and reliable); level (it would aim at the most common exam range in Europe, CEFR levels A2–C1); content and target audiences (overall reading comprehension for young adults and adults in the context of English for general purposes); length (about

40 items, 45–60 minutes); text type, topic, and domain (texts ranging from 15 to 515 words, covering everyday and more academic topics, in different domains); and item type (multiple matching, true and false, and multiple choice test types). Due to space limitations, the development of the test is not discussed further in this article.

The authors had obtained permission from the Finnish Matriculation Board to use some items from exam sessions already released. The Finnish items were reported to be aimed at the B1–C1 range of the CEFR levels (Kaftandjieva & Takala, 2002). Additional items expected to cover A2 and the lower band of the B1 level were developed by the authors on the basis of the CEFR descriptors on reading scales, and by drawing heavily on the work of the Dutch CEFR Construct Project (Alderson et al., 2006). A booklet containing a total of 48 items was put together and piloted in Cataluña, Spain, and in Finland by over 300 students from upper secondary schools (aged approximately 18 years old) and young adults in higher secondary schools and language schools. Data were collected at the item level (student responses) and student level. Teachers were given the CEFR reading descriptor pool and asked to assign a CEFR level to their students. The final research project booklet was assembled with 41 items, and the results and statistics reported are based on those 41 items.

Test statistics

The analysis of the item booklet responses was based on classical test theory, and the results (see Table 1) showed that the quality of the test was reasonably good for both sub-samples as well as for the total sample.

The item statistics are also similar for the two samples. The Spearman correlation between item difficulty for both samples is 0.72, and the paired samples *t*-test indicates that there is no statistically significant difference between item difficulty for both sub-samples (t = 1.49; p = .144). The summary statistics for the test booklet are reported in Table 2.

Table 1: Test statistics of the total sample

Test statistics	Total
Sample	334
Items	41
Mean (raw score)	26.60
SD (raw score)	7.04
Reliability (α)	0.86
SEM	2.63

Table 2: Item statistics

Item statistics		Total
Difficulty	Min	29%
	Mean	**65%**
	Max	97%
Discrimination	Min	0.14
	Mean	**0.33**
	Max	0.59

Data analysis at the item level was also very relevant for the project, as it was used extensively to throw light on the results of the judgements and to help the decision-making process. Reckase (2010) argues persuasively for the importance of paying more attention to item development and analysis.

Standard-setting process

Any standard-setting process implies the operationalization of a verbal description or standard (in this case, a collection of CEFR level descriptors) into test content (in this case, items) to be able to translate a construct into a numerical score. Cizek and Bunch (2007) aptly summarize the ingredients involved in the process and its complexity:

> [M]uch more is required of a defensible standard setting process than choosing and implementing a specific method, and any listing of steps masks the reality that the key to successful standard setting lies in the attention to decisions about many consequential details. For example, those responsible for standard setting must attend to identification and training of appropriately qualified participants, effective orientation and facilitation of the standard setting meeting, monitoring and providing feedback to participants, and well-conceived data collection to support whatever validity claims are made. (p. 35)

The following sections provide an account of how such requirements were addressed in practice in the workshops.

Workshops

The workshops were initially targeted at EALTA members interested in standard setting and, therefore, restricted primarily to professionals in the field. Each workshop lasted for two and a half days. Due to the tight schedule, it was necessary to ask participants to do some preparatory work before the workshop, and also some additional activities at the end of each day of the workshop.

We were well aware of the great variety of standard-setting methods and decided to apply four methods in which the judgements of the panel members play a vital role, as explained earlier. One of the advantages of the methods we chose is that they do not require extensive preparatory work.

Panel (judges)

All participants in the workshops acted as judges. The majority of them were testers with some teaching expertise. Background information was collected and is summarized in Table 3.

The judges in the three workshops (83 in total) constitute a fairly unique pool of participants from the viewpoint of familiarity with the field, expertise, and international background. This has implications for estimating the validity, representativeness, and generalizability of their judgements.

Familiarization and training

The importance of training in the widest possible sense, not only in relation to the interpretation and operationalization of the standard, but also in relation to the standard-setting process, is well recognized in the literature (e.g., Berk, 1995; Cizek, 2001; Cizek & Bunch, 2007). The two versions of the manual (Council of Europe, 2003, 2009) also stress the crucial importance of training, which is described in two separate phases: familiarization with the standard (i.e., the CEFR descriptors at different levels across the proficiency continuum) and standardization of judgements (operationalizing level descriptors into items).

A considerable amount of time was devoted to training, both before and during the workshop, and Kaftandjieva's (2004, p. 29) recommendations regarding planning, organizing, and conducting training were followed. Prior to attending the workshop, participants were asked to do some preparatory work and activities. One month before the start of the workshop, participants were sent full instructions and documentation, with background reading tasks and work to be done plus some other homework. Participants were asked to read the most relevant sections in the CEFR and the manual to decrease the time needed for them to become familiar with the CEFR during the workshop.

Table 3: Background of the pool of judges (panels)

Seminar	Number of judges	Nationality	Experience in item writing	Experience in standard setting	Teaching background
Barcelona	34	18	50%	62%	> 5 years
Turku	24	2	50%	62%	> 5 years
Budapest	25	1	48%	81%	> 5 years

They were also asked to read two chapters from a recent book on standard setting (Cizek & Bunch, 2007). Finally, to prepare them to be able to estimate the difficulty and level of reading items, they were asked to access the website for the Dutch CEFR construct project and work through the training module to get used to taking into consideration the interactions in estimating the difficulty of texts, items, and tasks. The rest of their homework consisted of rating 56 CEFR reading comprehension descriptors into the six CEFR levels without consulting CEFR-related materials. They sent in their ratings for analysis a week before the workshop, and their results were integrated into planning the familiarization session on the first day.

At the workshop proper, familiarization and training took most of the first day. Familiarization included an introductory review session on the main aspects of the CEFR, and continued with a detailed discussion and feedback on the results and analysis of the descriptor ratings. Discussion of the ratings of the most problematic descriptors (those which the participants had misplaced or for which they had rated descriptors at more than two different levels) helped identify the salient characteristics of the CEFR levels.

Training with items (called "Standardisation" in the 2003 Council of Europe manual) followed, and the participants were presented with reading comprehension items from the Council of Europe CD (2006), which they were required to judge in relation to the question presented in the manual (Council of Europe, 2003) that corresponds to the Basket procedure method: "At what CEFR level can a test taker already answer the following item correctly?" (p. 85).

Participants had the opportunity to discuss how they estimated difficulty levels and to analyze why their perceptions did not always match the empirical difficulty of the items. This activity was helpful in better understanding the meaning of the CEFR descriptors in relation to text and item characteristics, and it highlighted possible reasons for discrepancies between perceived difficulty levels as opposed to empirical difficulty levels.

Standard setting

There is an extensive (and growing) literature on standard-setting methods and their characteristics (e.g., Berk, 1995; Hambleton & Pitoniak, 2006; Kaftandjieva, 2004, 2010; Mills & Melican, 1988), and the interpretation and adoption of the cut scores arrived at through such methods has been widely discussed. New methods keep emerging, with some experts advocating the use of more than one method despite the cost and additional effort. Cizek and Bunch (2007) argue that context-related method selections, rigorous applications, and thorough documentation should suffice (the present authors disagree). In addition, they rather strongly dismiss multiple methods in the following quote:

> A man with a watch knows what time it is. A man with two watches is never sure. . . . Because there is no equivalent of an atomic clock in the field of standard setting, our recommendation is simply for practitioners to invest in a single watch of the greatest quality given available resources. (Cizek & Bunch, 2007, p. 319)

The authors decided to use the Basket procedure, commonly used in the first stage of standard setting with the CEFR in Europe, as the first standard-setting method in the workshops, but to supplement it with three other methods (contrary to the view by Cizek & Bunch, 2007). Prior to the judgement session, the participants were asked to respond to all the items in the test booklet, just as a student would, and the participants were given sufficient time to do so. Once finished, they were asked to judge the level of each item in response to the question: "At what CEFR level can a test taker already answer the following item correctly?"

The judges (panellists) got feedback on the second day of the workshop, before embarking on a second (widely used) standard-setting procedure (a modified Angoff). For this method, the participants were given the following instruction and then asked to judge items: "Out of 100 examinees bordering on A2 and B1, B1 and B2, and B2 and C1, how many will answer each of the following items correctly?" The results of the judgements, using these two methods and focusing on the items (test-centred), was contrasted at the end of the seminars with the results of two other standard-setting methods. The latter used empirical data (examinee-centred) rather than judges. The implications of the results for setting cut scores will be discussed below.

Judgement results

All workshop participants rated each of the 41 items in the test booklet on the basis of the 56 CEFR reading comprehension descriptors ranked in the homework exercise and discussed in the familiarization session. They also wrote their judgements on a rating form, which was then collected for analysis without further discussion. Each level rating (A1, A2, B1, B2, C1, C2) was coded (A1 = 1, A2 = 2, etc.), so that all 83 participant judges' judgements could be tallied to calculate the CEFR level of each of the 41 items. The analyses focused on the degree of agreement across workshops and judges for the different items, the distribution of the ratings along the CEFR level continuum, and the consistency of the judges' ratings. The results of these analyses were expected to throw light on the assignment of items to CEFR levels.

Variability – interjudge agreement

The frequency distributions of ratings per item were calculated to study differences across the ratings of judges and the CEFR, and the results showed that the majority of the items were assigned to the same CEFR level by more than 50% of the judges. The percent of perfect agreement (% of judges assigning an item to the model CEFR level) varied across items, ranging from 40% (for Items 27 and 29) to 61% (for Item 2). All but one item (Item 21) were assigned in two consecutive levels by more than three quarters of the judges; additionally, the ratings of at least 90% of the judges were in three consecutive CEFR levels for all items (% of adjacent agreement).

On the whole, the range of the ratings for all items varied between 3 and 5, which meant that the ratings for all items fell into at least four consecutive CEFR levels. Three items (Items 8, 19, and 29) had a range of ratings equal to 5, meaning that the ratings for these three items covered the whole range of the CEFR scale, but this range was a measure of variability greatly affected by outliers. The standard deviation (*SD*), on the other hand, was between 0.57 and 0.91, with an average of 0.75, which confirmed that the majority of ratings were in three consecutive CEFR levels for all test items.

It is worth noting that interjudge agreement can be considered satisfactory, as 40 out of 41 items were rated in two consecutive levels by more than three quarters of the judges. In addition to the percentage of exact agreement, another index of variability was calculated: Aiken's (1985) index of homogeneity (*H*). It is similar to *SD*, but more appropriate for ordinal levels of measurement. Aiken's index of homogeneity is a measure of internal consistency for rating data, and "when computed across raters, *H* is a measure of agreement among the raters (or judges) as to how a specific item should be rated (or judged)" (p. 140). In our study, *H* (which ranges from 0 to 1) is statistically significant for all 41 items, varying between 0.63 and 0.78, with a mean of 0.69.

Table 4 presents the statistical summary for all indices of variability discussed so far. The items in boldface are items with more homogeneous ratings, while the rest of the items are the more heterogeneous ones.

The inter-judge rank correlation was always positive and varied between 0.02 and 0.95 with an average of 0.67. In fact, for more than 75% of pair comparisons, the correlation was equal or greater than 0.60.

Table 4: Variability indices

Indices	Min Value	Item no.	Mean	Max Value	Item no.
% agreement – 1 level (Exact)	40	27, 29	50	61	**2**
% agreement – 2 levels	72	21	83	95	**16**
% agreement – 3 levels (adjacent)	90	29, 30, 40	96	99	**1, 10, 16, 18, 25**
Range	3.0	**23**	3.5	5.0	8, 19, 29
IQR	0.0	**2, 19**	1.0	2.0	21, 27
SD	0.57	**16**	0.75	0.91	29
Aiken's *H*	0.63	**29**	0.69	0.78	16

IQR = Interquartile range

Intra-judge consistency

On average, the items assigned by the judges to the lower CEFR levels are easier than those assigned to higher levels, which confirms that the judges had – on the whole – an adequate perception of item difficulty.

Spearman's rank correlation between item difficulty (% correct) and CEFR level rounded mean of judgements was 0.71, and Spearman's rank correlation and the mean of judgements (without rounding) was even higher (0.81). In fact, the intra-judge consistency of the aggregated rating of judges on items was comparable to the intra-judge consistency of the teachers who rated their students for the test during the pilot phase. This leads to the conclusions that (a) the level of intra-judge consistency for the three workshops is rather high (although, unfortunately, we are not aware of published data to compare our outcomes with), and (b) the mean of judgements (without rounding) seems to be more informative than the rounded mean, and should be taken into account.

Figure 1 shows how the item judgements relate to the empirical difficulty of the items. There is a big overlap between the confidence band for level B1 and the confidence bands for its adjacent levels (A2 and B2). This may be due to the fact that B1 is the level to which the judges assigned the smallest number of items (5). The number of items assigned to the other levels was A2 – 10 items; B2 – 19 items; and C1 – 7 items. It is important to note that, although the C1 band is much narrower, its number of items is not much larger than for B1 (7 and 5, respectively).

A pattern emerges from this whereby judges find it much easier to judge either very easy or very demanding items than to judge mid-level items. This observation also commonly arises in relation to rating the level of *can do* descriptors. One conclusion that may be drawn

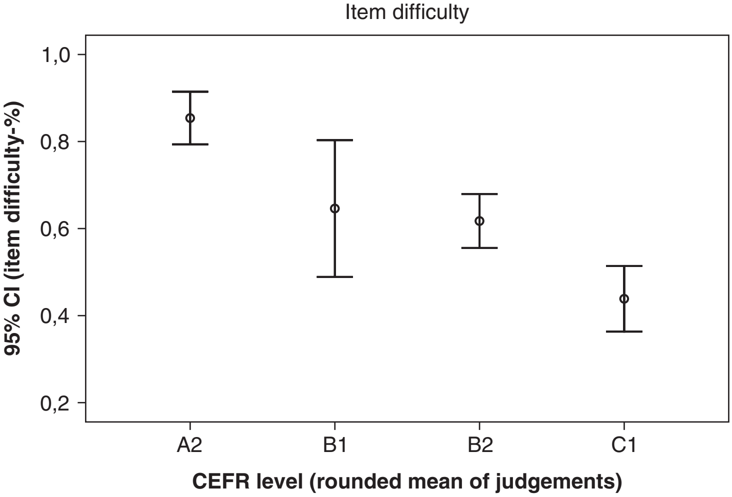

Figure 1: Confidence bands for the assigned CEFR levels

It is probable that some of the items assigned by judges to level B1 do not belong to this level, and that some of the items assigned to the adjacent levels in fact belong to level B1.

from this is that it is advisable to devote more attention to A2–B2 levels in rater training, as the judges need to engage in detailed discussion of what the criterial features of different levels are (cf. Hawkins & Filipović, 2012).

With this information, it is not possible to base the cut score decision on the level assignment reached after the judgements. Figure 1 shows very clearly that the judges did not identify a clear boundary between A2 and B1, or between B1 and B2. Level B1 (which corresponds to the Threshold level as defined by the Council of Europe in the late 1970s, and one of the most commonly referred to levels in the literature and in real life) seems to be undefined and underrepresented. To further explore the inconsistencies between the judgements and the empirical levels before assigning final levels to items, two scatter plot graphs were produced.

Figure 2 presents the scatter plot of items based on their empirical difficulty (% correct) and the mean of the judgements. The markers are based on the rounded mean of judgements and the shaded items are, in our view, misplaced by judges (5, 7, 8, 16, 17, 19, 20, 21, 26, 27, 28, 29, 31, 33, 34, 39).

This graph is informative as it depicts very clearly the mismatch – at the item level – of some judgements, and the corresponding

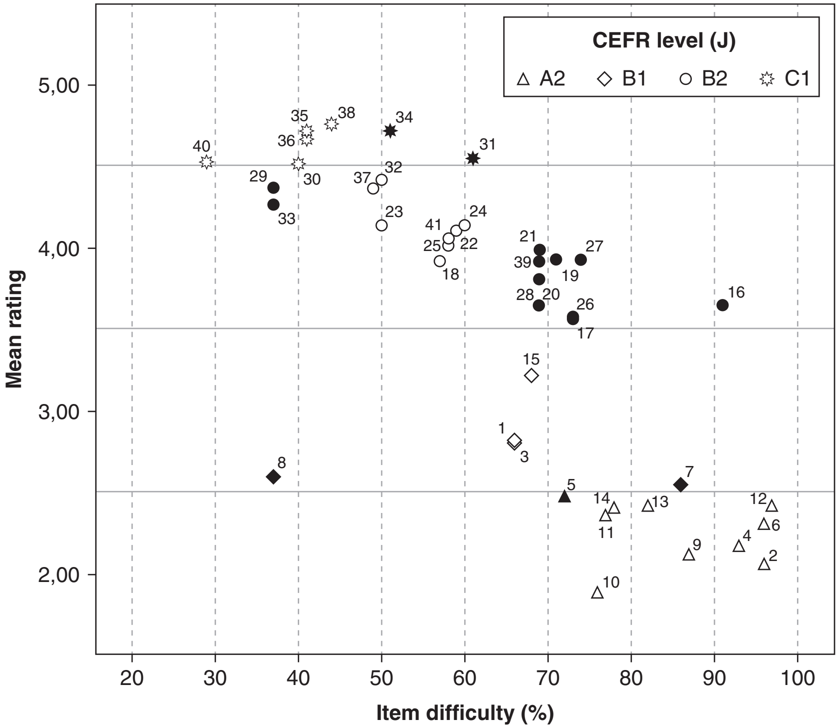

Figure 2: Item difficulty (facility)/mean rating

empirical difficulties. At this point, it is necessary to explore possible explanations for the misplacement of some items as well as to consider how adequate arguments can be presented for each item's CEFR level allocation. Figure 2 also shows that if standard setting were done on the basis of these judgements, and if cut scores were decided on the basis of the levels assigned by the pool of judges, the validity of the outcome for test takers would be questionable, especially for levels B1 and B2 (most test takers assigned a B1 level probably could have correctly answered items assigned to B2, which are in fact occasionally easier than the ones assigned to B1).

Data such as we have collected can be used to explore judges' perceptions of item difficulty, but this would require a separate study. We reiterate that judgements play an equally important (pre-data) role in enhancing the content relevance of tests and examinations, and in providing an opportunity for feedback (pre-data ratings vs. post-data ratings).

Assigning final levels to items

The figures in the preceding section show how, despite the quite thorough training undertaken, we consider 39% of the items to have been misassigned. The pool of 83 judges assigned 16 out of the 41 items in the test booklet to levels that do not match their real empirical difficulty, and in some cases such misplacements were extremely consistent! This seems to confirm the reported difficulties of the judges to assess the real difficulty of items and their tendency to overestimate the difficulty of multiple choice items (this is the case in 13 out of the 16 misplaced items). There was thus a need to find a coherent interpretation of item characteristics that, related to the judgement levels, could justify a final CEFR level assignation to items.

Reckase (2009b) has indicated the importance of the consistency of "the standard" and its operationalization. In fact, he suggests that the research on standard-setting methods per se is off the mark: "It is my belief that the inconclusive nature of standard setting research is due to the lack of a coherent theory of standard setting that can guide the research and provide a structure for interpreting the results" (2008, p. 13).

To explore possible arguments for modifying the judges' decisions, the item level statistics discussed so far were revisited for the allegedly 16 misassigned items. The discussion in this section presents – on the basis of test content and test results – possible reasons for the inaccuracies in the difficulty estimation by the pool of judges in the project. On the basis of this discussion, it is likely that decisions may be made on more solid grounds. Due to space considerations, we cannot fully present the detailed item analysis that we have carried out; rather, the discussion will centre on one short example testlet.

"Getting tough on planning"

The Government issued tough new planning rules this week to protect the countryside from urban sprawl. The rules instruct local councils to prevent building on greenfield sites until options for building on previously developed land have been exhausted. With 4.4 million homes needed by 2016, the guidelines also call for smaller houses with less parking space. The Council for the Protection of Rural England hailed the new approach as a "historic watershed."

16. What is the purpose of the new rules?

A. To protect the rights of home owners
B. To use land more sensibly
C. To improve relations with the environmentalists

17. Why do the rules appear justified?
A. Houses need larger parking space
B. The need for building sites is great
C. There is a threat of water running out

Items 16 and 17 are discussed together because they belong to the same testlet, based on a short text on government policy and involving two multiple choice items. The difficulty values of the two items are .91 and .73, respectively.

The judges "stayed at the level" for both items, but in this case they overestimated their difficulty, possibly influenced by the abstract nature of the text, its heading, and some infrequent words such as "sprawl," "hailed," and "watershed." In fact, neither the text type nor the language in the text is considered in the CEFR to reflect A2 reading comprehension descriptors, which focus on "familiar names, words, and very simple sentences" found in "notices and posters, catalogues, very short, simple texts, simple everyday material, advertisements, prospectuses, menus, timetables, and short simple personal letters." The judges, who took their work seriously, focused on the text and referred to the CEFR descriptors during the judgement sessions. They were, however, unduly influenced by the low text demands of CEFR A2 descriptors. The judges should have also more carefully considered the cognitive demands of the items themselves, as recommended by the Dutch CEFR Construct project (Alderson et al., 2006). These demands were, in fact, rather low. Both items asked a "why" question, which could be answered by identifying the relevant information in two clearly marked and quite simple sections of the text: "protect the countryside" and "4.4 million homes needed by 2016."

The considerations presented thus far could be expanded to cover all possible facets of difficulty to facilitate the interpretation of human judgements, but we hope that what has been noted so far is enough to explore possible interpretations for the decisions collected at the three workshops.

Final assignment of levels to items

On the basis of the discussions in this section, we were faced with the need to take decisions on CEFR item level labelling. In some cases, as seen, we considered that there were sufficient arguments to make defensible decisions on item level labels, but in other cases there were no arguments that allowed for a decision on whether to follow the judges' assigned level or the empirical difficulty. Taking into consideration test statistics (see Table 1), and also bearing in mind the findings of the Dutch CEFR construct project (Alderson et al., 2004) on the

Table 5: Final item levels (the numbers represent the order of the items in the test booklet)

	Based on the judgements (rnd)	Total	Based on the judgements, and moderated by empirical difficulty	Total
A2	2, 4, 5, 6, 9, 10, 11, 12, 13, 14	10	2, 4, 7, 6, 9, 10, 11, 12, 13, 14, 16	11
B1	1, 3, 7, 8, 15	5	1, 3, 5, 15, 17, 19, 20, 21, 26, 27, 28, 39	12
B2	16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 32, 33, 37, 39, 41	19	8, 18, 22, 23, 24, 25, 31, 32, 34, 41	10
C1	30, 31, 34, 35, 36, 38, 40	7	29, 30, 33, 35, 36, 37, 38, 40	8

difficulties that judges have in assigning CEFR levels to items, the authors assigned final levels to items by adjusting judgements to empirical difficulty. Our results are presented in Table 5.

The implications of the changes in the table above are very important, especially in the B1/B2 range, as can be seen in the contrast between Figure 3 and Figure 1.

For some, the above reasoning and decision making may seem iconoclastic behaviour, especially after all the efforts made to organize, collect, and document the valuable judgements from the workshop participants. We think that, considering that judgement is an inherent part of all standard setting and that such judgement is bound to show some variation, the decisions are plausible and defensible, as they correspond much better to the picture presented by the item analysis based on real student performance.

Setting cut scores: the final step in standard setting

Having decided on item level labelling, the final step in standard setting is to determine *cut scores,* which are score points that divide the examinees who know and are able to do what is stated at a certain level from those who do not and cannot. In linking tests/exams, the number of cut scores is one less than the levels aimed at. In the case of the CEFR, for example, 2 levels presupposes 1 cut score; 3 levels, 2 cut scores; 4 levels, 3 cut scores; 5 levels, 4 cut scores; 6 levels, 5 cut scores. The number of credible cut scores is crucially dependent on test reliability.

As stated earlier, there are dozens of standard-setting methods and new ones emerge every year. The revised manual (Council of Europe, 2009) describes 10 methods that seem relevant for CEFR-related standard setting. The Basket method (the simplest method of standard setting) developed for the EU's (2001) DIALANG project (as cited in

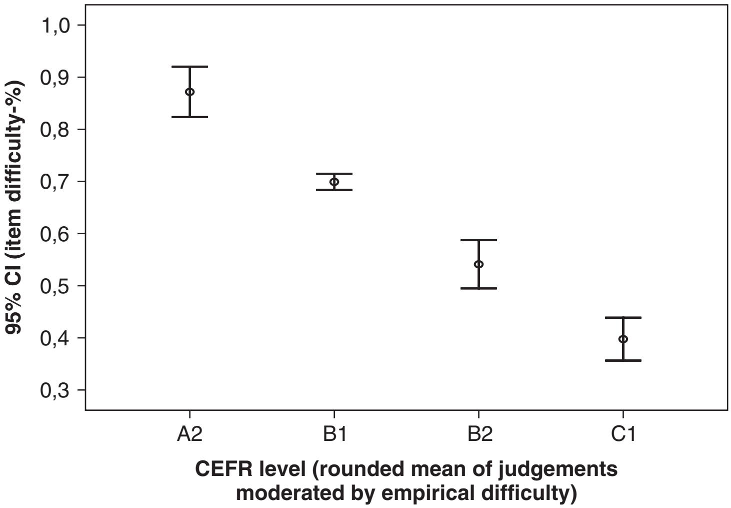

Figure 3: Confidence bands for the assigned CEFR levels

Council of Europe, 2009) was the first method used in this research. Each cut score equals the sum of items that a rater considers the examinees to be able to answer correctly at all levels of competence that are below the respective cut score. Only one round of ratings is needed, and cut scores are based entirely on judges' ratings (although this has serious implications for rater training; see Kaftandjieva, 2010, pp. 122–135, for a critical discussion of this issue). In addition, at least one item at the level below or above must be considered to have been answered correctly. Cut scores can only be established if the number of items rated as below the level in question is neither zero nor the maximum number of all items. In practice, it may turn out that it is not possible to precisely indicate the lower and upper ends of the scale; it may only be possible to state that, for example, scores ranging from x1 to y1 represent level A2 or below, and that x2–y2 is at level C1 or above.

In the Modified Angoff method, one of the most widely used methods and the second method in the study (see the section "Standard setting" above), judges are asked to indicate the proportion of the threshold (minimally acceptable) at which examinees should be expected to correctly answer each item for a given performance level. The judgements are aggregated over items and judges. The outcome is typically a summed score on the set of items that represents the cut point.

These two methods are test-centred standard-setting methods. Two examinee-centred methods were also used but not elaborated upon here: the Contrasting Groups method and the Borderline Group method. The use of these two methods requires the rating of examinees in terms of the CEFR levels and requires test scores from a sample of students. In practice, only the teachers who have taught the examinees can do this. The ratings for the Contrasting Groups method and the Borderline Group method were provided by the Finnish and Catalan teachers of the test-takers at the time of the piloting of the test used for the project.[1] A fuller description of these methods can be found in the Council of Europe's (2009) manual (pp. 67–68).

Table 6 shows the cut scores set by applying the two examinee-centred methods used in the workshops (Basket and Modified Angoff) and the two student-centred methods on the basis of teacher ratings of examinees (Contrasting Groups and Borderline Group) and presented to the judges at the end of the workshops. The last row includes the cut scores that should be applicable on the basis of the information presented so far: the judgements made by the 83 participants in the three workshops, the empirical data, and the judgements that teachers participating in the pilot made of their own students.

Translating this into score bands for the levels covered gives the following distribution of levels:

- A2 or lower: scores 0–11
- B1: scores 12–23
- B2: scores 24–33
- C1: scores 34–41

Table 6 shows two common observations: (a) different standard-setting methods lead to somewhat different cut scores and (b) the Basket method tends to produce lower cut scores than do the other methods. Kaftandjieva (2010, p. 133) notes that the Basket method has some distinct problems: lack of sufficient consistency with empirical

Table 6: Cut scores set using four standard-setting methods

Standard setting method	A2/B1	B1/B2	B2/C1
Basket(test-centred)	10	15	34
Modified Angoff(test-centred)	14	26	34
Contrasting Groups(student-centred)	19	26	32
Borderline(student-centred)	19	28	32
Final proposal for cut score	11	23	33

data, considerable distortion of the cut scores toward the end of the interval in which the test results vary, and large standard errors of the cut scores. Kaftandjieva therefore recommends that, due to problems of internal validity, the Basket procedure should be used only in tests intended for formative purposes. One central check concerns the size of the standard error of cut scores (SE_J) in relation to the standard error of measurement (*SEM*). There are several proposals of varying strictness/leniency:

- $SE_J \leq 2$ items (out of 100) (Norcini, Lipner, Langdon & Strecker, 1987)
- $SE_J \leq$ ¼ *SEM* (Jaeger, 1991)
- $SE_J \leq$ ½ *SEM* (Cohen, Kane & Crooks, 1999)
- The most lenient criterion: $SE_J < SEM$

Using the Basket ratings data, it was found that the standard error of the mean (SE_J) was 0.7 for the cut score A2/B1, 1.0 for B1/B2, and 0.8 for B2/C1. The corresponding *SEM*s were 2.6 in all three cases (and conditional standard errors of measurement, *CSEM*, were 2.3, 2.9, and 2.6, respectively). Thus, the SE_J was always smaller than one-half of *SEM* or *CSEM*, and the cut scores fulfil the quality criterion.

The validity of the cut scores needs to be analyzed not only by taking into account standard errors, but also from several validity viewpoints. Common forms of validation checks and procedures in standard-setting validation were carried out (Cizek and Bunch, 2007; Council of Europe, 2009). Feedback by the participants on a questionnaire adapted from Cizek and Bunch (2007, p. 62) provided strong support for procedural validity.

With regard to internal validation, it was found that

- *consistency within the method* proved adequate. The *intrajudge consistency* r (mean) was 0.48 (66% all cases were > .50), and
- *decision consistency* was, not unexpectedly, the highest between the two examinee-centred methods (the Borderline Group method and the Contrasting Group method) based on the same teacher rating data: 92% of all cases. The Basket method was the least satisfactory; its decision consistency with the Angoff method and the Contrasting Group method was 44%, and only slightly higher with the Borderline Group method, at 46%. The Angoff method had clearly higher decision consistency with the examinee-centred methods: Borderline Group 66% and Contrasting Group 75%.

With regard to external validation, the CoE manual stated that "the essential requirement for real validation is the availability of a criterion test which can be trusted as a good indicator of the CEFR levels" (Council of Europe, 2009, p. 120). We regarded this as an unwarranted claim that also raises the question of how to produce such a valid test, and the even more difficult question of how to access such a test and get permission to use it. In fact, there are other ways to explore external validation: by using other standard-setting methods and comparing the outcomes, by using other sources of information, and by analyzing the reasonableness of the cut scores.

We used four different standard-setting methods, representing both test-centred and examinee-centred methods, and reported on the results. We were also able to draw on prior work on standard setting in the two contexts (Finland and Catalonia) and use that to reflect on the cut scores. The panel members also discussed the reasonableness of the cut scores on the basis of their own contextual knowledge. On the basis of all this external validation, the final cut scores were determined.

Conclusions

The project reported on in this article had a specific limited goal: to provide hands-on experience in some of the many types of approaches in standard setting, using a purposely designed test of reading comprehension. The focus was on the training of judges in assessing test items against the levels of the CEFR and producing a set of cut scores. We acknowledge that the analysis of data could be continued and we could define other objectives. We will not attempt to define these options here, nor to spell out the obvious limitations of the project. Instead, we will summarize what we consider the main points that bear on different aspects of the standard-setting process that all testers should address and that teachers should also be aware of when faced with the need to decide on final scores in their classroom assessments.

- The standard (i.e., the CEFR level descriptors in this article) and the importance of its operationalization(s) (the items, the reading comprehension test) have been emphasized, especially highlighting problems with descriptors that do not help identify "the level" of proficiency required by the text or an item. This issue has been researched elsewhere (see, e.g., Alderson et al., 2006, and Fulcher, 2004, for complementary, albeit opposed views) and may be highly controversial, but the experience gained in this project suggests that more work on the interpretation of the standard, plus more elaborated descriptions (test specifications) and additional illustration with calibrated samples (items) are needed.

- The lack of precision of human judgement in relation to the estimation of item difficulty has also been studied. After thorough (and well documented) training, judges get more consistent, and the agreement indices improve, but not always in the right direction. However, this does not mean that judges are of limited use. Judges do play a key role, and their training needs to be improved. We believe that judges not only should act as data providers, but also need to be given feedback on their ratings and need to have an opportunity to discuss their ratings; no matter what kinds of sophisticated statistical analyses may be used (IRT – Item Response Theory), judges are needed to interpret the operationalization of levels to give meaning to any testing instrument.
- We have also argued that teachers usually have very limited data to draw on and need to rely more on their own ability to judge items. Thus the Basket method, a term used in Europe for an Item-Descriptor Matching method, is likely to be the most appropriate in their situation. The interpretations presented in this article give possible reasons for the overestimation or underestimation of difficulty, and highlight the need to pay attention to the influence of the text and the other items in the testlet in the estimation of difficulty. Verbal protocol analyses of judges' behaviour would also be very useful. Teachers are in an ideal position to discuss items with their students, and, armed with this feedback, they can improve their skill in rating item difficulty in terms of a standard. In the case of productive skills, teachers' tasks can be much facilitated by exemplars ("benchmark" performances that are provided with rationales for level assignment).

To conclude, we feel reasonably satisfied with the results of the case study project, and our initial hesitation about "meddling with" the judgements disappeared when we became aware that it was, in fact, our duty to use the rich data available (this included data at item level) to its full potential, to take a detailed look at the key component – the items. Moderating judgements (i.e., adjusting item levels) on the basis of the data from a test of good quality is not only appropriate but useful and even necessary.

The work reported in this article has demonstrated that – indeed – there is no gold standard out there to be found, in the same way that there is no ideal methodology or teaching approach. We have also emphasized that the contexts in which cut scores are to be set can vary considerably. In high-stakes situations, the standards have to be constructed and well documented. The basic requirement is that the chosen method be appropriate for the context. We hope that the reader will not misinterpret Kaftandjieva's (2004) perceptive and sobering conclusion in Section B of the Reference supplement:

> There is no "gold standard," there is no "true" cut-off score, there is no best standard setting method, there is no perfect training, there is no flawless implementation of any standard setting method on any occasion and there is never sufficiently strong validity evidence. In three words – nothing is perfect. (p. 31)

We would rather end on a positive note, suggesting that if sufficient work is done on the elaboration and illustration of the standard, standard-setting endeavours will be not only less costly but more efficient. Quoting Reckase and Chen (2012):

> Standard setting is a complex process with many components. A poorly designed process with insufficient time for implementation will not likely produce credible standards. The required resources need to be put into the process so that it can produce defensible recommendations. (p. 163)

We agree with this recommendation for large-scale, high-stakes standard setting. We believe it also basically applies to classroom assessment as well: If teachers make the effort to assess their tests (items) in sufficient detail, their skill in setting standard-related cut scores will improve for summative purposes, which will also enhance their chance of using the tests for more formative purposes, assessing the tests/items for improving learning.

Kaftandjieva's posthumous monograph (2010) shows that new methods can be developed and their relative merits can be compared empirically. Reckase has also continued to work on more sophisticated approaches to standard setting (2009a, 2010), a clear indication of the possibility of progress to be made.

Correspondence should be addressed to **Neus Figueras**, Departament d'Ensenyament, Generalitat de Catalunya, Via Augusta 202. 08021 Barcelona, Spain. E-mail: nfiguera@xtec.cat.

Note

[1] The instruction given to teachers was as follows: "Based on your experience, please assess in terms of CEFR the level of language proficiency in reading comprehension for every student in your class. Please use a

single level mark *only* in case you are almost 100% certain that the student belongs to this level. If you are not *completely certain* about the level of language proficiency for a given student, then use double level mark (i.e., A2/B1 or B1/B2, etc.)."

References

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). *Final Report of the Dutch Construct Project*. (available on request: sjtakala@hotmail.com).

Alderson, J.C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2006). Analysing tests of reading and listening in relation to the Common European Framework of Reference for Languages: The experience of the Dutch CEFR construct project. *Language Assessment Quarterly*, *3*(1), 3–30. http://dx.doi.org/10.1207/s15434311laq0301_2

Berk, R. (1995). Something old, something new, something borrowed, a lot to do. *Applied Measurement in Education*, *8*(1), 99–109. http://dx.doi.org/10.1207/s15324818ame0801_8

Byram, M. & Parmenter, L. (Eds.). (2012). *The Common European Framework of Reference: The globalisation of language education policy*. Bristol, UK: Multilingual Matters.

Byrnes, H. (Ed.). (2007). The issue: The Common European Framework of Reference for Languages: Perspectives on the making of supranational language education policy. *Modern Language Journal*, *91*(4), 641–645. http://dx.doi.org/10.1111/j.1540-4781.2007.00627_1.x

Cizek, G.J. (Ed.). (2001). *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Erlbaum.

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Cohen, A., Kane, M., & Crooks, T. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education*, *12*(4), 343–366. http://dx.doi.org/10.1207/S15324818AME1204_2

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge University Press. Retrieved from http://www.coe.int/t/dg4/linguistic/cadre1_en.asp

Council of Europe (2003). *Relating language examinations to the common European framework of reference for languages: learning, teaching, assessment (CEF) Manual*. Preliminary Pilot Version. Language Policy Division, Strasbourg.

Council of Europe. (2004). *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the CEFR*. Language Policy Division. http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Council of Europe. (2009). *Manual for relating language examinations to the Common European Framework of Reference for Languages (CEFR).* The revised version. http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Figueras, N., & Noijons, J. (Eds.). (2009). *Linking to the CEFR levels: Research perspectives*. CITO, Council of Europe, & European Association for Language Testing and Assessment. www.ealta.eu.org/resources.htm

Fulcher, G. (2004). Deluded by artifices? The Common European Framework and harmonization. *Language Assessment Quarterly*, *I*(4) 253-266.

Haertel, E., & Lorie, E. (2004). Validating standards-based test score interpretation*s*. *Measurement*, *2*(2), 61–103.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: American Council on Education/Praeger.

Hawkins, J.A., & Filipović, L. (2012). *Criterial features in L2 English: Specifying the reference levels of the Common European Framework. English Profile Studies 1*. Cambridge, UK: Cambridge University Press.

Jaeger, R.M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice*, *10*(2), 3–6. http://dx.doi.org/10.1111/j.1745-3992.1991.tb00185.x

Kaftandjieva, F. (2004). *Section B: Standard setting*. Council of Europe Reference Supplement to the Preliminary Pilot Version of the Manual for Relating Language Examinations to the CEFR. Language Policy Division. http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests. A comparative analysis of six recent methods with an application to tests of reading in EFL*. European Association for Language Testing and Assessment. www.ealta.eu.org/resources.htm

Kaftandjieva, F., & Takala, S. (2002). *Relating the Finnish Matriculation Examination test results to the CEF scales*. Paper presented at the Helsinki Seminar on Linking Language Examinations to CEFR for Languages. June 31-July 2, 2002. http://kiesplang.fi

Martyniuk, W. (Ed.). (2010). *Aligning tests with the CEFR: Reflections on using the Council of Europe draft manual*. Cambridge, UK: Cambridge University Press.

Mills, C., & Melican, G. (1988). Estimating and adjusting cutoff scores: Features of selected methods. *Applied Measurement in Education*, *1*(3), 261–275. http://dx.doi.org/10.1207/s15324818ame0103_7

Norcini, J., Lipner, R., Langdon, L., & Strecker, C. (1987). A comparison of three variations on a standard-setting method. *Journal of Educational Measurement*, *24*(1), 56–64. http://dx.doi.org/10.1111/j.1745-3984.1987.tb00261.x

Reckase, M.D. (2009a). *Multidimensional item response theory*. Berlin, Germany: Springer. http://dx.doi.org/10.1007/978-0-387-89976-3

Reckase, M.D. (2009b). Standard setting theory and practice: Issues and difficulties. In N. Figueras & J. Noijons (Eds.), *Linking to the CEFR levels: Research perspectives*. Arnhem: CITO-EALTA. Retrieved from http://www.ealta.eu.org/resources.htm

Reckase, M.D. (2010). NCME 2009 presidential address: What I think I know. *Educational Measurement: Issues and Practice*, *29*(3), 3–7. http://dx.doi.org/10.1111/j.1745-3992.2010.00178.x

Reckase, M.D., & Chen, J. (2012). The role, format, and impact of feedback to standard setting panelists. In G.J. Cizek (Ed.), *Setting performance standards: Foundations, methods, and innovations* (2nd ed., pp. 149–164). New York, NY: Routledge.